Video Accessibility for the Visually Impaired

Ilmi Yoon [1] **Umang Mathur** [1] **Brenna Gibson Tirumalashetty** [1] **Pooyan Fazli** [1] **Joshua Miele** [2]

Abstract

Video accessibility is crucial for blind and visually impaired individuals for education, employment, and entertainment purposes. YouDescribe is a web-based platform that enables sighted volunteers to add audio descriptions to YouTube videos thus making them accessible to visually impaired users. Creating good descriptions requires much effort, and it is impossible for volunteers to catch up with all the videos published daily. This work builds on top of YouDescribe and facilitates video accessibility by automating the description generation process for online videos and generating well-structured training data to advance the state of the art in video understanding.

1. Introduction

Thousands of videos are uploaded to online video platforms daily. Online videos have become a part of our daily lives. The blind and visually impaired often miss out on the visual information conveyed through videos. The vast majority of online video material is currently not accessible to millions of blind and visually impaired people who would significantly benefit from improved access to videos for education, employment, and entertainment purposes. Video understanding is an emerging field in which models are trained to extract semantic information from video data and annotate or describe for intended users. This work is relevant to about 253 million visually impaired people in the world, of whom 36 million are blind. (Bourne et al., 2017; Fricke et al., 2018).

An effective way to bridge this gap is through adding audio descriptions to videos. Audio description verbalizes the visual components of a video, so that people with visual impairments are able to follow the events on the screen through understanding settings, facial expressions, gestures, on-screen text, costumes, lightning, or any other relevant information that can be described in a narrative audio track synchronized with the video and can be turned on or off as needed. Over the past 5 years, organizations, such as the Media Access Group[1] and American Foundation for the Blind[2] have paved the way to make television programs and feature films accessible to the blind. With the increasing availability of online videos for education and entertainment on YouTube, Hulu, and Vimeo, there is a need to make these contents accessible as well.

The Smith-Kettlewell Eye Research Institute[3] devised a platform known as Descriptive Video Exchange (DVX) that facilitates recording and distributing audio descriptions for online videos. YouDescribe[4] is a web-based client for DVX that enables sighted volunteers to add audio descriptions to YouTube videos. It follows a crowdsourcing model where any sighted volunteer can record and upload audio descriptions for YouTube videos upon request. These audio annotated videos can be shared across the internet thus making them accessible to all. It has created a community of more than 2000 volunteers and many blind or visually impaired users. More than 28000 audio description files have been created since the launch of the website in April 2017. While the current free-form video description has its value, describing video is an open-ended task that can be extremely daunting for inexperienced volunteers who are often limited by their own vocabulary and visual experiences. Inexperienced describers also have difficulty producing informative and thorough descriptions. The quality of descriptions can vary significantly depending on the volunteers, and the fact that good descriptions require high quality creative writing often discourages lots of good-hearted volunteers from getting started. Furthermore, it is not feasible for volunteers to keep up with the rate of videos published online.

Video accessibility is still in its infancy, and one of the biggest challenges for advancing the field is the lack of well-structured training data. Labeling videos for training models is complex, and there is no clear standard to do so (Abu-El-Haija et al., 2016). In addition to providing a service for video accessibility, generating video descriptions that are

[1]Department of Computer Science, San Francisco State University [2]Amazon Lab126. Correspondence to: Ilmi Yoon <ilmi@sfsu.edu>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

