Non-Visual Accessibility Assessment of Videos

Ali Selman Aydin
Stony Brook University
Stony Brook, NY, United States
aaydin@cs.stonybrook.edu

Yu-Jung Ko
Stony Brook University
Stony Brook, NY, United States
yujko@cs.stonybrook.edu

Utku Uckun
Stony Brook University
Stony Brook, NY, United States
uuckun@cs.stonybrook.edu

IV Ramakrishnan
Stony Brook University
Stony Brook, NY, United States
ram@cs.stonybrook.edu

Vikas Ashok
Old Dominion University
Norfolk, VA, United States
vganjigu@odu.edu

ABSTRACT

Video accessibility is crucial for blind screen-reader users as online videos are increasingly playing an essential role in education, employment, and entertainment. While there exist quite a few techniques and guidelines that focus on creating accessible videos, there is a dearth of research that attempts to characterize the accessibility of existing videos. Therefore in this paper, we define and investigate a diverse set of video and audio-based accessibility features in an effort to characterize accessible and inaccessible videos. As a ground truth for our investigation, we built a custom dataset of 600 videos, in which each video was assigned an accessibility *score* based on the number of its *wins* in a Swiss-system tournament, where human annotators performed pairwise accessibility comparisons of videos. In contrast to existing accessibility research where the assessments are typically done by blind users, we recruited sighted users for our effort, since videos comprise a special case where sight could be required to better judge if any particular scene in a video is presently accessible or not. Subsequently, by examining the extent of association between the accessibility features and the accessibility scores, we could determine the features that significantly (positively or negatively) impact video accessibility and therefore serve as good indicators for assessing the accessibility of videos. Using the custom dataset, we also trained machine learning models that leveraged our handcrafted features to either classify an arbitrary video as accessible/inaccessible or predict an accessibility score for the video. Evaluation of our models yielded an F_1 score of 0.675 for binary classification and a mean absolute error of 0.53 for score prediction, thereby demonstrating their potential in video accessibility assessment while also illuminating their current limitations and the need for further research in this area.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility theory, concepts and paradigms**; *Accessibility systems and tools.*

KEYWORDS

video accessibility;non-visual accessibility

ACM Reference Format:
Ali Selman Aydin, Yu-Jung Ko, Utku Uckun, IV Ramakrishnan, and Vikas Ashok. 2021. Non-Visual Accessibility Assessment of Videos. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482457

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482457

1 INTRODUCTION

Videos are increasingly becoming a first-choice medium to share information in several domains including education, news, and social media. Websites such as Youtube allow users all over the world to create, share, and consume videos of different kinds such as how-to videos, tutorials, lectures, highlights, events, and even presentations. For instance, more than 500 hours of content gets uploaded every minute on the Youtube website which has in excess of 2 billion monthly users as of 2021 [40].

Video format is inherently multi modal, where the information is conveyed to a user via a combination of both visual content and complementing audio. As the default audio present in a video typically by itself cannot provide the full information contained in that video, it needs to be extended to cover the information conveyed by the visual content as well in order to make the video accessible to blind users who can only listen to the video content. In this regard, prior works exist that either provide accessibility guidelines for video content creators[38, 39], or propose automated methods for creating video descriptions[4].

However in practice, the guidelines are rarely followed as the whole process requires significant manual effort, thereby making it expensive, time consuming, and selective. Even the automated methods have not yet achieved mainstream acceptance. As a consequence, videos found on websites vary significantly in how accessible they are to blind screen-reader users. At one extreme, there exist videos where all information is conveyed visually, e.g., a video showing a nature scene with no sound, or with a background music would likely convey no information to the blind user, and in the other extreme, there are videos where the audio covers all the necessary information in the videos, e.g., a narrator accurately describing a nature scene would make it possible for the video to be followed by blind users. Most videos however exhibit accessibility between these two extremities where the audio partially covers the information present in the videos.

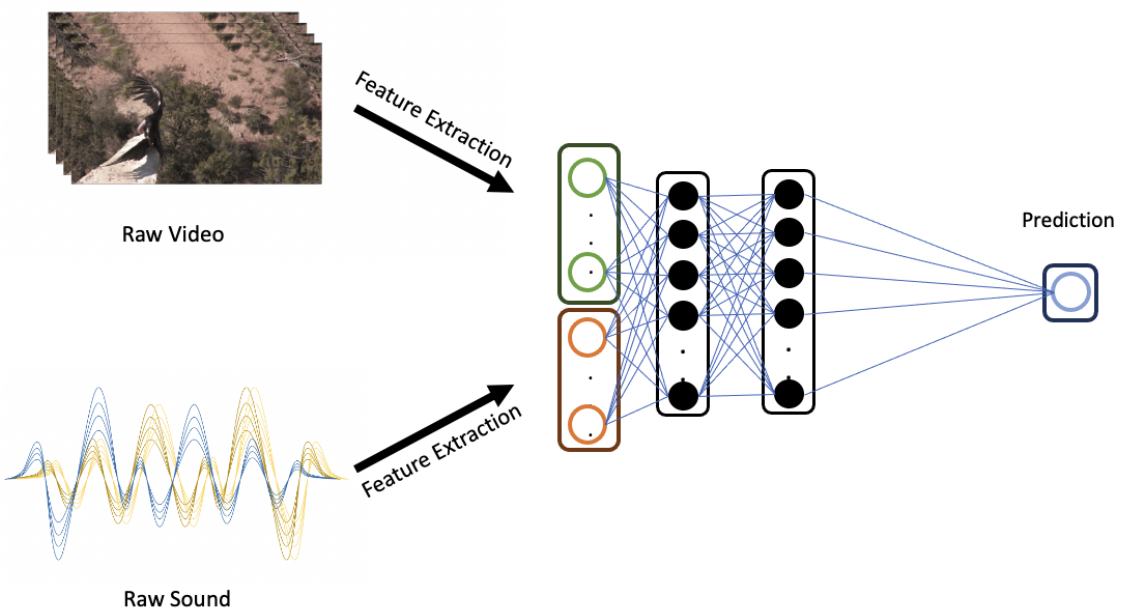


Figure 1: Accessibility analysis using handcrafted features. From left to right: (i) Two main sources of information, namely the video and the audio. (ii) Handcrafted feature computation. (iii) Use of features for predicting accessibility scores.

