2023
The Journal on
Technology and
Persons with
Disabilities

You Described, We Archived: A Rich Audio Description Dataset

Charity Pitcher-Cooper[1], Manali Seth[2], Benjamin Kao[2], James M. Coughlan[1], Ilmi Yoon[2]
Smith-Kettlewell Eye Research Institute[1]
Computer Science, San Francisco State University[2]
cpc@ski.org, mseth@mail.sfsu.edu, bkao1@mail.sfsu.edu, coughlan@ski.org, ilmi@sfsu.edu

Abstract

The You Described, We Archived dataset (YuWA) is a collaboration between San Francisco State University and The Smith-Kettlewell Eye Research Institute. It includes audio description (AD) data collected worldwide 2013-2022 through YouDescribe, an accessibility tool for adding audio descriptions to YouTube videos. YouDescribe, a web-based audio description tool along with an iOS viewing app, has a community of 12,000+ average annual visitors, with approximately 3,000 volunteer describers, and has created over 5,500 audio described YouTube videos. Blind and visually impaired (BVI) viewers request videos, which then are saved to a wish list and volunteer audio describers select a video, write a script, record audio clips, and edit clip placement to create an audio description. The AD tracks are stored separately, posted for public view at https://youdescribe.org/ and played together with the YouTube video. The YuWA audio description data paired with the describer and viewer metadata, and collection timeline has a large number of research applications including artificial intelligence, machine learning, sociolinguistics, audio description, video understanding, video retrieval and video-language grounding tasks.

Keywords

Video Accessibility, Blind and Low Vision, Audio Description, Artificial Intelligence, Machine Learning, Sociolinguistics.

Journal on Technology and Persons with Disabilities
Santiago, J. (Eds): CSUN Assistive Technology Conference

Introduction

Thousands of videos are uploaded to online video platforms daily; unfortunately, the visual content of these videos is inaccessible to blind and visually impaired (BVI) individuals and a lack of accessible video content has a profound negative impact on learning and community connections (Packer et al.). According to the World Health Organization, at least 2.2 billion people globally are visually impaired (World Health Organization). Over eight million Americans (3.3 % by population) are visually impaired and may rely on a screen reader, screen magnifier or have a form of color blindness ("Accessibility Statistics | Interactive Accessibility," "53 Web Accessibility Statistics [Updated for 2022]").

The video platform YouTube is the second largest search engine with 500 hours of video content uploaded every minute, and 5 billion videos watched per day ("30 Eye-Opening YouTube Facts"). It is a major video content source for: Film & Animation, Music, Autos & Vehicles, Travel & Events, Pets & Animals, Sports, People & Blogs, and Gaming. An effective way to bridge the video accessibility gap is to add audio descriptions, an additional narration track providing visual information unable to be inferred from audio cues alone such as setting, facial expressions, gestures, on-screen text, style of dress, or any other relevant information, synchronized with the video that can be turned on or off as needed.

YouDescribe (YD) is a unique, free, crowdsourcing tool for adding audio descriptions to YouTube videos ("YouDescribe - Audio Description for YouTube Videos"). It was first launched in 2013 as a project of The Smith-Kettlewell Video Description Research and Development Center by scientist Dr. Joshua Miele. The current version of YouDescribe has been in continuous use since a major renovation in 2017. With the accelerating inclusion of short-form (rapidly produced, under 5 minute) videos via social media, the need for timely AD has only become more critical. Audio description made with YouDescribe is a complementary service to professional quality AD; it strives to be accurate, practicable, timely and fun. While most recent movies and TV shows have professional AD available, AD for home movies, educational videos for work and school, and short-form content is in very high demand.