[1] http://access.wgbh.org/

[2] https://www.afb.org/

[3] https://www.ski.org/

[4] https://www.youdescribe.org

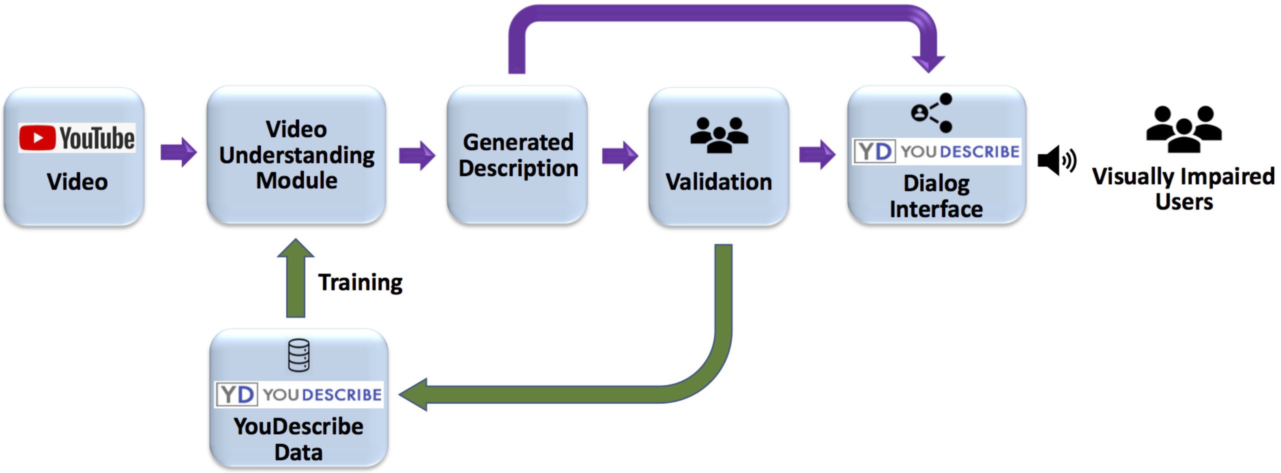


Figure 1. The workflow of the proposed framework

moderated by sighted volunteers will help build the corpus of labeled video data.

In summary, this work addresses the following major issues:

- Enhancing video accessibility for blind and visually impaired individuals.
- Generating well-structured training data to advance the state of the art in video understanding.

2. Proposed Framework

We propose a framework (Figure 1) where video understanding assist human volunteers to produce high-quality audio descriptions to YouTube videos. The goal is to enhance video accessibility for blind and visually impaired individuals. To implement a prototype of the framework, we use publicly available APIs, such as Microsoft Azure Cognitive Services[5] and IBM scene classifier model[6] (Zhou et al., 2018).

Through interviews with a representative group of end users, we have designed a keep-minimal-and-rest-on-demand description generation system that keeps the *baseline descriptions* minimal mainly focusing on background and primary foreground objects and people in the scene, while *on-demand descriptions* are available when user pauses the video and asks questions. The questions can be about the location of the scene (e.g., residential neighborhood, New York City, etc.), people in the scene (e.g., facial expression, gender, etc.), and other relevant information. The baseline descriptions will be synchronized with YouTube videos while on-demand descriptions will be handled through the dialogue interface.

We have designed a user interface to be used on YouDescribe (Figure 2). This is similar to the existing YouDescribe volunteer interface where users are taking notes on the right side panel while watching the video for the first time. In the new interface, sighted users will see scene segments, baseline descriptions, and on-demand descriptions and can make their own notes given the provided descriptions. Compared to free-form descriptions, this will help to structure the descriptions and improve their completeness. With the added structure and suggested content, the task of describing a video becomes much easier and can be scaled to substantially more videos on YouTube.

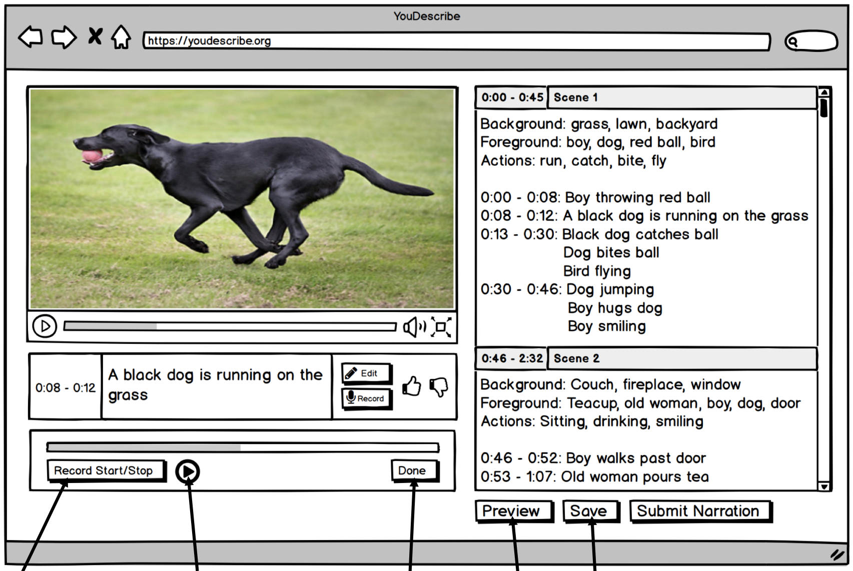
As sighted volunteers can not catch up with all the available videos on YouTube, visually impaired users wish to be able to access non-validated descriptions with confidence scores as well as an interface to tag unlikely descriptions.