If the accessibility of a video can be expressed in some quantified form (e.g., a score), the blind users can then use this cue to compare and select more accessible videos to watch among the several alternatives available in search results. Without this cue, they have to follow a tedious and frustrating trial-and-error approach where they have to test each video in the search results by listening to a portion of it before deciding whether to continue listening or move on to test the next search result video [27]. Therefore in this paper, we explore a statistical approach for quantifying the degree of accessibility exhibited by any arbitrary video. Specifically, we explore the following research challenge: *Can we quantify the degree of accessibility of a video in the form of a rating or score, and then explain or justify this rating?*

In this regard, we first constructed a dataset comprising subjective accessibility evaluations of 600 videos, where multiple sighted raters evaluated the accessibility of each video via a Swiss-system tournament [17] thereby resulting in a final accessibility score (i.e., number of *wins*) for each video at the end of the tournament. Leveraging this dataset, we then investigated a diverse assortment of handcrafted visual and audio features with regard to the strength of their associations with the accessibility scores, i.e., what features better correlate with high/low accessible scores. Using these custom features, we also trained: (i) a binary classifier that can predict whether a video is accessible or not with a performance of 0.675 F_1 score; and (ii) a neural network based prediction model (see Figure 1) that can provide reasonable estimates (mean absolute error of 0.53) of the accessibility score of unseen videos in the dataset. The use of custom handcrafted features in the models facilitates explainable predictions, i.e., the features can be used to justify the predicted scores or the assigned accessible/inaccessible labels.

Our contributions are as follows:

- We collect annotations for a dataset for quantifying video accessibility that consists of subjective accessibility evaluations of 600 short videos
- We perform statistical analysis of associations between accessibility of a video and a set of handcrafted visual and audio features that represent the video, both similar to the ones existing in the literature, and new ones
- We design and perform evaluation of predictive models that can either classify a video as accessible/inaccessible or generate accessibility scores/labels automatically for the video.

2 RELATED WORK

Our contributions in this paper closely relate to the following extant literature: (i) general accessibility evaluation frameworks; (ii) video accessibility; and (iii) multimedia feature extraction.

2.1 Accessibility Evaluation and Diagnostics

Evaluating accessibility of software, tools and websites is immensely beneficial for both users and developers, therefore there exist plenty of works that facilitate such evaluations[11, 35]. However, many of these methods focus predominantly on assessing the accessibility of websites. For instance, [19] proposed two metrics to evaluate webpage accessibility. The first metric attempts to quantify *navigability* by considering factors such as estimated time it takes to navigate to page sections of interest, use of headings in HTML source, and the accessibility of links. The second metric attempts to quantify *listenability* aspect by considering factors such as existence of alternative ('ALT') text, and repetition of content. On the other hand Gonzalez et al. [21] proposed a system named KAI for not only measuring the accessibility of webpages for people

with visual impairments, but also producing an accessibility report for different sections of a webpage. A comparative study of seven different accessibility metrics is presented in [37]. Other than the work by Asakawa et al. [8] that focuses on accessibility of online Flash content, all other aforementioned techniques, to the best of our knowledge, do not focus on evaluating video content.

Several automated accessibility diagnostics tools or accessibility checkers also presently exist that can analyze an arbitrary webpage or a PDF document, and subsequently generate a detailed accessibility report highlighting the issues that need to be fixed by the web developers of that page. For example, Darvishy et al. [13] and Doblies et al. [14] both proposed a tool for diagnosing and correcting accessibility problems in PDF documents. On the other hand, WAVE (Web Accessibility Evaluation Tool) [2] can pinpoint accessibility issues with webpages. In the context of mobile application development, GSCXScanner for iOS [1] and Lint in Android Studio [3] can assist the developers in evaluating the code structures and then improve the accessibility of their applications.

2.2 Video Accessibility

Prior work on video accessibility exists mainly in the form of guidelines for creating accessible videos. One such example is the W3C guidelines on video accessibility [39]. W3C guidelines outline considerations for accessibility for video creators not only for people with visual impairments, but also people with hearing difficulties. These guidelines suggest that the videos should contain audio descriptions when there is visual content that is essential to convey the meaning of the video [38]. It also introduces various video description methods and ways of creating video descriptions [38].

In addition to creation guidelines, there also exist prior works that focus on improving the accessibility of existing videos. For example, Yuksel et al. [41] present an approach based on creating video descriptions for improving the video accessibility for both people who are blind and those with low-vision. Better utilization of the audio modality via annotations has also been proposed to improve video accessibility [15]. A more recent work [9] describes a method that leverages the concept of visual saliency as a guiding signal for detecting the important regions in the video and then selecting magnifying these regions for improved low-vision interaction with the video.

Compared to the sizeable literature on improving video accessibility, research on video accessibility assessment and evaluation remains an under-studied topic. As an example work in this direction, the work of Acosta et al. [5] concerns accessibility of educational videos produced by universities. Their manual analysis revealed widespread accessibility issues of these videos. While the analysis performed here is manual, as opposed to our goal of investigating the possibility of an automated system, this work identifies the nature and the scope of many video accessibility problems.

