Cost-effective and Collaborative Methods to Author Video's Scene Description for Blind People.

Rosiana Natalie
Singapore Management University
Singapore, Singapore
rnatalie.2019@phdcs.smu.edu.sg

ABSTRACT

The majority of online video content remains inaccessible for blind people due to the lack of audio descriptions. Content creators have traditionally relied on professionals to author audio descriptions, but their service is costly and not readily available. In this research, I introduce four threads of research that I will conduct for my Ph.D. dissertation, aimed to create methods and tools that are both time- and cost-effective in providing good quality audio descriptions. They are: (i) The development and evaluation of mixed-ability collaboration authoring tool, (ii) The formative study to uncover the feedback pattern from the reviewer, (iii) the evaluation and generation of real-time supports for novice authors to write AD, and (iv) the design, development, and evaluation of a system that demonstrate the utility of semi-automatically authoring AD. I believe these four research threads help me to uncover a cheaper solution to generate audio description. Hence, motivating the content creator to include this accessibility feature in the video production process and making the existing and upcoming videos accessible.

KEYWORDS

Audio Description, visual impairment, video accessibility, collaborative writing

ACM Reference Format:
Rosiana Natalie. 2022. Cost-effective and Collaborative Methods to Author Video's Scene Description for Blind People.. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3491101.3503814

1 INTRODUCTION

In the era of hundreds of hours of videos is uploaded online on YouTube [13] and a 20-40% increment of video streaming in the year 2020 due to the COVID19 pandemic [10], it is evident that video has become an essential way of daily communication amongst people. Yet, despite the great demand for videos, many blind people still cannot directly consume the content due to its visual nature. One solution to make videos more accessible for people with visual impairments is for content creators to provide Audio Descriptions (AD). AD verbally explains visual events that are not audible to blind users.

Despite the existence of AD, the availability of accessibility guidelines to follow (e.g., WCAG [15]), and the mandate from the anti-discrimination-related regulations to the provision of videos with audio descriptions [19], many videos remain inaccessible. For instance, according to the American Council of the Blind, only about 3, 042 out of 75 million videos come with AD [20, 27]. Also, this problem is partially attributed to the cost and the lack of availability of professional audio describers. For instance, the estimated cost of generating AD is about US$12 to US$75 [21, 31], which has disincentivized many casual content creators from adding AD to their videos.

My Ph.D. work is the first step toward making every video that has been and will be created in the future accessible for blind people. My research work aims to develop methods and tools to generate AD that are more affordable than hiring professional audio describers. These methods and tools should allow novices, who are relatively cheaper to hire than professionals, to generate good quality AD. Using the combination of mixed-ability collaboration, computer vision (CV), and natural language processing (NLP), I will design, develop, and evaluate methods to support novices to take part in the AD generation process. Ultimately, these efforts should encourage casual content creators to make their videos accessible and give those with visual impairments a helping aid to consuming video content.

This dissertation study will focus on answering these questions:

(1) *Is it cost-effective to involve novices in generating AD?*
(2) *What are the dimensions of AD qualities and "how good" are novice-created AD along those dimensions?*
(3) *How can automated adaptive feedback from machine learning output (specifically, computer vision and natural language processing) help novice audio describers generate good audio descriptions?*
(4) *Can humans and machines collaborate to generate good audio descriptions?*

To answer these questions, I am working on four research threads: (i) the development and evaluation of mixed-ability collaboration authoring tool, (ii) The formative study to uncover the feedback pattern from the reviewer, (iii) the evaluation and generation of real-time supports for novice authors to write AD, and (iv) the design, development, and evaluation of a system that demonstrate the utility of semi-automatically authoring AD

2 BACKGROUND

My work proposes technical solutions with friendly user interfaces that streamline and automate parts of the authoring process to

