Toward Automatic Audio Description Generation for Accessible Videos (Supplementary Materials)

Yujia Wang
Beijing Institute of Technology
George Mason University

Wei Liang*
Beijing Institute of Technology

Haikun Huang
George Mason University

Yongqi Zhang
George Mason University

Dingzeyu Li
Adobe Research

Lap-Fai Yu
George Mason University

ABSTRACT

In this supplementary, we provide: **1)** demographics of sighted users participating in our system evaluation; **2)** user-interface illustration of our designed and implemented program; **3)** analysis of user feedback in pre-study interview; **4)** cross-video type evaluation analysis of our generated audio descriptions in terms of additional information requirement, audio description confusion, redundancy, and grammar errors; **5)** additional interesting findings.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools**; *Accessibility technologies.*

KEYWORDS

audio description, video description, audio-visual consistency, video captioning, sentence-level embedding, accessibility

ACM Reference Format:
Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos (Supplementary Materials). In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3411764.3445347

1 SIGHTED PARTICIPANTS DEMOGRAPHICS

With an IRB approval, we recruited 20 sighted participants (p01-p20). They reported normal or corrected-to-normal vision, no color-blindness, and normal hearing. As shown in Table 1, the sighted participants represent a diversity of backgrounds in terms of gender (10 female, 10 male), age (range is 19 to 56 with a mean of 39.05), occupation (from people who are retired, to those who are students, managers, journalist, professor, etc.). These participants had a range of accessing platforms, as well as varied online video experiences.

*Corresponding author.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445347

2 USER INTERFACE

Due to COVID-19, we conducted all the interviews and studies virtually via Zoom. We designed and implemented a program as shown in Fig. 1. For BVI participants, we asked them to raise their hand and give evaluations. For the sighted participants, we sent them the executable program and instructed them to perform the same operations as BVI participants by clicking different buttons, *e.g.*, clicking "A" for *Additional information*, clicking "C" for *Confusing description*, etc.

3 PRE-STUDY INTERVIEW

Before the study, we asked both sighted and BVI participants about their experience of accessible videos. The specific questions can be found in the Appendix.

7 out of 20 sighted users who wanted to be eyes-free complained about the lack of audio descriptions of videos in social media platforms (*e.g.*, YouTube, Instagram), but stated the movies, series, etc. with audio description in video streaming platforms, like Netflix and Amazon Prime Video, were satisfactory. Therefore, most participants (12) use audio description service only when watching movies, series, or documents on video streaming platforms. Participants who like to watch sports games (4) stated that they would like to use broadcast or live texts to access the game when can not watch the video, *e.g.*, when driving the car, because such video not always equipped with audio descriptions, especially on-live.

The most common accessibility issue mentioned by BVI participants was the lack of audio descriptions for online videos except for movies. They noted that most videos that edited, uploaded, or published by users did not include audio descriptions. Some participants said they could sometimes guess the video content based on what they heard, but they commonly said that they could only describe the video roughly. P23 noted that she sometimes gets completely misunderstanding of the video content due to excessive imagination with the video title read through the screen reader. Our system is designed on the considerations of these barriers to further explore the results of the research question.

4 CROSS-VIDEO TYPE ANALYSIS

To understand how participants perceived the generated audio description of different video types, we further analyzed the three operations' results of different combinations of participants and validation data. As demonstrated in Table 2, for human or animal-related video types, *e.g.*, animals, film, activity, comedy, and DIY,

Table 1: Demographics of sighted participants.

ID	Gender	Age	Occupation	Video Platforms	Video Types
P01	M	25	Student	YouTube, Netflix	Music, sports
P02	M	32	Student	YouTube	Current affairs, popular science
P03	F	27	Student	YouTube, Netflix	Movie, series, TV shows
P04	M	28	Manager	YouTube	Documentary, sports, pets
P05	M	27	Student	Netflix	Movie
P06	F	26	Student	YouTube	Music, Vlog, cooking Recipe
P07	F	56	Retired	TV	Music, fitness
P08	F	29	Student	YouTube, Netflix, Instagram	Movie, series, fashion Vlog
P09	F	30	Programmer	YouTube	Pets
P10	M	29	Engineer	YouTube	News, sports, travel Vlog
P11	M	57	Professor	TV	News, documentary
P12	F	57	Retired	YouTube, TV	Fitness, DIY, news
P13	F	19	Student	YouTube	Game, movie
P14	M	28	Journalist	YouTube, TV	News, sports
P15	M	24	Student	YouTube, TV	Sports, comedy, game
P16	M	37	Manager	YouTube	Music, sports, news
P17	F	27	Office Assistant	Netflix	Movie, cartoon
P18	F	24	Student	YouTube	Movie, Vlog
P19	F	35	Office Assistant	YouTube, Netflix	Movie, parenting, shopping Vlog
P20	M	25	Student	YouTube	Education videos