Perhaps the closest research related to our work in this paper is by Liu et al. [27] who explore the same problem, but differ from us with regard to the methods and formulations. First, their work is a macro-level assessment that focuses on entire videos, similar to the ones that could be found on online video platforms, whereas our work is more fine-grained in that we focus on individual scenes in a video or short videos. Second, the core heuristics used in their analysis was determined from the findings of a user study with blind participants, where these participants assessed the accessibility of different videos. In contrast, we recruited sighted users for our analysis, since videos comprise a special case where sight is potentially more suited to judge if any particular scene in a video is presently accessible or not; without the benefit of visual confirmation, blind users are likely to miss several inaccessible parts of a video during accessibility assessment. Nonetheless, some of their findings do seem comparable with our observations, and we report these details later in Section 3.2 and Section 4.3.

2.3 Feature Extraction

Automatically computing the handcrafted features used in our analysis and models requires the use of some state-of-the-art techniques in computer vision and audio analysis. We discuss some of the techniques relevant to our work next.

2.3.1 Object Detection. One way to understand the factors that impact the accessibility of a video is by handcrafting different visual and audio features that represent the video and then determining how these features correlate with the accessibility (score/rating) of the video. Object detection, which has witnessed significant improvement in recent years due to advent of deep learning techniques, is one of the main sources of information for computing the visual features that represent a video. Object detection has been a well-studied topic for the past decade [42]. YOLO is one of the widely used object detection frameworks [10, 32–34], so we leverage this framework to compute many of the visual features (see Table 1) in our work. We also leverage a cloud-based service as a secondary source[6], as detailed in Section 3.

2.3.2 Audio Event Detection. Audio event detection is a vital part of our system pipeline for understanding the relationship between the video and audio modalities. Stowell et al. [36] provide a survey of pre-deep learning era work on audio event detection and datasets. Deep learning advancement has led to increased use of neural network models for audio event detection, typically using convolutional layers. In our work, we leverage a recent model proposed by Kong et al. [25] to extract audio events from videos.

3 DESIGN

Our accessibility analysis and the proposed rating system leverage existing techniques in computer vision, audio analysis and natural language processing. Our system focuses on short scenes instead of full-length videos, as scene level analysis captures detailed fine-grained information, that could be then generalized in a bottom-up manner to cover the entire duration of a video.

3.1 Techniques for Video Analysis

3.1.1 Object Detection. One of the main components of our video analysis toolkit is object detection. Since object detection is a well-studied problem in computer vision, there are many publicly available solutions that we can leverage in our system, even in the form of cloud-based services. Our object detection pipeline processes videos frame-by-frame. Specifically, let a video V be a collection of individual frames, $V = I_0, I_1, ..., I_N$ where N is the number frames. The object detector generates object proposals for each frame in

the video, resulting in zero or more bounding box coordinates, with each bounding box having an associated object class label. For our analysis, this predicted object class label is of more interest than the bounding box coordinates. In our analysis, we used two existing tools for object detection: 1) YOLO [34] object detector, trained on COCO dataset [26], and 2) Amazon Rekognition, a cloud-based image and video labelling service [6]. We used the object detection information from these sources to compute features that capture the nature and variety of objects that appear in a video, which we explain later in Section 3.2.

3.1.2 Audio Event Detection. Detection of audio events in a video facilitates a better understanding of the video. For example, if the only audio event in a documentary scene is music, the non-visual accessibility of the video will likely be very low, since all information in the scene is visual. On the other hand, the presence of different speech events in a video could enhance the accessibility of a video as the speech could contain cues about the visual information in the video. Having an understanding of audio events requires the use of an audio event detection model. In our work, we use the model proposed in [25]. This model generates audio class predictions for each time point of the video. Also, the model was trained on [20] dataset, which supports 527 output classes, thereby enabling a detailed analysis of the audio events in the videos. We utilize the model predictions for computing audio features such as event types that involve manually determined meta-class labels (*i.e.,* audio events pointing to a person, such as speech) and also features that provide cumulative descriptive statistics such as the total number of audio events belonging to different classes.

3.1.3 Optical Flow. Optical flow captures the nature of movement in a video, and it has found use in numerous applications[18]. We utilized optical flow to quantify the total amount of movement in a video. Specifically, we computed optical flow maps for each frame in a video using Farneback method [16], which allowed us to derive features related to the extent of movement in the video.

3.1.4 Transcription. Although finding associations between the different classes may potentially provide us with a general understanding of the co-occurrences of the audio and object classes, the information provided by these associations does not often provide a complete picture. For example, it is a very common occurrence that a video contains a narrator who may never actually appear in the video, but provides informative content about the visual content in the video. To better understand the relationship between the speech content and the visual content, we transcribed the videos in our dataset and used features derived from these transcriptions. For transcription task, we used Amazon Transcribe [7], which is an automated service for video and audio transcription.

3.1.5 Text Analysis. Analysis of speech content could reveal details about the relationship between the audio and the visual content. For example, if the detected objects in a video are also described in the speech, it may lead to higher accessibility. To extract features related to speech content, we used Natural Language toolkit – NLTK [28]. We utilized NLTK specifically for part-of-speech tagging, which assigns part-of-speech tags (e.g., singular noun, verb) to all the words in the transcribed speech content.

3.2 Handcrafting Features for Assessing Video Accessibility

To understand the extent to which various visual and audio aspects of a video impact its accessibility, we handcrafted different features (see Table 1) and then examined the correlations between these features and the accessibility scores (obtained from sighted users in a study described later in Section 4.3), so as to uncover the reasons impacting these scores, and also identify potential sources of accessibility issues. In contrast, using embedding features (such as the ones from residual networks[22]) extracted from the visual/sound modalities of a video could result in better model performance for a sufficiently large dataset, but it would be challenging if not impossible to leverage these features for providing justifications or explanations for the model predictions. In sum, an *explainable* model for predicting video accessibility is essential and more useful than a *blackbox* model that only outputs the accessibility ratings.

