Video Accessibility for the Visually Impaired

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

[1]Department of Computer Science
San Francisco State University

[2] YouDescribe.org
Amazon.com Inc

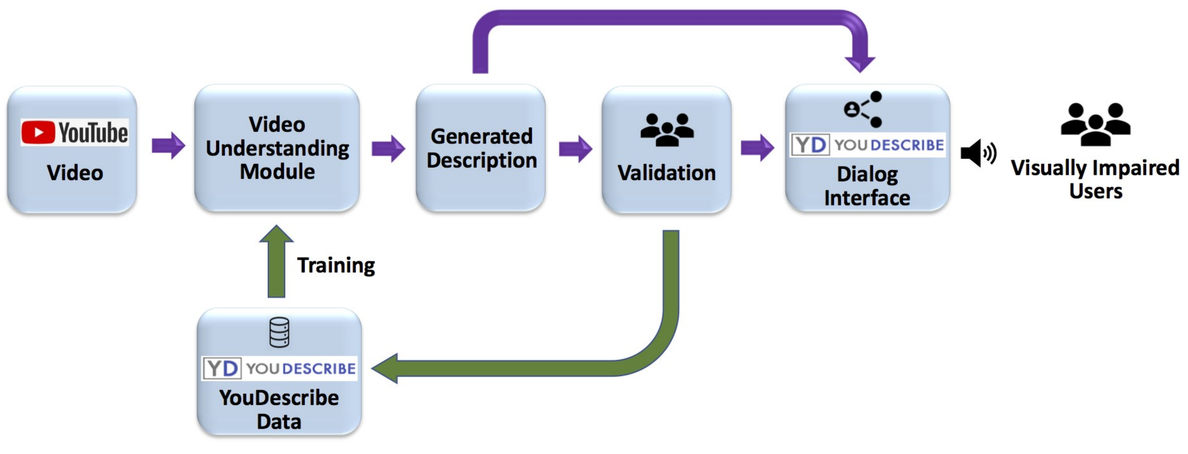
Abstract

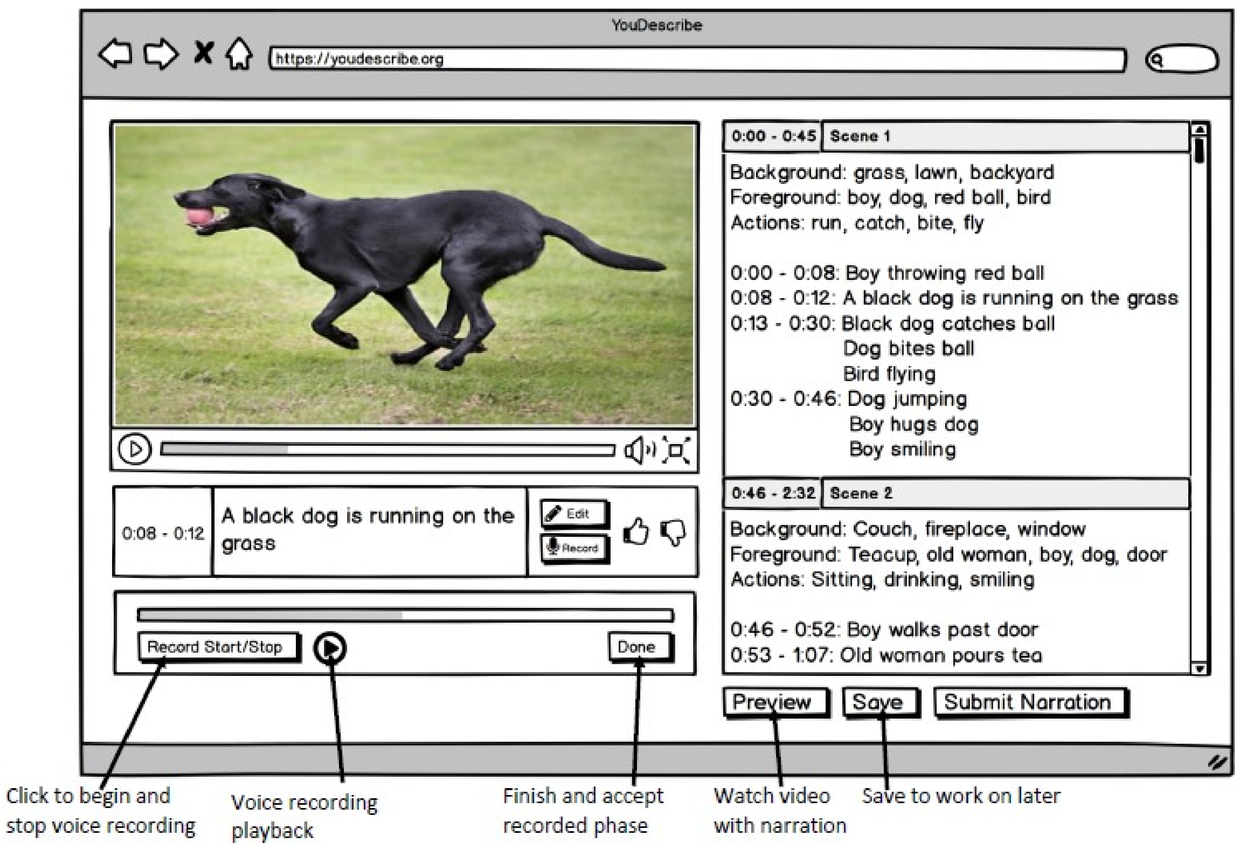
Video accessibility is crucial for blind and visually impaired individuals for education, employment, and entertainment purposes. YouDescribe is a web-based platform that enables sighted volunteers to add audio descriptions to YouTube videos thus making them accessible to visually impaired users. Creating 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.