Discussion

Audio description is an art that requires critical thinking about content importance, as well as a precise vocabulary to create harmonious descriptions with the fewest number of syllables to fit the space imposed by comprehension-critical soundscape and dialog. More than crowdsourcing, YouDescribe is a community of people who want accessible video, combined with volunteers who have a vested interest in creating useful content. At YouDescribe, viewers have an active role in requesting content that is important and enjoyable to them. The YouDescribe AD tool focuses on functionality for individual viewers and ease of use for volunteer describers. Training is encouraged for all new audio describers through in-person or virtual training provided by: The Smith-Kettlewell Eye Research Institute, a formal class provided by a describer organization, college class, apprenticeship, or through YouDescribe's accessible, text-based tutorials with a complimentary educational YouTube channel ("Accessibility: A Guide to Building Future User Interfaces"; "Literature and Disability"; "The Art of Writing"; "YouDescribe - Audio Description for YouTube Videos"; "YouDescribe - YouTube"). While it is possible to use YouDescribe through trial and error, the text-based training materials are brief and cover both how to use YouDescribe as well as standard industry-wide AD guidelines. The tutorial page is one of the most common visitor-accessed pages, second to the landing page. Visitors are provided step-by-step instructions on where the control buttons are located, how to request wish list items, essential describer training and general troubleshooting. The Smith-Kettlewell Eye Research Institute has sponsored training since 2013 with inventor Joshua Miele, Professor Yue-Ting Siu and trained audio describer Charity Pitcher-Cooper. Populations brand new to the concept of audio description are generally able to complete their first 3-minute AD, on a video with content well known to them in about 1.5 hours and describers already familiar with AD take about an hour for their first description made with YouDescribe (Pitcher-Cooper and Brabyn). Experienced describers often choose research-heavy projects like AD for museum pieces or full TV shows and feature films requiring many hours of background study, as well as lengthy recording time.

The You Described, We Archived (YuWA) dataset protects the privacy of the viewer and describer community and provides only anonymized data. For the purposes of this paper, we are concentrating on metadata for visitors and volunteers who used the current YouDescribe web interface officially launched May 17th, 2017. Audio descriptions made before Spring, 2017 were created with a similar but not identical interface launched in 2013. That interface became unstable for use and was completely rebuilt in 2017. The YuWA data repository includes all YouDescribe related AD from 2013-2022 and can be sorted to include or exclude important YouDescribe milestones.

Google Analytics for https://youdescribe.org was implemented in July, 2020 and reflects visitor trends for the past two years (Fig. 1). Of the 35 thousand visitors, 65 percent of visitors are from the US with notable traffic from Canada, United Kingdom, and Australia (Fig. 2). Around 14,000+ users have accessed the website directly, 11,000+ were referral traffic (the segment of traffic that arrives at YouDescribe from a link on another domain), and 11,000+ users visited via Organic Search, Social or Email combined (Fig. 3). Describers range in age from 13 and above. Approximately 40% of the web app traffic is male, and 60% is female (Fig. 4).

(These statistics are a general overview of visitor trends; Google Analytics for age and gender isn't available for all visitors and not all collected data is accurate; for example, those under 18 have no tracked data, and collected account information may not align with a visitor's gender). Our largest number of audio descriptions, 5895 videos out of a total 6,484, are recorded in English; language tags for English include Australia, Belize, Canada, Ireland, Jamaica, New Zealand, South Africa, Trinidad, United Kingdom, and the United States. The most compatible browser for the YouDescribe website is Google Chrome. As of August 2022, Chrome is the leading internet browser in the world with a global market share of 65.52 percent ("Most Popular Web Browsers in 2022 | Oberlo"). YouDescribe visitors also enjoyed viewing in Safari (Apple), Edge (Microsoft), and Firefox (Mozilla).

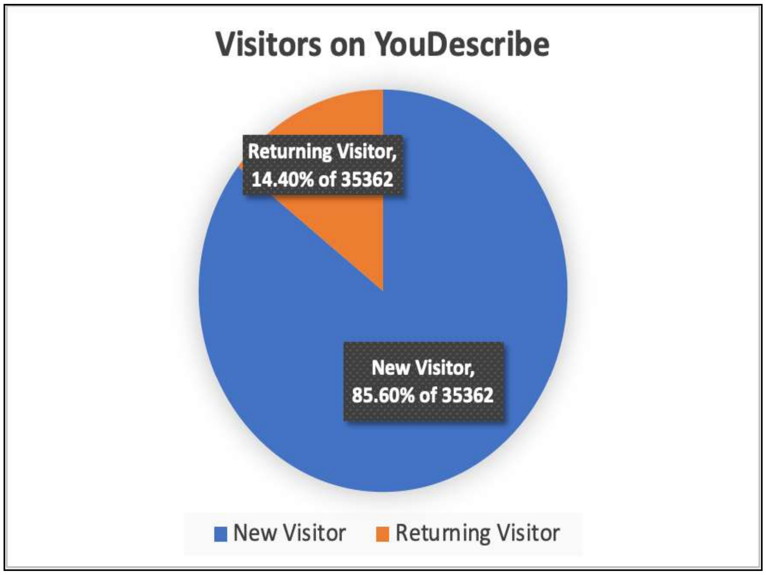

Fig. 1. Total, New and Returning Visitors (07/2020-09/2022, Google Analytics).