The design of features shown in Table 1 was based on manual exploratory analysis of the videos in a custom built video dataset (described later in Section 4.1). This set of features provides us the means to analyze the accessibility of videos from various aspects. For example, the positive correlation between the transcript length and the accessibility could suggest ample speech content in a video is more likely to result in more accessible videos. On the other hand, a negative correlation between the number of detected object types and the accessibility score could mean that more object types in a video could imply more entities that need to be mentioned and explained in a video, possibly leading to greater accessibility challenges than having a lower number of object types.

Some of the features described above share similarities with those proposed in a very recent contemporary related work [27]. Specifically, features f_1, f_2, f_3, f_6, f_9 and f_{13} are similar to some of the metrics proposed in that prior work [27], however exact implementation and representation of many of these features differ significantly between the two works. Also notice that some of the features described in Table 1 capture similar information (e.g f_1 and f_4), hence are highly likely to be correlated.

Lastly, note that Table 1 does not include all the features we initially considered as some of these features did not exhibit a strong or significant relationship with the user-generated accessibility ratings, and therefore have been excluded from the table for brevity. Specifically, in addition to the features in Table 1, we had also considered features based on video saliency and motion vector information. However, in a correlation test, we did not find any meaningful relationships between these features and the assigned accessibility scores, and therefore these features were removed from further consideration.

4 EXPERIMENTS

In this section, we describe (i) the dataset we built by selecting videos from two other commonly used datasets for visual saliency prediction[23] and action recognition[29] respectively; (ii) the accessibility evaluation annotation procedure with sighted users; (iii) observations related to our handcrafted features; and (iv) a user

Feature ID	Feature name	Corr.	Significance	Explanation
f_1^*	Speech ratio	0.4217	$p < 0.001$	Speech duration, normalized by the video length
f_2^*	# of object predictions	0.1685	$p < 0.001$	Total number of objects prediction events in the video
f_3^*	# of object prediction types	-0.1042	$p = 0.011$	Total number of different types of object predicted in a video
f_4	Does speech audio event exist?	0.4147	$p < 0.001$	Binary feature describing if speech event detected
f_5	Avg. Sum of optical flow magnitude	-0.1206	$p = 0.003$	Sum of optical flow vectors for each frame, averaged over frames
f_6^*	Transcription number of words	0.5476	$p < 0.001$	Number of words in the transcribed speech of the video
f_7	Does music audio event exist?	-0.1320	$p = 0.001$	Binary feature describing if music event detected
f_8	Music ratio	-0.2229	$p < 0.001$	Music duration, normalized by the video length
f_9^*	Nouns ratio	0.3741	$p < 0.001$	Frequency of the nouns as result of POS tagging
f_{10}	Counts ratio	0.2160	$p < 0.001$	Frequency of the counts as result of POS tagging
f_{11}	Pronouns ratio	0.3629	$p < 0.001$	Frequency of the pronouns as result of POS tagging
f_{12}	# of person detections	0.1918	$p < 0.001$	Number of person detection events
f_{13}^*	Obj. detection-transcription match	0.2823	$p < 0.001$	Number of times a detected object name appears in the transcription
f_{14}	Speech-person coexistence	0.1730	$p < 0.001$	If a video has speech audio event and a person has been detected at the same time

Table 1: Notable features extracted, along with their Spearmans's correlation with the aggregate user-generated accessibility ratings. Notice that some of the metrics proposed in [27] are aimed to capture information similar to the ones captured by the features f_1, f_2, f_3, f_6, f_9, and f_{13} (Highlighted with *), although the exact realizations of these features are different. Also, all features were normalized if applicable.

study with users having visual impairments to understand the relationship between their perceptions of accessibility and the accessibility evaluations previously obtained from sighted users.

4.1 Video Dataset

In order to perform a statistical analysis to determine the associations between handcrafted features and video accessibility, it is imperative to first quantify accessibility over a representative sample of videos. Non-visual accessibility of videos can be highly subjective to quantify, hence we conducted a data-collection study to obtain aggregate ratings of video accessibility.

We first compiled a dataset by sampling videos from LEDOV dataset[23] and AviD dataset[29], which were originally collected for benchmarking video saliency detection and action recognition tasks respectively. The diverse topics of the videos and presence of audio made LEDOV dataset a suitable choice for this task. The videos chosen from this dataset depict a wide range of scenarios including nature scenes, sports/artistic performances, playing instruments, interviews, conversations and other similar settings. To make annotations feasible, we restricted our focus to videos in English language. Furthermore, we removed videos that do not contain any sounds (i.e., no audio channel). We also removed videos longer than 20 seconds to remove outliers and be consistent with other sources. In total, we collected 399 videos from LEDOV dataset.

AviD dataset [29] contains a diverse set of action videos. Since the dataset contains around 450k videos, we randomly sampled videos from this dataset subject to a few constraints. First, we filtered out long videos, and as in case of LEDOV dataset, we focused on videos containing English speech. Sampled videos from AviD dataset belonged to action classes such as playing an instrument, playing sports, outdoor events, and instructional videos. In total, we sampled 201 videos from AviD dataset, which along with 399 videos from LEDOV dataset resulted in a total of exactly 600 videos in our dataset. Overall, the combined dataset consisted of videos that have 10 seconds duration on average (Max: 20 seconds, min: 3 seconds, standard deviation: 3.2 seconds). Majority of the videos in the combined dataset consisted of single scenes, with some videos containing more than one scene with the same theme (e.g., a snowboarding performance shown at different angles).

4.2 Accessibility Annotation of the Dataset

We conducted a user study with 9 sighted participants to obtain accessibility ratings for videos in our dataset. First, the users were introduced to the task of interest, which is answering the following question: *If you were only hearing this video, how well would you understand this video?*. The participants were then introduced to sample videos that belong to both extremes with regard to accessibility (i.e., full narration vs. no narration, e.g., a video that depicts a natural scene and the music played vs. a video where the scene is perfectly described by the narrator in a detailed manner), and video(s) that fall in between these extremes. Next, the users were introduced to the annotation interface, where pairs of videos were shown to user for making comparisons with aforementioned question. The users had to choose between the following three options to make this comparison: (i) first video, meaning first video had higher accessibility, (ii) second video, meaning second video had higher accessibility, or (iii) equal, denoting an inconclusive comparison. Pairwise comparisons have been previously used for similar annotation tasks, such as predicting video interestingness [24], and visual quality assessment [30, 31].