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3503814

Variable	Description
Descriptive	SD provides a pictorial and kinetic description of objects, people, and settings and explains how people act in the scene. [1, 11, 30]
Objective	SD illustrates objects, people, or relationships between them in an unbiased manner using objective language. SD should not include a speculative description. SD should also avoid making their own inference about what is occurring in the scene without proper evidence. [1, 9, 11, 30]
Succinct	An audio generated from textual SD should fit in a gap without a dialogue or a natural pause in the video. [11, 30]
Learning	SD should convey the video's intended message to the audience. [11]
Sufficient	SD should depict all the scenes and provide sufficient information for the audience to comprehend the content of a video while not being overly descriptive.[22]
Accurate (^)	SD should not provide the incorrect information of what is shown in the scene [11, 30]
Referable (^)	SD should use language that is accessible to everyone with different disabilities. The use of demonstrative pronouns like "this", "there", "that" is not Referable as it is not understandable for people with visual impairments because these pronouns need to be complemented with visual help. [11]
Interest	SD should make the video be interesting for the audience by writing a cohesive narrative. The tone of the description should reflect the tone of the video. [12, 26, 32]
Clarity	SD should communicate descriptive information in a language and manner that are easy to follow for people with visual impairments. [32]

Table 1: Scene description quality codebook. Note: ^ = these variables are assessed only by the sighted evaluator. The complete codebook can be accessed at [24]

minimize the time and cost of creating ADs. In the status quo, these are the current solutions proposed by other researchers to create affordable ADs. For example, Campos et al. designed a way to automatically script audio descriptions using the pre-existing video scripts (i.e., the blueprint of a chronological rundown of the scenes) and subtitles [8]. Kobayashi et al. [16] created a script editor that allows novice authors to easily edit the audio description and modify synthetic speech parameters. Pavel et al.'s Rescribe automated the editing and revision processes to fit the content into limited space for audio descriptions using dynamic programming [29].

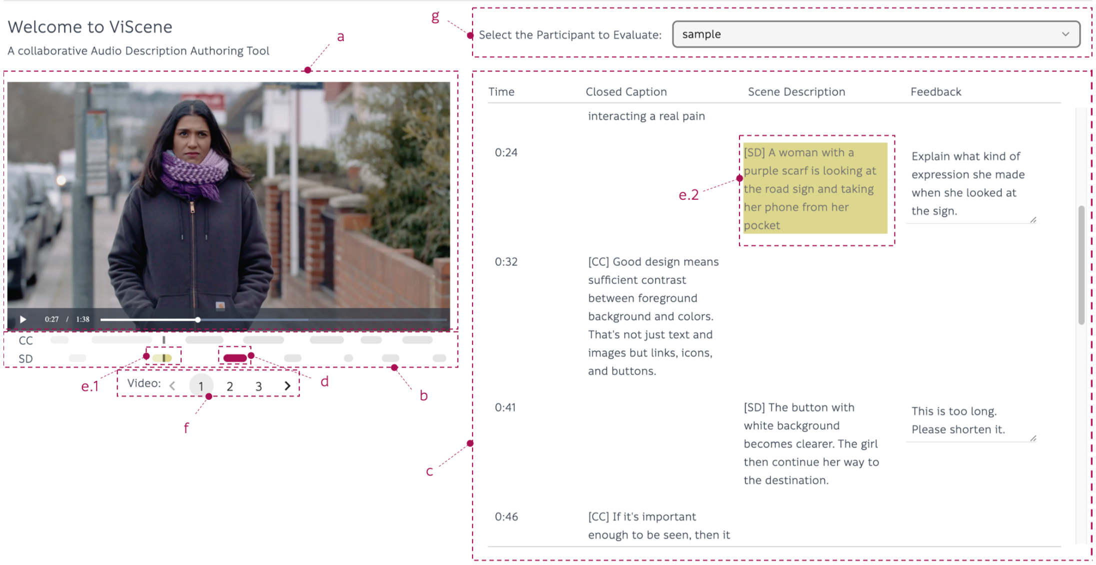
Another viable option to achieve the overarching goal of making all online videos accessible and improving current methods is by automatically generating AD using CV techniques. It is the most promising way of offering a cheap and scalable approach to generate AD. For example, Wang et al. [33] were the first to build and evaluate a system to automate AD for blind people fully. However, the result of the fully automatically generated AD still lacks useful information for blind people (e.g., the character actions, gender, places) and concluded that the quality of the automatically generated AD is subpar.