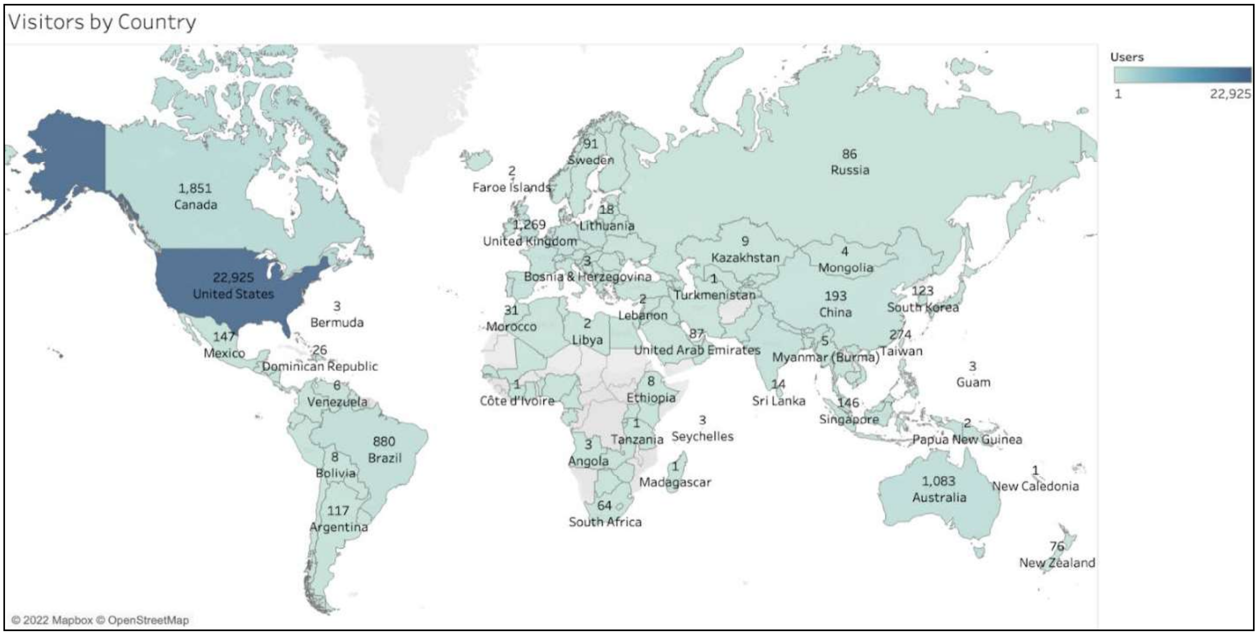

Fig. 2. Choropleth map of visitors by country (07/2020-09/2022, Google Analytics).

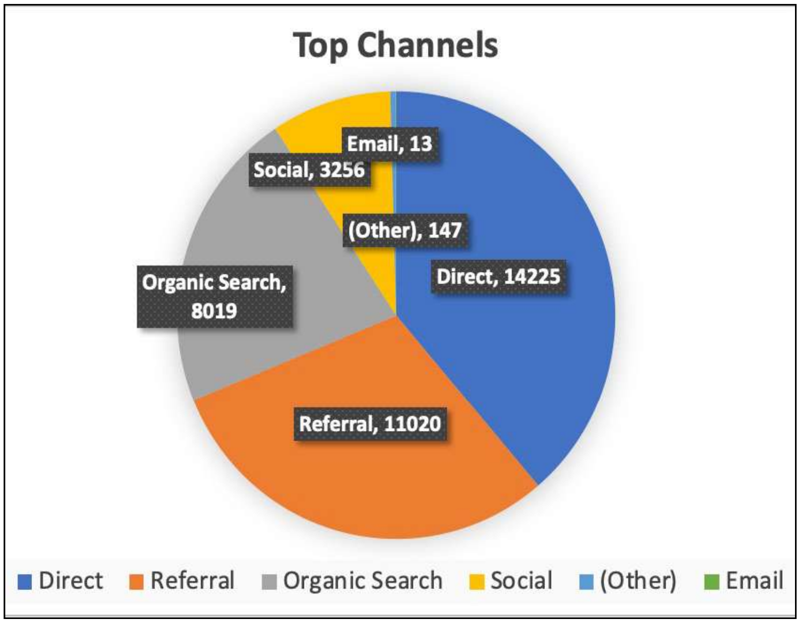

Fig. 3. Top Channel Traffic for YouDescribe (07/2020-09/2022, Google Analytics).