The workflow of the framework is described below:

1. **Input Data:** Videos for which descriptions have been requested are forwarded to the model for further processing.
2. **Scene Segmentation:** For effective description generation, we segment the video into a sequence of scenes.
3. **Key Frame Extraction:** As each scene segment has a different time span, key frames are sampled to maintain the right amount of granularity of input data for the model being trained.
4. **Caption Generation:** Each key frame is processed by the model to generate captions which best describe the

[5] https://azure.microsoft.com/en-us/services/cognitive-services/directory/vision/

[6] https://developer.ibm.com/exchanges/models/all/max-scene-classifier/



0:08 - 0:12 A black dog is running on the grass

0:00 - 0:45 Scene 1

Background: grass, lawn, backyard
Foreground: boy, dog, red ball, bird
Actions: run, catch, bite, fly

0:00 - 0:08: Boy throwing red ball
0:08 - 0:12: A black dog is running on the grass
0:13 - 0:30: Black dog catches ball
Dog bites ball
Bird flying
0:30 - 0:46: Dog jumping
Boy hugs dog
Boy smiling

0:46 - 2:32 Scene 2

Background: Couch, fireplace, window
Foreground: Teacup, old woman, boy, dog, door
Actions: Sitting, drinking, smiling

0:46 - 0:52: Boy walks past door
0:53 - 1:07: Old woman pours tea

Click to begin and stop voice recording

Voice recording playback

Finish and accept recorded phase

Watch video with narration

Save to work on later

Figure 2. Proposed interface for the sighted volunteers. On the right side, the relevant objects and suggested caption per scene are generated by the model while the text box under the video displays the text synchronized with the video at display. User will edit, and the changes will be reflected on the right side panel.

video at that instance. It also detects any text in the key frame, people with ID (so reappearing person will be handled properly), gender, emotion, hair color, age, objects with their bounding boxes, and environment category. People are recognized if they are known celebrities while others are labeled as unknown. We will identify unknown people from the video dialogues and provide correct names. We also plan to train a CNN + LSTM architecture to take a key frame as input and output a caption (Xu et al., 2015; Venugopalan et al., 2015).

5. **Summarizing the Descriptions**: Text summarization is used to combine the descriptions of all key frames into a single coherent summary of the entire scene.

6. **Revising or Validating the Descriptions**: Through the interface, volunteers revise or validate the model-generated descriptions.

7. **Retraining the Model**: The discrepancy between the model-generated and revised narrations shall be recorded. The revised versions shall be used as inputs to retrain and improve the accuracy of the model.

8. **Dialogue Interface**: Through the dialogue interface, baseline and on-demand descriptions are presented to blind and visually impaired users. Users can also tag unlikely descriptions for further investigation by the sighted volunteers.

Figure 3 shows two sample videos along with the extracted scenes and key frames, generated captions, their confidence scores, detected objects, people, faces, and locations in the scene.