A promising strategy to improve automatically generating the AD is by introducing human-in-the-loop (HITL). [35, 36]. In these studies, the authors let novice describers edit the AD, and it was evident, as evaluated by blind raters, that novices can improve the AD quality [35]. Even in the scenario where novice describers may not be aware of what constitutes a good AD, a nudge in the right direction (e.g., in the form of feedback) would help them improve the quality of their edits further. That is why in my research, I aim to empower novices by providing helpful information to the automatically generated AD via guidance from automatic and real-time feedback.

All in all, while the prior studies on providing the audio description are in the right direction, they are still far from making all online videos accessible. These methods are still costly to create good quality AD, or the quality of the generated AD was still subpar. In my study, I have demonstrated the feasibility of supporting novices to create affordable AD with good quality using feedback [23, 24]. The prior research [3, 14, 17] inspires this study have shown the benefit of the adoption of peer evaluation to improve one's writing. Moreover, the mixed-ability collaboration in my work which is a novel framing in the context of AD is inspired by the prior studies that show how blind and sighted people, in pairs, can co-create an accessible workspace [5] and household [4]. I believe that with the help of commentary feedback from their peers (i.e., reviewers), novices can transform the nature of creating AD for non-accessible videos.

3 RESULT TO DATE

To create more cost- and time- efficient method to produce AD, my dissertation comprises of four threads of research: (i) The development and evaluation of mixed-ability collaboration authoring tool, (ii) The formative study to uncover the feedback pattern from the reviewer, (iii) the evaluation and generation of real-time supports for novice authors to write AD, and (iv) the design, development, and evaluation of a system that demonstrate the utility of semi-automatically authoring AD. By discovering the outcome of these threads, I would like to motivate more content creators to create a habit to accompany their videos with AD. Hence, by the extrapolation, all the existing and upcoming videos can become accessible.



Welcome to ViScene

A collaborative Audio Description Authoring Tool

Select the Participant to Evaluate: sample

Time	Closed Caption	Scene Description	Feedback
	interacting a real pain		
0:24		[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket	Explain what kind of expression she made when she looked at the sign.
0:32	[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.		
0:41		[SD] The button with white background becomes clearer. The girl then continue her way to the destination.	This is too long. Please shorten it.
0:46	[CC] If it's important enough to be seen, then it		

Figure 1: ViScene's interface. (a) the video pane; (b) closed captions (CC) and scene descriptions (SD) bars; (c) a table with Time, CC, SD, and Feedback columns; (d) SD succinctness feedback, (e) CC/SD text-segment correspondence visualization, (f) video selector, and (g) author dropdown selector (for reviewers).

3.1 Mixed-ability Collaborative Audio Description Authoring.

In my preliminary study, I showed that it was possible to make AD affordable by harnessing the ability of novices. My goal is to create a volunteer-based AD authoring system like youdescribe.org [34], where any online volunteer can create AD for YouTube videos. I developed a Viscene (Fig 1), a collaboration tool that allows a novice to author audio descriptions, receive feedback from sighted or blind reviewers and revise the scene descriptions [23]. To extend this study, I have also conducted a user study with 60 novices to generate the scene description (SD)– a textual description of a scene that was later transformed into speech using text-to-speech (TTS) technology. In addition, I evaluated the utility of feedback in mixed-ability collaboration by analyzing the revised AD with the help of comments from sighted or blind reviewers [24].