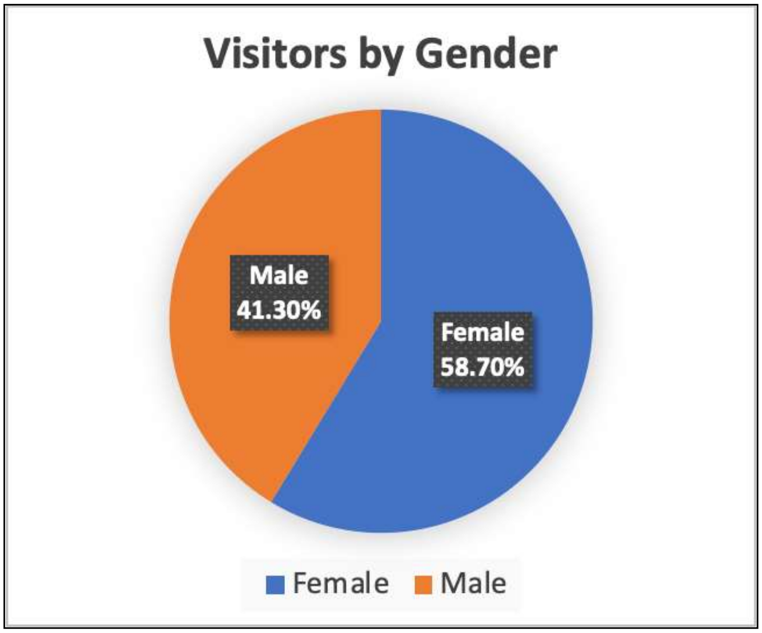

Fig. 4. Visitors by Gender (07/2020-09/2022, Google Analytics).

Our wish list, a curated list of what videos viewers want to have described, weighted for when the video was requested and how many “up-votes” each video has (videos requested multiple times go to the top of the queue), is one of the most important features to understand the needs of the community. All YouDescribe training encourage volunteers to choose videos from the wish list. The maximum number of times that a video is requested is 32. The most requested video category is ‘People & Blogs’ with about 500 requests, followed by 450+ for ‘Entertainment’ and 400+ for ‘Film & Animation’. In addition to adding videos to the wish list, viewers rate the quality of the AD on a one-to-five-star basis (1 being poor, 5 being excellent) and those ratings are matched to the describer’s Google ID and posted publicly alongside of the description. In the case of multiple descriptions for the same YouTube video ID, the highest rated AD is listed first. In addition to the public facing ratings, viewers can select feedback from a list to help new describers with specific improvements on: audio quality, diction, balance of inline and extended tracks, a need for less or more description, voice tone matching the tone of video, description given before action/audio cues, and a lack of description for onscreen text. About 21% of the AD has been rated by viewers, and 79% of videos remain unrated (Fig. 5). Of the videos rated, 59% are excellent, 25% are very good, 8% are good, 5% are fair, and only 3% are rated as poor (change accordingly from the modified figure) (Fig. 6). The quality and utility of the YuWA data set would be improved with a greater number of videos being rated by the BVI community.

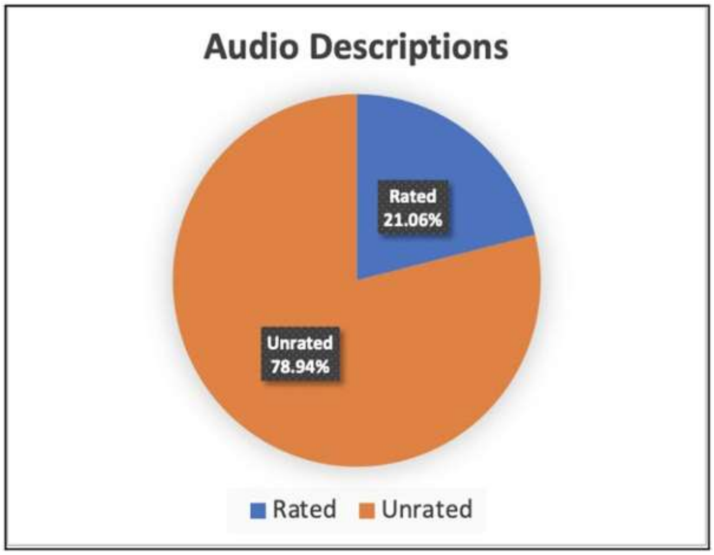

Fig. 5. Audio Descriptions rated and unrated (03/2017-09/2022, Live Dataset).