Proposed Workflow & Architecture

The workflow of the framework is described below:

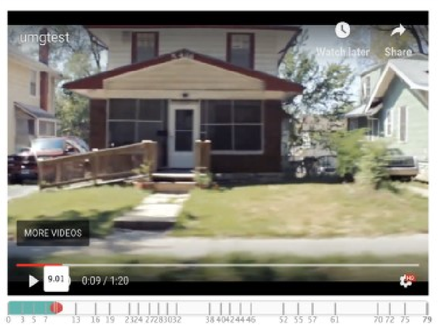
1. Input Data: Videos for which descriptions have been requested are forwarded to the model for further processing.
2. Scene Segmentation: For effective description generation, we segment the video into a sequence of scenes.
3. Key Frame Extraction: As each scene segment has a different time span, key 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 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.





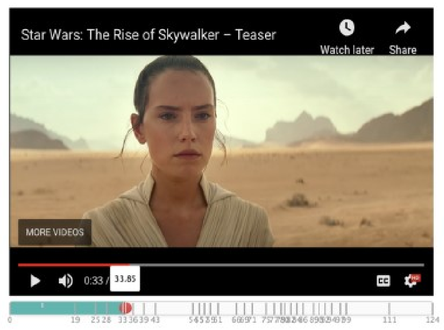
Current Implementation

Information extracted from video/frame as input to Video Understanding Model



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 %

Dialog Interface for on-demand description

Intent name	Training Sentences	Possible Output
fetch_number_of_persons	• how many people can be seen in the current frame of the video ? • how many persons are visible in the video at this instant ? • can you tell me the number of people on the screen at this instant ? • how many humans are in this scene ? • what is the total count of people on the screen ? • give me the number of persons on the screen right now. • give me the number of persons on the screen right now. • how many persons are visible on the screen right now ? • tell me the number of people in the scene. • how many people are there in the scene ?	**Case 1:** Confidence score for all 5 'person' objects is greater than 0.8 **Output:** The scene contains 5 people **Case 2:** Confidence score for all 3 'person' objects is greater than 0.8 and for 2 'person' objects it is between 0.5 to 0.8 **Output:** I'm not sure but The scene contains a group of 3 to 5 people. . . **Case n:** **Output:**
surrounding_detection	• what do the surroundings look like in the current scene • are the characters indoor or outdoor • describe the surroundings • describe the environment • describe the location	**Case 1:** driveway (0.67), residential_neighborhood(0.24), garden(0.09) **Output:** It is an outdoor scene and the background seems like a driveway but could also be a residential_neighborhood.

Dataset and Plan for User Study

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).

User study is planned on "Documentary style" and "How-to-video style" for baseline description and on-demand description generation that will be edited by sighted volunteers and will be tested by blind users.