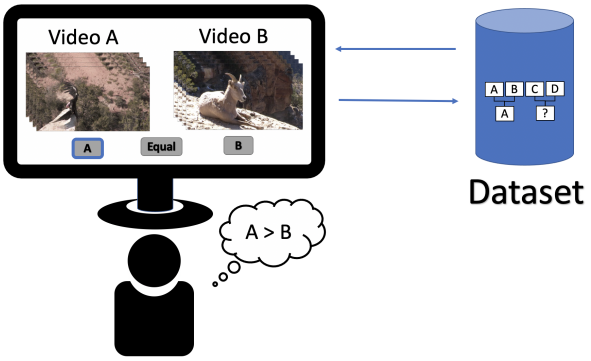


Figure 2: Annotation pipeline. The participants are shown pairs of videos chosen from the dataset, for which they provide one of the three options (A, B or equal). The videos with the same scores are paired together in the next round.

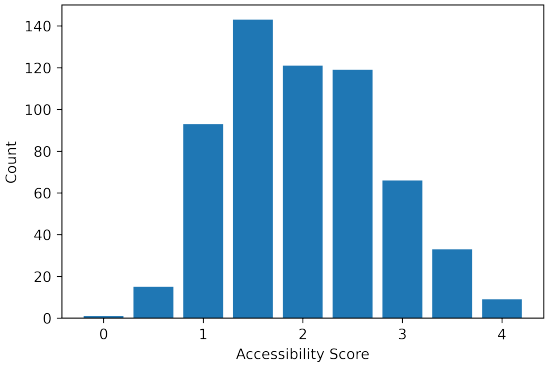


Figure 3: Number of samples with respect to accessibility scores.

One crucial aspect of this annotation process is the cost of annotation with videos. Video annotation is hard to scale even when the video duration is limited, mainly due to the time overhead involved in annotating a large number of videos. Ideally, we would like to minimize the number of comparisons while obtaining reliable ratings. Therefore, for the purposes of our data annotation study, we decided to use the Swiss system [17], which has been widely used in dataset construction for visual quality assessment [30, 31]. Swiss system considers the ranking process as a tournament, where each comparison between pairs of samples is a match. We start with randomly chosen pairs, after which the winner samples are paired together and losing samples are paired together for consecutive rounds. This allows for an approximate ranking to be obtained in significantly lower number of steps compared to $O(n^2)$ approach of comparing all pairs of videos. In our study, we added 1 point to the score of the winning video and 0.5 to each videos for a draw, and compared each video 4 times (except in the cases of bye, where such videos are automatically given a score of 1 for the round), resulting in accessibility scores between 0 and 4, with increments of 0.5. Figure 3 shows the histogram of accessibility scores at the end of the labeling.

4.3 Accessibility Analysis

The correlations between handcrafted features and the accessibility ratings are shown in Table 1. Based on these correlations, below are some of the key relationships that we observed to be important in assessing the accessibility of a video from our dataset.

Object Detection. We found out object detection results are linked to the accessibility ratings in various ways. For instance, we found out that having more object types in a video correlated with a lower corresponding accessibility rating. This could likely be due to the increase in the quantity of visual information that cannot all be explained via audio for ensuring accessibility of the video content. Also, we observed that increased references to the video objects in the transcribed speech positively affected the accessibility ratings, which presumably is due to more visual information being made available via audio to blind screen reader users.

Speech Event-Person Co-Occurrence. We found out that accessibility of a video was higher when a speech-related audio event coincided with the presence of a person (determined using object detection) in the video. This is best explained by the positive correlation value (0.173) between the feature f_{14} and accessibility score as shown in Table 1.

Motion. Average sum of optical flow magnitude was found to be negatively correlated with the accessibility score. More movement in a video could indicate that more actions will need to be explained in audio for accessibility. However, note that this by itself is not a very strong inference – it needs to be supplemented with a higher level information source, such as results from an action recognition model.

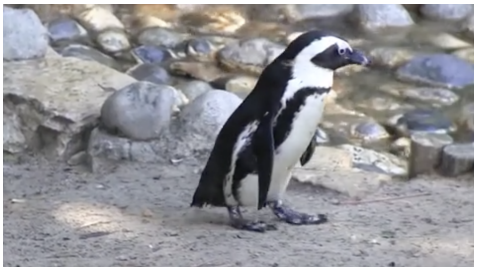
Audio Recognition. Existence of a detected speech event was found to be a major signal that predicts accessibility, as suggested by the high corresponding correlation value (0.4147) in Table 1. This is in a way an unsurprising observation – unless the point of interest in a video is also expressed in the audio stream with a well-known sound(*e.g.,* piano sound for a video where the main event is a piano being played), it is hard for the listener to understand the video with the remaining audio information.

Speech Analysis. We found a strong relationship between the length of the transcribed speech, both in terms of number of characters and words, and the accessibility ratings provided by the users. This implies that more speech content generally captures more context about the visual content of the video, which makes it easier for blind users to comprehend the events in the video.

Part-of-Speech Tagging Analysis. Part-of-Speech tagging revealed several signals that could be used to predict accessibility, which is expressed via f_9, f_{10} and f_{11} features in Table 1. For example, we found a positive relationship between the frequency of nouns and the accessibility score, regardless of the relevance of these nouns to the objects detected in the video. Although it does not establish a causal relationship, this suggests that more nouns could mean higher number of entities being referred to in a video, hence a higher chance of visual content or related entities being explained via audio. Similarly, higher number of counts and

(a) From LEDOV[23] dataset.