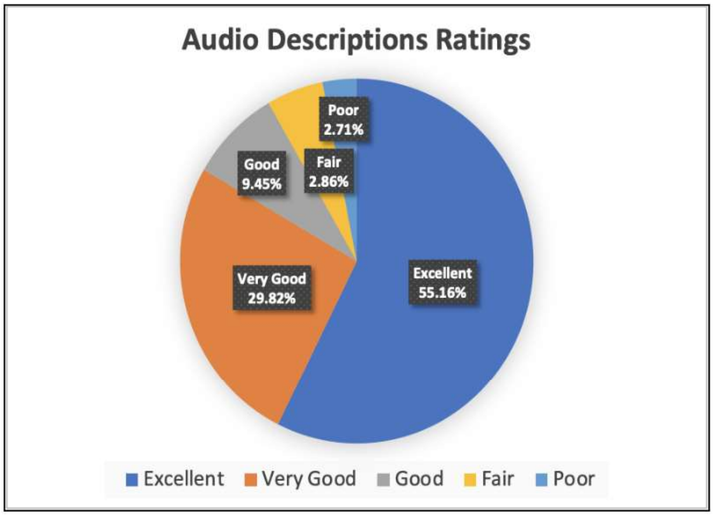

Fig. 6. Audio Descriptions ratings: excellent to poor (03/2017-09/2022, Live Dataset).

As of Sept. 2022, there are 3,000+ describers, 6,400+ audio descriptions with approximately 1,000+ ADs added in 2022, and on average 90+ videos described each month, depicting a general trend of more volunteers providing descriptions each year. The dataset covers a vast domain of videos in 32 titled categories including the 15 most popular: Film & Animation, Music, Autos & Vehicles, Travel & Events, Pets & Animals, Sports, People & Blogs, Gaming, Comedy, Entertainment, How-To & Style, News & Politics, Nonprofits & Activism, Education, and Science & Technology. YuWA has 5,500+ videos with a total duration of about 310 hours and an average video duration of about 5.5 minutes. Currently there are 76,000+ audio clips (about 10 audio clips per 5-minute description with about one audio clip per 30 seconds). The audio clips are transcribed with Listen by Code and Google Cloud's Speech-to-Text API. As speech to text services improve, the data set may be updated to reflect those changes. Similarly, as new YouDescribe AD content is published, additional data may be added to the YuWA database.

The following are some specialty data sets that are of interest to researchers in various fields, highlighted to showcase the possible utility of audio description across disciplines. YuWA includes a specialized dataset for 75+ descriptions of an identical video made for Carnegie Mellon's Human Computer Interaction Institute course. "Students in Special Topics: Accessibility: A Guide to Building Future User Interfaces" have uploaded the video to their own YouTube Channel (each video has the same content but a different YouTube ID), and described an employment graph. The shortness of the video (21 seconds), along with the 75 aggregate descriptions collected over 5 years has a number of possible study applications in audio description research, comparative linguistic studies, or to test against computer generated descriptions with multiple trainers. While other videos at YouDescribe have multiple descriptions, they are listed in order of quality rating, and attached to a single YouTube ID making them easier to be cataloged.

We also have a premium set of describer data that can be compared to the full data set, or future describer data sets. The audio description in this specialized set was done exclusively by describers aged 13-18, earning community volunteer hours for a Northern California service organization. A 1–2-hour training was supplied by a Smith-Kettlewell sponsored trainer, and club members received personalized and general feedback over the year from trainer Charity Pitcher-Cooper as desired. Describers had additional support from an adult coordinator who was also present for the describer trainings. More experienced student describers (1+ years) supported new describers with their audio descriptions. Many describers from this group continued creating AD after age 18, and their post service club member descriptions are included. Because YouDescribe is used all over the world, in any language, and our volunteers have a variety of different describer backgrounds, this specialized data subset offers a known age range, and audio description training cohort that can be compared to other describers.