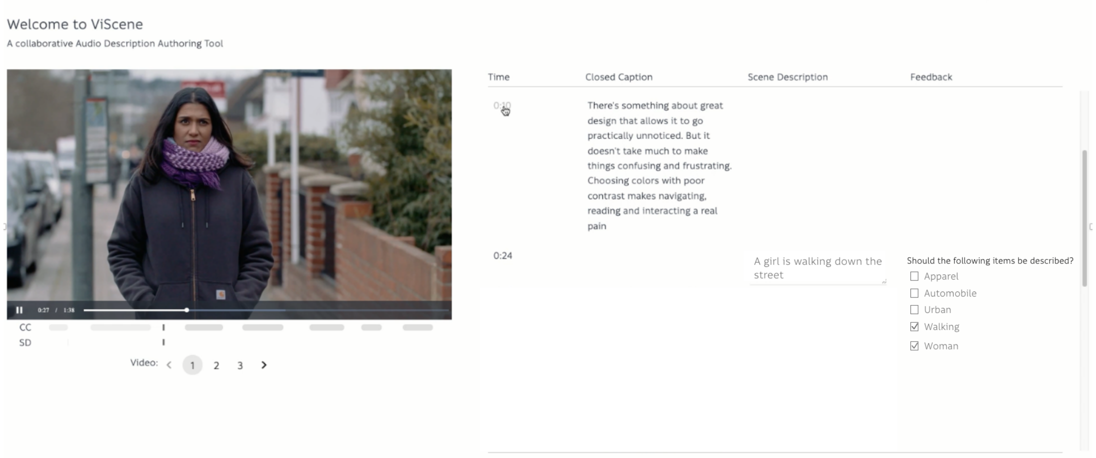
The AD evaluation followed my proposed concise codebook, grounded in experts' guidelines that define nine qualities to assess SD: Descriptive, Objective, Succinct, Learning, Sufficient, Accurate, Referable, Interest, and Clarity [6, 9, 11, 22, 26, 28, 30, 32, 34] (Refer to Table 1). From the results, I observed the usefulness of feedback to improve the nine qualities of novice written SD. Specifically, novices can enhance the quality of AD in terms of its descriptiveness, objectivity, referability, and clarity qualities. My study suggested that author-reviewer collaboration, especially that with a sighted author and a blind reviewer, is a promising method to enhance the SD quality.

However, the fully manual authoring and reviewing process is time-consuming and prevents us from adding ADs to online videos. In an era where 500 hours of videos are uploaded every minute on YouTube alone [13], it is currently improbable that ViScene would be able to generate SDs for every single video. In my prior study, the average time to create AD for videos, where each video is around one minute long, is 50-56 minutes. By extrapolation, it would take almost a day to generate an SD for a video that lasts 24 minutes, rendering ViScene not scalable, e.g., for long videos. Thus, as I describe next, automating some of the processes (e.g., reviewing and authoring) in generating AD is critical and has high potential to reduce the time and cost.

3.2 The formative study to uncover the feedback pattern from the reviewer

The authoring and reviewing processes employed in this version of ViScene were all manual. The SD authors did not receive the commentary feedback in real-time, which can disrupt the overall process. I am currently investigating the efficacy of automatically generating some feedback, and we see a lot of room for innovation in this area.

Thus far, I have conducted a preliminary study to uncover the feedback pattern from the reviewer [25]. My collaborator and I performed a content analysis by open coding the reviews we got from sighted and blind reviewers in my prior study. We analyzed in total 1,120 reviews from both sighted and blind reviews. We conclude that study by summarizing four themes in the reviewers' comment: (i) Quality; commenting on different AD quality variables, (ii) Speech Act; the utterance or speech action that the reviewers used, (iii) Required Action; the recommended action that the authors should do to improve the AD, and (iv) Guidance; the additional help that the reviewers gave to help the authors. We believe that the result informs the design of the future system to generate an automatically generated review. For example, we found that sighted and blind reviewers mostly gave feedback on Descriptive, Sufficient,



Welcome to ViScene

A collaborative Audio Description Authoring Tool

Time	Closed Caption	Scene Description	Feedback
0:10	There's something about great design that allows it to go practically unnoticed. But it doesn't take much to make things confusing and frustrating. Choosing colors with poor contrast makes navigating, reading and interacting a real pain		
0:24		A girl is walking down the street	Should the following items be described? □ Apparel □ Automobile □ Urban ☑ Walking ☑ Woman

Figure 2: Proposed prototype for ViScene Interface with the Adaptive Feedback. This interface inherited all the components of ViScene’s previous version (Figure. 1), except for the Feedback column it is now shown in checklist format. This format will allow the participant to know which item has/has not been described

and Clarity variables. Thus, for the subsequent study, I will explore ways to generate feedback targeted to improve these variables.