(b) From AViD[29] dataset.



(c) From LEDOV[23] dataset.

(d) From AViD[29] dataset.

Figure 4: Sample videos from the dataset. (a) and (b): sample videos with high rating, with both have $4/4$ score. (c) and (d): sample videos with low ratings, with both having 0.5 score.

pronouns could also possibly capture several entities or persons in the visual content, thereby increasing the accessibility of the video.

Music. Both the existence of music (f_7) and the proportion of the video covered by a music event (f_8) features were found to be negatively correlated with the accessibility score. We found many examples in our dataset where the videos had background music that was completely unrelated to the visual content, thereby lowering the accessibility of these videos. In fact, this was one of the main motivations that led us to consider features related to background music in our analysis. An exception to this scenario comprised instances where the music event detected was accompanied by an instrument or instruments being played in the video.

Comparison with Contemporary Video Accessibility Assessment [27]: We remark the above findings from our analysis share many similarities with those of a contemporary related research work [27]. For instance, the positive impact of feature f_1 (Speech ratio) on video accessibility was also observed in that prior work although by a different approach – by showing that the *% Non-Speech* feature that captured the proportion of non-speech duration in the video was negatively correlated with video accessibility. Similarly, the positive correlation between the feature f_9 (Nouns ratio) and f_{13} (Object-transcription match) and accessibility rating too was also equivalently captured as negative correlations between the features *low lexical density speech* and *% visual entities not in speech* in the prior work [27]. However, as the ground truth accessibility ratings in that work were obtained from people with visual impairments, they were unable to determine (with statistical significance) the type of correlation between the *number of visual entities/min* feature and the accessibility ratings, although their initial model suggested a negative correlation. In our work however, as the ratings were obtained from sighted users, we did not face this problem; we instead observed that there was a positive correlation between f_2 (Number of object predictions) and accessibility ratings. However, we observed that the feature f_3 (number of object prediction types) exhibited a negative correlation with the ratings.

4.4 Videos with High/Low Accessibility

In this section, we provide and discuss examples of a few videos in our dataset that were assigned high and low accessibility scores respectively by sighted users.

4.4.1 Videos with High Accessibility. We report the findings of a qualitative analysis of the videos in our dataset that were rated high (i.e., 3.5 or 4.0) by the sighted users. One of the common aspects we observed regarding the highly-rated videos was the existence of speech. Examples of this ilk included speech videos, where the videos contain a single person speaking with a constant background. Similarly, videos where the scenes were vividly and accurately described by a narrator in detail, also had high accessibility ratings.

Figure 4a and 4b depict a couple of example videos that were rated as highly accessible. The video in Figure 4a was taken from a scene where two pandas are shown with a background narrator speaking about the pandas. Although the exact scene itself is not described, the audio content being closely related to the visual content is presumably why the sighted raters concluded that the video is highly accessible. Figure 4b is from a scene where snow shoveling tips are being instructed by a background narrator while a man is shown shoveling snow in the video. The speech content, along with the shoveling sound together convey the visual content of the video via its audio, which may have led users to provide a high accessibility rating for this video.

ID	Age & Sex	Diagnosis	Media	Video Habits
P1	48/F	LCA	Mobile, PC	Everyday
P2	37/F	Congenital Glaucoma	Mobile	Everyday
P3	55/M	Optical atrophy	Mobile	Once-twice a week
P4	46/M	Diabetic Retinopathy	Mobile, TV, PC	Everyday
P5	41/M	Congenital cataracts	TV, PC	Everyday
P6	58/F	Congenital cataracts	PC, Mobile, TV	Every other day

Table 2: User study participant demographics.

4.4.2 Videos with Low Accessibility. Similar to the videos with high accessibility scores, we also observed certain patterns among the videos with low accessibility rating (i.e., videos with 0 or 0.5 score). For example, one class of videos had background music that was totally unrelated to the video content. Another type of videos in this category where those where the background sound was not discernible enough to be associated with a particular source.

Two examples of videos with poor accessibility ratings are shown in Figure 4c and Figure 4d. The video for Figure 4c is from a scene where background sound could be interpreted as coming from various sources, which possibly led users to provide low ratings to these videos. Figure 4d contains background music as the only audio theme, which makes it impossible for the user to comprehend the visual content just from the audio.

Overall, the examples we have seen in Figure 4 highlight the importance of the concordance between the audio and the visual channels of a video for improved accessibility. Notice how the example videos although of similar nature, as in the case of Figure 4a and Figure 4c, had contrasting accessibility scores, purely due to the nature of their audio content.

4.5 Evaluation with Visually-Impaired Users

We conducted a pilot study to better understand the video listening experience of users with visual impairments. Towards this, we recruited 6 users with visual impairments (3 male, 3 female) to better understand their habits of video interaction (See Table 2). All users except P1 relied on listening as the only way to consume videos. In the study, we asked the participants to listen to recordings of 30 randomly chosen videos from the dataset. The videos were presented to users in random order. For each video, the users were asked to describe the video, and their perception of how well they understood the video (On a Likert scale from 1 to 7). Below are some of our findings from this study:

Findings. Due to the differences between the methodology of collecting the ratings (i.e., pairwise comparisons vs Likert scale), we do not report correlation information between the ratings. We however observed significant differences between the ratings provided by sighted and visually-impaired users in certain specific videos. For example, a participant misinterpreted the sound of the wind outdoors in a kite running video as the sound of fire and gave a high accessibility score of 7. In another example, an interview video where only a person is shown speaking (which received a high accessibility score of 3.5/4 in our earlier evaluation with sighted users) was instead deemed inaccessible by two visually-impaired participants in this study – the video received low scores of 2 and 3 respectively. Lastly, the video that received the lowest score by the sighted participants (0.5/4) was instead considered average in terms of accessibility ratings given by the visually-impaired participants.