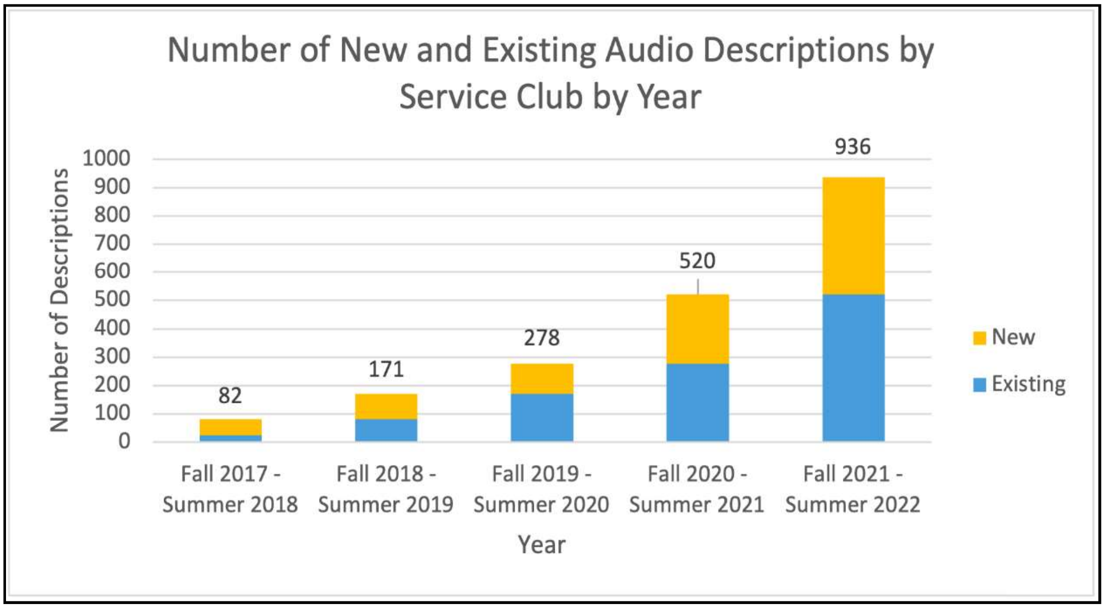

Fig. 7. Audio Descriptions by Service Club (Fall 2017 - Summer 2022, Premium Dataset).

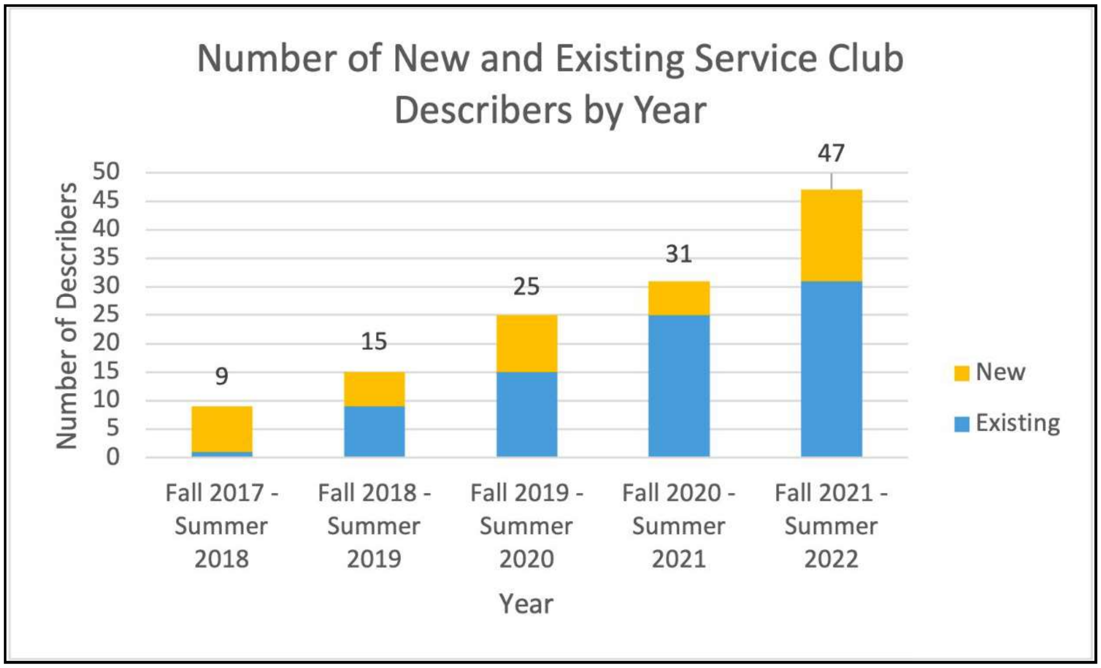

Fig. 8. Service Club Describers (Fall 2017 - Summer 2022, Premium Dataset).