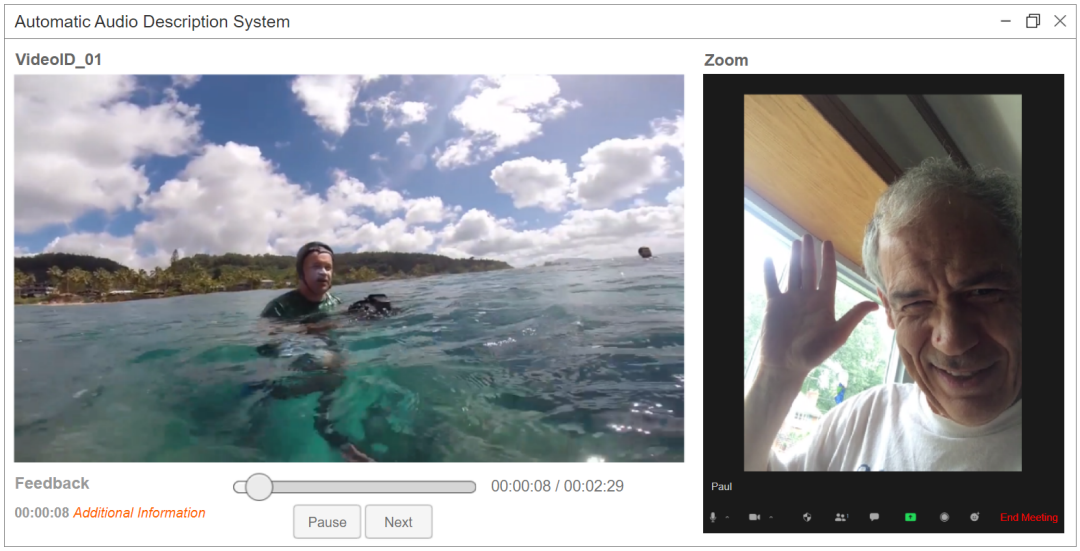

Figure 1: The user interface of our automatic audio description system that used in the evaluation experiments. The video/audio player is on the left. The participants raise their hands during a Zoom online meeting to add annotations. The screenshot on the right shows one of our BVI participants, Paul, who gave permission to show his photo taken during the study. ©Paul D'Addario

both sighted and BVI users have the same demand on the description quantity. For physical-related videos, including sports, dancing, and fitness, 78% of the sighted participants mentioned that such videos should be briefly described the players' actions, or only be described at the greatest moment, while BVI users would like to have more detailed descriptions as much as possible.

For sighted participants, when listening to videos, the description perplexities are significantly lower than that of watching videos. They noted that due to the absence of visual information, they were skeptical of the described video content. 50% of the sighted users also claimed their priors of some specific video types, like comedy. They would picture the video content in mind based on the

Table 2: Description perception statistics of different evaluations.

User & Data	Video Type								
	Music	Animals	Film	Activity	Comedy	DIY	Sports	Dancing	Fitness
	Insertion QTY (User demands vs. Our results)								
SV	>	>	>	>	>	>	<	<	<
SA	<	>	>	>	>	>	<	<	<
B	<	>	>	>	>	>	>	>	>
	Confusion								
SV	50.00%	42.86%	86.67%	56.52%	72.73%	88.89%	34.78%	5.00%	23.08%
SA	33.33%	40.00%	69.23%	18.75%	83.33%	80.00%	38.10%	7.69%	7.69%
B	17.39%	45.00%	31.25%	25.93%	27.27%	83.33%	47.35%	0	0
	Redundancy & Grammar Errors								
SV	11.43%	28.57%	6.67%	4.35%	9.09%	22.22%	26.09%	70.00%	30.77%
SA	67.67%	16.00%	15.38%	31.25%	33.33%	60.00%	19.05%	11.54%	30.77%
B	0	0	0	0	0	0	0	0	4.17%

video title and what heard, which explains the higher complexity of sighted users accessing *Audio Set*. The perplexity results of BVI users are relatively lower than that of sighted users (accessing *Video Set*), since they usually changed the initial thoughts other than confused about the description, which is different from the views of sighted users.

For the evaluation of redundancy and grammar errors, there are significant different trends in physical-related videos (*i.e.* decreased *R* on *Audio Set* compared to that on *Video Set*) and human-related videos (*i.e.* increased *R* on *Audio set* compared to that on *Video Set*). As P16 explained, *"For sports video, I didn't mind the redundant descriptions because I would like to know the real-time status of the players. On the contrary, I couldn't get enough information from the same descriptions of human-related videos, because there would many uncertain factors to influence my understanding, like the scene changes and new character appear. I would like to know more details rather than the same descriptions."*

5 ADDITIONAL FINDINGS

While we designed our study to focus on description evaluation, additional interesting findings emerged.

Perception of Detailed Sound. During the analysis of the data with respect to operation *A*, *i.e.* description insertion, we made a general observation that both sighted and BVI participants are sensitive to changes in background music, covering changes in speed, rhythm, instruments, pitch, etc. As P25 noted, *"This video has only background music and no dialogue or scene noise. I think the scene has changed when the music changes, and there should be a corresponding sentence to describe what was going on."*

For detailed sound, under the same circumstance, *i.e.* , inability to perceive visual information, BVI participants are more sensitive to detailed sounds than sighted participants. For instance, BVI participants stated that *"I heard the noise of a shopping cart. The shopping cart seems to be rolling on through at the beginning of the video."* The noise of the shopping cart lasts 5 seconds in very low volume even covered with other vehicle noises. However, no sighted participants who listened to this video noticed such detailed sound.

Inference of Visual Content. Several participants in our study stated that they would usually like to infer the people's appearance appearing in the video. Interestingly, their inferences achieve high accuracy in three video types, *i.e.* music video, movies, and physical related videos (*i.e.* sports, dancing, and fitness videos). For instance, P24 stated, *"I would say the actor is maybe the tall guy with like a dark jacket on." "I think the person who playing the violin is female." "There is man in blue and he is playing with the stick. I'm guessing he is older."* All these inferences are consistent with the visual content. For physical related videos, both sighted participants and BVI participants can easily describe the cloth style of players or dancers based on their prior knowledge.