4 CURRENT AND NEXT STEP

4.1 Semi-automatically Authoring and Reviewing Audio Descriptions

Currently, I am exploring methods to generate automated adaptive feedback for the novice author. I envision ViScene to analyse the SDs written by the novice author and provide feedback on areas that they should describe. The feedback given in such a manner is "adaptive" since it get customed to the novice authors’ input. The amount of feedback varies following the amount of information expressed in the SDs generated by the novice authors. I leverage computer vision and natural language processing techniques to generate automated adaptive feedback. I make use of an existing scene understanding service (i.e., Amazon Rekognition [2]) to generate labels of objects, actions, and settings in the scene. I then clustered the labels based on their semantic meaning and selected a label from each cluster as the cluster representative. Next, given the participants’ SD, I match the SD with each of the cluster representatives to find the cluster that matches with the SD. The system gives feedback when the novice authors do not describe some of the labels within the matched cluster. I then present the feedback in the form of a checklist of labels that need to be described in the SD (see Fig. 2 for the system prototype.)

After I have finished developing the above system, I plan to evaluate its efficacy by conducting a user study with novice authors as the participants. I will conduct a between-subject study and consider the feedback types (i.e., without-feedback, human-feedback, generic-feedback, and adaptive-feedback) as the factor. In the control group (i.e., without-feedback condition), the novice authors will not receive any feedback. The novice authors will receive feedback from a sighted reviewer for human-feedback condition. For generic feedback, the novice authors will receive automatically generic feedback, such as "Do not forget to explain the objects you see in the video," "Do not forget to explain the SD clearly." Lastly, for adaptive feedback, the participants will receive the feedback generated using my automated adaptive feedback system.

To evaluate how much my adaptive feedback system has helped the novice authors generate good quality SDs, I will measure the quality of the generated SDs following the codebook I developed and presented in [24]. I will also assess the time to generate the SD, as well as the number of edits the novice authors performed. It is also important to know if the proposed method for generating SD is favorable to the user. Hence, I will administer a semi-structured interview with the participants to gauge the participants’ experience and use System Usability Scale (SUS) [7] to measure the system usability.

4.2 Fully-automatically Generating Audio Descriptions and Human in the loop.

Another fertile research topic would be automating the authoring of SDsS itself. We are excited to see a few efforts this past year in this direction (e.g., [33, 35, 36]), though there is much to be done for achieving high-quality audio descriptions. Inspired by Yuksel et al. [35, 36], my goal is to evaluate if the automatically generated AD can be useful as a first draft that can be efficiently post-edited by human authors to improve the AD quality. Then, I will explore and incorporate the state-of-the-art CV technique to generate AD automatically, such as [18], a memory module to augment the transformer architecture to generate multi-sentence video. After that, I will let the novice author post-edit the automatically generated AD and improve its quality.

To highlight the difference between our proposed system and the one created by Yuksel et al. [35, 36], I will also incorporate automatically generated feedback explained above. Thus, while the novice is post-editing, s/he will be continuously receiving feedback on what to revise. Not only it aims to guide the novices, but also it has the potential to reduce the number of iterations a participant needs to go through to generate AD (i.e., from three stages: (i) authoring, (ii) waiting for feedback, (iii) revision, the iteration is reduced to only one iteration: a revision with real-time feedback). Thus, it will also reduce the overall AD generation cost.

5 ACADEMIC STATUS

I am currently at my third-year PhD student at Singapore Management University under the supervision of Dr. Kotaro Hara. This is a 4-to-5-year PhD program, and I am enrolled in a Full-time program. I expect to present my dissertation in 2024.

REFERENCES

[1] 3PlayMedia. 2020. Beginner's Guid to Audio Description. https://go.3playmedia.com/hubfs/WP%20PDFs/Beginners-Guide-to-Audio-Description.pdf. Accessed: 2021-01-13.

[2] Amazon. 2021. Amazon Rekognition. https://aws.amazon.com/rekognition/?blog-cards.sort-by=item.additionalFields.createdDate&blog-cards.sort-order=desc.. Accessed: 2021 - 10 -09.