Conclusions

High-quality datasets that can effectively contribute to advancing artificial intelligence (AI) and deep learning (DL) on video understanding, video retrieval, or video-language grounding tasks are increasingly valuable as computation power and algorithms grow in power. The YuWA data set has many possible applications given the large volume of data collected over the past 9 years. For example, The Movie Audio Description (MAD) scalable dataset which focuses on video-language grounding tasks was presented by King Abdullah University of Science and Technology (KAUST) (Soldan et al.). The MAD database and the YuWA dataset share a similar description origin with respect to the crowdsourcing of data. Crowdsourced data is considered “noisy,” as humans make mistakes and recording instruments can sometimes be inaccurate, the data collected has some error bound to it. Noisy data can significantly impact the prediction of any meaningful information. MAD attempted to overcome the hidden biases often found in video-language grounding tasks and increase the generalization capabilities from visual features for long-form video, aiming to have high-quality temporal localization of detecting activities to output beginning and ending timestamps of the language in untrimmed videos. Ultimately, they compared a model trained on the much smaller quality-controlled LSMDC-G training set (which is only 32% the size of the MAD training set), with the same model trained on the full MAD dataset that yielded a relative quality improvement of 20%. As a control, when the MAD model was trained on just a small, 32% sub-section of the MAD training set, it performed worse than the LSMDC-G training set. In this case, MAD researchers showed the success of using a large, diverse, crowdsourced data set, like YuWA, and achieved better results than smaller, less noisy, data sets.

The Visual Geometry Group from the University of Oxford introduced QuerYD, an audio description video dataset, for performing retrieval and event localization in videos based on a subset of YouDescribe data (Oncescu et al.). Oncescu et al's training research illustrated that the content descriptions made with YouDescribe for the BVI community are more relevant than dialogue alone and more detailed than previous description attempts, which can be observed to contain many superficial or uninformative descriptions. The QuerYD data set has been influential for a number of applications and is cited in ten scholarly publications. However, QuerYD focuses on the AD speech to text outputs without the inclusion of metadata on viewers or describers which greatly limits its scope for applications specifically for audio description research, and sociolinguistic applications. In contrast, the YuWA open-source dataset has two audio tracks which include the original audio and high-quality audio descriptions provided by the volunteers without any background disturbance with a precise and direct relationship to video segments. It provides audio narrations which are much more detailed and relevant than the standard narration transcriptions used by other ML training datasets. This is possibly due to the addition of an extended audio clip tool. Due to the application to scholarly and work critical audio descriptions, the YD tool has two kinds of track styles: inline (which is played over the video soundtrack in dialog and soundscape critical pauses) or extended (where the source video is paused, allowing for a longer description to be inserted). Professional AD does not utilize extended track timing in order to keep pace with the video content. Because YouDescribe's purpose includes detail-oriented fan-base observations, as well as incredibly content-critical educational, and work-related AD, describers have the option to record longer, more detailed tracks with the extended play mode. Keeping in line with international audio description standards, describers are urged to use the greatly preferred inline track style as much as possible and extended only when necessary. The large volume of data, plus the dedication of the volunteers, combined with the structure of the tool itself (the addition of an extended audio track) makes the YuWA data set both unique and versatile. This organized, curated, open-source dataset will most certainly be used for AI-based audio description projects.

Currently, a project headed by San Francisco State University Computer science program called YouDescribeX seeks to generate audio descriptions either automatically or semi-automatically by making use of the Human-in-loop Machine Learning (HILML) approach to video description, automating video text generation and scene segmentation and then allowing humans to edit the AI generated output (Bodi, et al). A comparison of YouDescribeX AD to the YuWA dataset has the near-future potential to build a user-friendly AD tool, and provide audio descriptions for a much larger number of YouTube videos than would be possible with human-generated audio descriptions alone.

Moreover, the YuWA dataset has included detailed, anonymized metadata as well as useful summary statistics allowing for potential sociolinguistic study of topics such as describer register (a variety of language used for a particular purpose or in a particular communicative situation) and viewer satisfaction; such metadata could enable the creation of AD-specific word selection guidelines helping both humans and computers find the most descriptive phrases with the shortest number of syllables, combined with impeccable track placement.

The YuWA Dataset described here is hosted in the You Described, We Archived repository at https://github.com/youdescribe-sfsu/You-Described-We-Archived

This work was supported by NEI- R01EY020925, 2010-2014; Department of Education-H327J110005 2011-2013; NIDLIRR- RERC on Blindness and Low Vision- 90RE5024 2016-2022 and 90REGE0018 2022-2027; Ability Central (formerly The Disabilities Communication Fund)- Improving Access to Described Video 2016-2017, Training the Next Generation of Audio Describers 2017-2018, and Machine Learning Enhanced Video Accessibility for the Blind 2020-2022.