We observed during the study the participants were indeed aware of the fact that they were missing some information when the main audio theme in a video was music. However, in some cases, the participants were also able to distinguish between the cases where an instrument was being played in a video (hence possibly high-accessibility) and the cases where music was being played in background which was unrelated to the visual content of the video. When asked, one participant *P4* stated that ambient noise in the video was helpful in distinguishing between such cases.

4.6 Automatic Evaluation of Video Accessibility

We formulated the video accessibility evaluation task in two ways: (i) As a classification task by binarizing the accessibility scores into two classes; and (ii) As a prediction task by learning a regression model and computing the mean absolute error (MAE).

4.6.1 Accessibility Evaluation by Classification. In the binary classification task, the accessibility ratings were collapsed into two groups – accessible and inaccessible. Specifically, all videos which had an accessibility rating of at least 2.5 were labelled as accessible, and those with ratings below 2.5 were treated as inaccessible. This binning scheme resulted in 37.8% of the videos being labeled as accessible, and the remaining 62.2% as inaccessible. We trained several classifiers to learn this binary classification task with different combinations of our handcrafted features, and found out that a support vector machine (SVM[12]) classifier with RBF kernel and $C = 3.5$ yielded the highest F_1 score of 0.675 (precision=0.746, recall=0.550) for the positive class (i.e., accessible) after 5-fold cross validation, averaged over 10 runs. This shows that there is still scope for improvement in this classification task, and it could benefit from a larger annotated dataset and more expressive features.

4.6.2 Accessibility Evaluation by Regression. For this task, we trained a multi-layer neural network (3 fully connected layers with ReLU nonlinearities in between) as the regression model that can predict an accessibility score for a given video. This model accepts input in the form of a vector of handcrafted features, which were previously described in Section 3.2. For ground truth, we leveraged the accessibility ratings produced by sighted users. 5-fold cross validation averaged over 10 runs resulted in a Mean absolute error (MAE) of 0.53. Note that most ground truth ratings were between 0 and 4, while only 10 videos had a rating of exactly either 0 or 4.

Note that prior related work [27] also trained and evaluated a prediction model based on regression to assess video accessibility. However, in their work, they used a linear regression model whereas we performed both a neural regression task and a classification task.

5 DISCUSSION

5.1 Video diagnostics

An advantage of handcrafted features is the possibility of deducing the causes underlying the predictions made by either the classification or the regression model. In our approach, providing explanations for the predictions is as important as providing an accessibility rating or class, since the underlying reasons could have implications for both consumers and creators of videos. For example, accurately predicting a video as inaccessible while also explaining to the video creator that a particular video scene does not contain any speech event, can immensely help the creator 'fix' the accessibility issues by supplementing that scene with video descriptions.

Our approach is simply based on comparing handcrafted features computed for a particular video against the distribution of the overall dataset. A similar comparison approach was used in [27].From this comparison to the overall distribution, we can identify sources of accessibility issues by detecting undesirable values for the various features. For instance, for features that positively correlate with accessibility rating, a low value may indicate a potential reason for accessibility problems. Similarly, a high value for a negatively correlated feature can also point to a potential source of the accessibility problems. This insightful knowledge will enable both video content creators and consumers to be informed about the potential reasons behind a prediction, thereby permitting better allocation of resources for improving accessibility (*e.g.*, adding video descriptions).

5.2 Limitations & Future Work

We discuss some of the aspects of our work which can be further improved, and future directions that can be explored next.

Dataset size & variety. Although the two datasets we used for labeling consisted of diverse sets of scenes and actions, the cost of manual filtering and annotation limited the number of videos that made their way into our final dataset. This created an inevitable dataset bias. For example, although we believe the features described in Table 1 will generalize to many videos outside our dataset, the exact correlation values are still highly dependent on the videos included in the dataset. A more general analysis and accurate model for evaluating accessibility will require larger and diverse datasets from which complex relationships can be derived, an observation that was also made by prior related work [27]. Also, more comparisons per video in the dataset will result in more fine-grained and reliable scores.

Dataset Artifacts. Some of the videos included from the AViD dataset contained blurred faces to preserve anonymity. Even though we observed that a lower confidence threshold for object detection mitigates this problem, this could have still impacted the performance of person detection in our work as it is possible that fewer than actual number of persons were detected during our analysis. Second, due to pre-processing, some of the videos have a still frame at the end. The still frame appears longer than a second for 16% of the videos, and more than two seconds for 1% of the videos. For these videos, we computed the audio and video features for the duration only when both video and audio are present and changing.

Better understanding of audio-video relationship. The analysis of relationship between the visual content and the audio in our approach was limited to explicit signals such as transcribed speech and detected audio events. While this approach has been demonstrated to be useful, it is not as powerful as a thorough analysis of understanding how much of the visual content is explained by the accompanying audio, which can be a research problem in itself. Further work in this regard can potentially result in better handling of the video accessibility evaluation task.

On-screen text. We did not attempt to analyze text that may sometimes appear in videos (*e.g.*, subtitles), which is another source of inaccessibility that we (and also prior work [27]) discovered during the study with users having visual impairments. One way to incorporate such text content into accessibility evaluation is to assess whether text content exists or not, and to analyze the relationship between the on-screen text and the speech/audio events, similar to that suggested in [27].

Improved feature extraction. Features extracted in our model were handcrafted and targeted at finding specific properties based on our manual observations regarding accessibility of videos. While these features indeed facilitate diagnostics, an accurate predictive model does not necessarily have to rely only on handcrafted features. Visual information that is not captured by our handcrafted features can possibly be captured by state-of-the-art deep learning methods and is very likely to boost the prediction performance.