[3] Fabricio Balcazar, Bill L Hopkins, and Yolanda Suarez. 1985. A critical, objective review of performance feedback. *Journal of Organizational Behavior Management* 7, 3-4 (1985), 65–89.

[4] Stacy M Branham and Shaun K Kane. 2015. Collaborative accessibility: How blind and sighted companions co-create accessible home spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2373–2382.

[5] Stacy M Branham and Shaun K Kane. 2015. The invisible work of accessibility: how blind employees manage accessibility in mixed-ability workplaces. In *Proceedings of the 17th international acm sigaccess conference on computers & accessibility*. 163–171.

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

[7] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[8] Virginia P Campos, Tiago MU de Araújo, Guido L de Souza Filho, and Luiz MG Gonçalves. 2020. CineAD: a system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society* 19, 1 (2020), 99–111.

[9] Audio Description Coalition. 2009. Standards for Audio Description and Code of Professional Conduct for Describers. https://audiodescriptionsolutions.com/wp-content/uploads/2016/06/adc_standards_090615.pdf. Accessed: 2020-11-6.

[10] Comcast. 2020. Comcast 2020 Network Report. https://update.comcast.com/wp-content/uploads/sites/33/dlm_uploads/2021/02/network-report-2020.pdf.

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

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

[13] James Hale. 2019. More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute. https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/. Accessed: 2020-11-5.

[14] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[15] World-Wide Web COnsortium Web Accessibility Initiative. 2016. Making the Web-Accessible. https://www.w3.org/WAI/. Accessed: 2020-11-6.

[16] 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.

[17] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the second (2015) ACM conference on learning@ scale*. 75–84.

[18] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2603–2614.

[19] Hoi Ching Dawning Leung. 2018. *Audio description of audiovisual programmes for the visually impaired in Hong Kong.* Ph.D. Dissertation. UCL (University College London).

[20] Market.US. 2020. Amazon Prime Video Statistics and Facts. https://market.us/statistics/online-video-and-streaming-sites/amazon-prime-video/. Accessed:2021-06-15.

[21] Chris Mikul. 2010. Audio description background paper. *Media Access Australia* (2010).

[22] 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.

[23] Rosiana Natalie, Ebrima Jarjue, Hernisa Kacorri, and Kotaro Hara. 2020. ViScene: A Collaborative Authoring Tool for Scene Descriptions in Videos. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

[24] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-ren Chan, Ebrima Jarjue, Hernisa Kacorri, and Kotaro Hara. 2021. Efficacy of Collaborative Authoring of Video Scene Descriptions. (2021). To appear in The 23rd International ACM SIGACCESS Conference on Computers and Accessibility.

[25] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-ren Chan, Ebrima Jarjue, Hernisa Kacorri, and Kotaro Hara. 2021. Uncovering Patterns in Reviewers' Feedback to Scene Description Authors. (2021). To appear in The 23rd International ACM SIGACCESS Conference on Computers and Accessibility.

[26] Netflix. 2020. Audio Description Style Guide v2.1. https://partnerhelp.netflixstudios.com/hc/en-us/articles/215510667-Audio-Description-Style-Guide-v2-1. Accessed: 2020-11-6.

[27] American Council of The Blind. 2021. Amazon Prime Video Audio Described Titles. https://acb.org/adp/amazonad.html. Accessed:2021-06-15.

[28] American Council of the Blind. 2021. Audio Description using the Web Speech API. https://acb.org/adp/education.html. Accessed: 2020-01-13.

[29] Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 747–759.

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

[31] Terril Thompson. 2019. Audio Description using the Web Speech API. https://terrillthompson.com/1173. Accessed: 2020-11-6.

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

[33] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos. (2021).

[34] YouDescribe. 2020. YouDescribe. https://youdescribe.org/support/tutorial. Accessed: 2020-11-6.

[35] Beste F Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Seung Jung Jin, Yue-Ting Siu, Joshua A Miele, and Ilmi Yoon. 2020. Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 47–60.

[36] 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.