Works Cited

"30 Eye-Opening YouTube Facts, Figures and Statistics You Should Know in 2022." *Cloud Income* 17 Apr. 2020, https://cloudincome.com/youtube-statistics/#youtubestats.

"53 Web Accessibility Statistics [Updated for 2022]." *Don't Do It Yourself*, 24 Mar. 2022, https://ddiy.co/web-accessibility-statistics/.

"Accessibility Statistics | Interactive Accessibility." *Interactive Accessibility*, www.interactiveaccessibility.com/accessibility-statistics.

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

"Audio Description Basics for Beginners: A Do's and Don'ts Video Tutorial. - *YouTube.*" www.youtube.com, www.youtube.com/playlist?list=PLNJrbI_nyy9uzywoJfyDRoeKA1SaIEFJ7.

Bodi, Aditya, et al. "Automated Video Description for Blind and Low Vision Users." *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 8 May 2021, 10.1145/3411763.3451810.

Carnegie Mellon University, Human Computer Interaction Institute. "Accessibility: A Guide to Building Future User Interfaces" HCI Undergraduate: 05-499, HCI Graduate: 05-899, 2017-2022. https://www.hcii.cmu.edu/course/special-topics-accessibility-guide-building-future-user-interfaces

"Lecciónes de Descripción - *YouTube*." www.youtube.com, www.youtube.com/playlist?list=PLNJrbI_nyy9sZmflPvSFv_WwuNVivLyBr. Accessed 6 Oct. 2022.

"Most Popular Web Browsers in 2022 [Sep '22 Update] | *Oberlo*." www.oberlo.com, www.oberlo.com/statistics/browser-market-share#:~:text=As%20of%20August%202022%2C%20Google.

Oncescu, Andreea-Maria, et al. "QUERYD: A Video Dataset with High-Quality Text and Audio Narrations." ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6 June 2021, 10.1109/icassp39728.2021.9414640.

Packer, Jaclyn, et al. "An Overview of Video Description: History, Benefits, and Guidelines." *Journal of Visual Impairment & Blindness*, vol. 109, no. 2, Mar. 2015, pp. 83–93, 10.1177/0145482x1510900204.

Pitcher-Cooper, Charity, and John Brabyn. "The Unstoppable Utility of YouDescribe." Wireless RERC 2021 State of Technology Forum Proceedings, 30 Sept. 2021, https://www.ski.org/unstoppable-utility-youdescribe-wireless-rerc-2021-state-technology-forum-proceedings.

Soldan, Mattia, et al. "MAD: A Scalable Dataset for Language Grounding in Videos from Movie Audio Descriptions." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, 10.1109/cvpr52688.2022.00497.

University of California, Berkeley. *Online Course Catalog*, English 165. "The Art of Writing: The Visible Made Verbal" Spring 2019 https://english.berkeley.edu/courses/5931.

University of California, Berkeley. *Online Course Catalog*, English 175. "Literature and Disability." FALL, 2020, Fall 2021 https://english.berkeley.edu/courses/7472.

"What Is YouDescribe? A Video Tutorial - *YouTube*." www.youtube.com, www.youtube.com/playlist?list=PLNJrbI_nyy9sjqZ-Wcn6sX868i9KtdNrT.

World Health Organization. “Blindness and Vision Impairment.” *WHO International, World Health Organization*: WHO, 11 Oct. 2021, www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

Xu, Hu, et al. “VideoCLIP: Contrastive Pre-Training for Zero-Shot Video-Text Understanding.” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, 10.18653/v1/2021.emnlp-main.544.

Xu, Hu, et al. “VLM: Task-Agnostic Video-Language Model Pre-Training for Video Understanding.” Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, 10.18653/v1/2021.findings-acl.370.

“YouCook2: Large-Scale Cooking Video Dataset for Procedure Understanding and Description Generation.” *YouCook2*, https://youcook2.eecs.umich.edu/.

“YouDescribe - Audio Description for YouTube Videos.” *YouDescribe*, https://youdescribe.org/.

“YouDescribe - *YouTube*.” www.youtube.com, www.youtube.com/channel/UCnXTm_yhrDJhUcH9hkwdisw

Zhu, Linchao, and Yi Yang. “ActBERT: Learning Global-Local Video-Text Representations.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020, 10.1109/cvpr42600.2020.00877.