3. Dataset

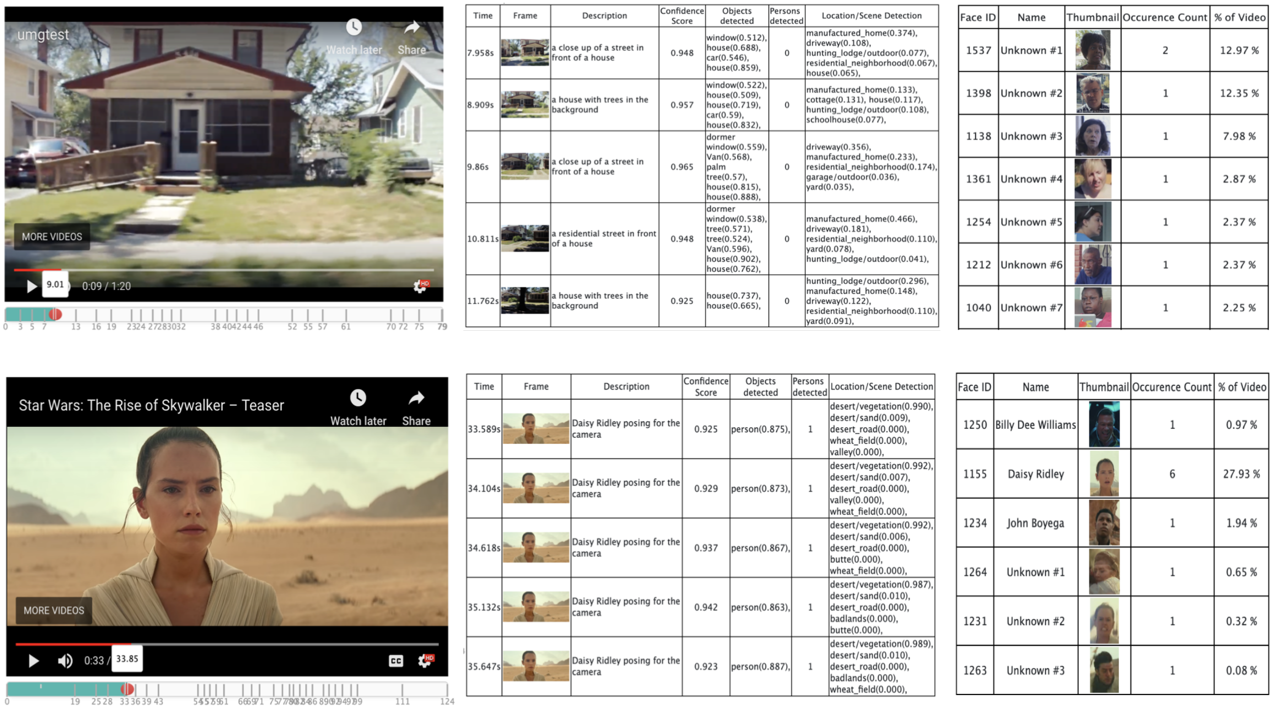
There are 28000+ audio files on YouDescribe.org as of January 2019. Some are in high quality (with user rating of 4 or 5) and some are in low quality (with user rating of 1-3).

4. Challenges

Comparing machine learning-generated audio and text content and human-generated content in their semantics is a challenging task in natural language processing (Anderson et al., 2016; Vedantam et al., 2015; Lavie & Agarwal, 2007; Lin, 2004). It usually works well on narrow contexts, for instance, sports news or product advertisement, but video accessibility is not limited to one context. Therefore, the model will take a longer time to train and need a large amount of data to perform well on every context that YouTube videos cover.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv:*,



Time	Frame	Description	Confidence Score	Objects detected	Persons detected	Location/Scene Detection
7.958s		a close up of a street in front of a house	0.948	window(0.512), house(0.688), car(0.546), house(0.859),	0	manufactured_home(0.374), driveway(0.108), hunting_lodge/outdoor(0.077), residential_neighborhood(0.067), house(0.065),
8.909s		a house with trees in the background	0.957	window(0.522), house(0.509), house(0.719), car(0.59), house(0.832),	0	manufactured_home(0.133), cottage(0.131), house(0.117), hunting_lodge/outdoor(0.108), schoolhouse(0.077),
9.86s		a close up of a street in front of a house	0.965	dormer window(0.559), Van(0.568), palm tree(0.57), house(0.815), house(0.888),	0	driveway(0.356), manufactured_home(0.233), residential_neighborhood(0.174), garage/outdoor(0.036), yard(0.035),
10.811s		a residential street in front of a house	0.948	dormer window(0.538), tree(0.571), tree(0.524), Van(0.596), house(0.902), house(0.762),	0	manufactured_home(0.466), driveway(0.181), residential_neighborhood(0.110), yard(0.078), hunting_lodge/outdoor(0.041),
11.762s		a house with trees in the background	0.925	house(0.737), house(0.665),	0	hunting_lodge/outdoor(0.296), manufactured_home(0.148), driveway(0.122), residential_neighborhood(0.110), yard(0.091),