6 CONCLUSION

In this paper, we analyzed a diverse set of handcrafted features that characterize accessibility of videos, and built prediction models for quantifying accessibility of videos. Towards this, we collected a labeled dataset of accessibility evaluations from sighted users, and then used handcrafted features, some of which exist in the literature, extracted from the videos in this dataset to find features that correlate either positively or negatively with video accessibility. These handcrafted features can not only be used as a means for predicting accessibility scores of videos, but also provide users with explanations regarding the factors that impacted the predicted accessibility score. Through a user study with 6 participants who were visually impaired, we found cases where the participants' perception of accessibility differed from the annotations provided by the sighted users. This work could pave the way for future video accessibility research with more data and use of more sophisticated machine learning models to understand in-depth the accessibility relationships between the visual and speech aspects of a video.

ACKNOWLEDGMENTS

This work was supported by NSF awards 1805076, 1936027, 2113485 and NIH awards R01EY030085, R01HD097188.

REFERENCES

[1] [n.d.]. iOS Accessibility Scanner Framework. https://github.com/google/GSCXScanner.
[2] [n.d.]. *WAVE Web Accessibility Evaluation Tool.* https://wave.webaim.org/.

[3] 2021. Improve your code with lint checks. https://developer.android.com/studio/write/lint.

[4] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)* 52, 6 (2019), 1–37.

[5] Tania Acosta, Patricia Acosta-Vargas, Jose Zambrano-Miranda, and Sergio Lujan-Mora. 2020. Web Accessibility evaluation of videos published on YouTube by worldwide top-ranking universities. *IEEE Access* 8 (2020), 110994–111011.

[6] Amazon. [n.d.]. Amazon Rekognition – Video and Image - AWS. https://aws.amazon.com/rekognition/.

[7] Inc Amazon Web Services. [n.d.]. *Amazon Transcribe – Speech to Text - AWS*. https://aws.amazon.com/transcribe/.

[8] Chieko Asakawa, Takashi Itoh, Hironobu Takagi, and Hisashi Miyashita. 2007. Accessibility evaluation for multimedia content. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 11–19.

[9] Ali Selman Aydin, Shirin Feiz, Vikas Ashok, and IV Ramakrishnan. 2020. Towards making videos accessible for low vision screen magnifier users. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 10–21.

[10] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 [cs.CV]

[11] Vicente Luque Centeno, Carlos Delgado Kloos, Jesús Arias Fisteus, and Luis Álvarez Álvarez. 2006. Web accessibility evaluation tools: A survey and some improvements. *Electronic notes in theoretical computer science* 157, 2 (2006), 87–100.

[12] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[13] Alireza Darvishy, Hans-Peter Hutter, and Oliver Mannhart. 2011. Web application for analysis, manipulation and generation of accessible PDF documents. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 121–128.

[14] Luchin Doblies, David Stolz, Alireza Darvishy, and Hans-Peter Hutter. 2014. PAVE: A web application to identify and correct accessibility problems in PDF documents. In *International Conference on Computers for Handicapped Persons*. Springer, 185–192.

[15] Benoît Encelle, Magali Ollagnier-Beldame, Stéphanie Pouchot, and Yannick Prié. 2011. Annotation-based video enrichment for blind people: A pilot study on the use of earcons and speech synthesis. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. 123–130.

[16] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*. Springer, 363–370.

[17] FIDE. [n.d.]. FIDE Handbook C. General Rules and Technical Recommendations for Tournaments / 04. FIDE Swiss Rules / C.04.1 Basic rules for Swiss Systems /. https://handbook.fide.com/chapter/C0401.

[18] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. 2015. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding* 134 (2015), 1–21.

[19] Kentarou Fukuda, Shin Saito, Hironobu Takagi, and Chieko Asakawa. 2005. Proposing New Metrics to Evaluate Web Usability for the Blind. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1387–1390. https://doi.org/10.1145/1056808.1056923

[20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[21] Julia González, Mercedes Macías, Roberto Rodríguez, and Fernando Sánchez. 2003. Accessibility metrics of web pages for blind end-users. In *International Conference on Web Engineering*. Springer, 374–383.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[23] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. 2018. DeepVS: A Deep Learning Based Video Saliency Prediction Approach. In *The European Conference on Computer Vision (ECCV)*.

[24] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. 2013. Understanding and predicting interestingness of videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27.

[25] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2019. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *arXiv preprint arXiv:1912.10211* (2019).

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[27] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 272, 14 pages. https://doi.org/10.1145/3411764.3445233

[28] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1* (Philadelphia, Pennsylvania) *(ETMTNLP '02)*. Association for Computational Linguistics, USA, 63–70. https://doi.org/10.3115/1118108.1118117

[29] AJ Piergiovanni and Michael S Ryoo. 2020. AViD Dataset: Anonymized Videos from Diverse Countries. *arXiv preprint arXiv:2007.05515* (2020).

[30] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication* 30 (2015), 57–77.

[31] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. 2009. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics* 10, 4 (2009), 30–45.

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs.CV]

[33] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. arXiv:1612.08242 [cs.CV]

[34] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[35] Camila Silva, Marcelo Medeiros Eler, and Gordon Fraser. 2018. A survey on the tool support for the automatic evaluation of mobile accessibility. In *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*. 286–293.

[36] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. 2015. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia* 17, 10 (2015), 1733–1746.

[37] Markel Vigo and Giorgio Brajnik. 2011. Automatic web accessibility metrics: Where we are and where we can go. *Interacting with computers* 23, 2 (2011), 137–155.

[38] W3C. 2021. *Audio Description of Visual Information | Web Accessibility Initiative (WAI) | W3C*. https://www.w3.org/WAI/media/av/description/.

[39] W3C. 2021. *Making Audio and Video Media Accessible | Web Accessibility Initiative (WAI) | W3C*. https://www.w3.org/WAI/media/av/.

[40] YouTube. [n.d.]. *YouTube for Press*. https://blog.youtube/press/.

[41] 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.

[42] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2019. Object Detection in 20 Years: A Survey. arXiv:1905.05055 [cs.CV]