Face ID	Name	Thumbnail	Occurence Count	% of Video
1537	Unknown #1		2	12.97 %
1398	Unknown #2		1	12.35 %
1138	Unknown #3		1	7.98 %
1361	Unknown #4		1	2.87 %
1254	Unknown #5		1	2.37 %
1212	Unknown #6		1	2.37 %
1040	Unknown #7		1	2.25 %

Time	Frame	Description	Confidence Score	Objects detected	Persons detected	Location/Scene Detection
33.589s		Daisy Ridley posing for the camera	0.925	person(0.875),	1	desert/vegetation(0.990), desert/sand(0.009), desert_road(0.000), wheat_field(0.000), valley(0.000),
34.104s		Daisy Ridley posing for the camera	0.929	person(0.873),	1	desert/vegetation(0.992), desert/sand(0.007), desert_road(0.000), valley(0.000), wheat_field(0.000),
34.618s		Daisy Ridley posing for the camera	0.937	person(0.867),	1	desert/vegetation(0.992), desert/sand(0.006), desert_road(0.000), butte(0.000), wheat_field(0.000),
35.132s		Daisy Ridley posing for the camera	0.942	person(0.863),	1	desert/vegetation(0.987), desert/sand(0.010), desert_road(0.000), badlands(0.000), butte(0.000),
35.647s		Daisy Ridley posing for the camera	0.923	person(0.887),	1	desert/vegetation(0.989), desert/sand(0.010), desert_road(0.000), badlands(0.000), wheat_field(0.000),

Face ID	Name	Thumbnail	Occurence Count	% of Video
1250	Billy Dee Williams		1	0.97 %
1155	Daisy Ridley		6	27.93 %
1234	John Boyega		1	1.94 %
1264	Unknown #1		1	0.65 %
1231	Unknown #2		1	0.32 %
1263	Unknown #3		1	0.08 %

Figure 3. Videos are segmented to tens of short scenes, where the scene segmentation is displayed right below the video. The tables on the right display the captions of key frames, their confidence scores, detected objects, people, faces, and locations in the scene.

1609.08675 [cs.CV], 2016.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the 14th European Conference on Computer Vision, ECCV*, pp. 382–398, 2016.

Bourne, R. R. A., Flaxman, S., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., Kempen, J. H., Leasher, J. L., Hans Limburg, K. N., Pesudovs, K., Resnikoff, S., Silvester, A., Stevens, G. A., Tahhan, N., Wong, T. Y., and Taylor, H. R. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review and meta-analysis. *Lancet Global Health*, 5 (9):e888–e897, 2017.

Fricke, T. R., Tahhan, N., Resnikoff, S., Papas, E., Burnett, A., Ho, S. M., Naduvilath, T., and Naidoo, K. S. Global prevalence of presbyopia and vision impairment from uncorrected presbyopia: Systematic review, meta-analysis, and modelling. *Ophthalmology*, 125(10):1492–1499, 2018.

Lavie, A. and Agarwal, A. METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, StatMT*, pp. 228–231, 2007.

Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the ACL workshop on Text Summarization Branches Out*, 2004.

Vedantam, R., Zitnick, C. L., and Parikh, D. CIDEr: Consensus-based image description evaluation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 4566–4575, 2015.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. Sequence to sequence - video to text. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV*, pp. 4534–4542, 2015.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning, ICML*, pp. 2048–2057, 2015.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.