Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users

Beste F Yuksel [1], **Pooyan Fazli** [2], **Umang Mathur** [2], **Vaishali Bisht** [2], **Soo Jung Kim** [1], **Joshua Junhee Lee** [3], **Seung Jung Jin** [1], **Yue-Ting Siu** [4], **Joshua A. Miele** [5], **Ilmi Yoon** [2]
[1]Department of Computer Science, University of San Francisco, San Francisco, CA
[2]Department of Computer Science, San Francisco State University, San Francisco, CA
[3]Computer Science, University of Washington, Seattle, WA
[4]Department of Special Education, San Francisco State University, San Francisco, CA
[5]Smith-Kettlewell Eye Research Institute, San Francisco, CA
ilmi@sfsu.edu (corresponding author)

ABSTRACT
Video accessibility is crucial for blind and visually impaired individuals for education, employment, and entertainment purposes. However, professional video descriptions are costly and time-consuming. Volunteer-created video descriptions could be a promising alternative, however, they can vary in quality and can be intimidating for novice describers. We developed a Human-in-the-Loop Machine Learning (HILML) approach to video description by automating video text generation and scene segmentation and allowing humans to edit the output. The HILML approach facilitates human-machine collaboration to produce high quality video descriptions while keeping a low barrier to entry for volunteer describers. Our HILML system was significantly faster and easier to use for first-time video describers compared to a human-only control condition with no machine learning assistance. The quality of the video descriptions and understanding of the topic created by the HILML system compared to the human-only condition were rated as being significantly higher by blind and visually impaired users.

Author Keywords
Video Accessibility; Video Description; Blind Users; Visually Impaired Users; Machine Learning; Human-in-the-Loop;

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

INTRODUCTION
The World Health Organization (WHO) estimates that roughly 285 million people worldwide are visually impaired, and 39 million of them are blind [11]. In the United States alone, there are 10 million people who are visually impaired, of whom 1.3 million are blind [11]. Video accessibility for blind or visually impaired users[1] becomes crucial as online videos on websites such as YouTube come to play an essential role in our daily lives. However, the vast majority of video material on the Internet is currently not accessible to these millions of visually impaired people who would significantly benefit from improved access to videos for education, employment, social, and entertainment purposes, thus creating an increasing digital divide between blind and sighted users. The United Nations Convention on the Rights of Persons with Disabilities states that persons with disabilities should be able to access services on the Internet (Article 21) and enjoy access to cultural material in accessible formats (Article 30) [55]. The Web Content Accessibility Guidelines recommends that audio descriptions be provided for all online videos [17]. Despite such international guidelines and standards, there is still a paucity of videos that are made accessible through video description for blind or visually impaired users. Video description can be defined as "Narration added to the soundtrack to describe important visual details that cannot be understood from the main soundtrack alone" [10]. Video description is particularly important as the speech communication channel is considered to be one of the most important modalities for blind and visually impaired users [25].

A significant bottleneck in video accessibility for blind and visually impaired users is the time and cost to produce video descriptions professionally. Professional video description typically has a minimum two-week turnaround time and can cost up to hundreds or thousands of dollars for a video [22], depending on its length. Alternatives to professional video description include tools such as MAGpie [1] and LiveDescribe [50], which allow anyone to describe videos. The website YouDescribe [73] is a free platform with around 2500 sighted volunteers that have described over 4000 YouTube videos. However, inexperienced describers can struggle to produce

DIS '20, July 06–10, 2020, Eindhoven, Netherlands

ACM ISBN 978-1-4503-6974-9/20/07…$15.00
DOI: http://dx.doi.org/10.1145/3357236.3395433

[1]In this paper, we use the terms blind and visually impaired users to refer to individuals who rely on audio descriptions, tactile graphics, or use magnification to typically access images.

high quality descriptions, with the quality of video descriptions on YouDescribe varying significantly depending on the volunteers.

Describing videos requires expertise to clearly but *succinctly* describe scenes within limited gaps in the original soundtrack. Video description can be daunting for inexperienced volunteers who can be limited by their own vocabulary, writing skills, or lack of training. Be My Eyes [23] is an application that connects blind and visually impaired people with two million sighted volunteers for visual assistance through live and generally short video calls. Informal feedback from Be My Eyes volunteers to the authors revealed that the difficulty of the task and skills required such as language and/or writing skills, combined with the public exposure of their work, were deterrents to them volunteering for video description. This informal feedback, the large discrepancy in volunteer numbers between YouDescribe and Be My Eyes, along with the fact that Be My Eyes is currently overwhelmed with fifteen times the number of sighted volunteers to blind or low-vision users [23], suggests that people are willing to help blind and visually impaired users, but describing videos is not something that are able to or willing to currently do easily.

In this paper, we turn to the use of Human-in-the-Loop Machine Learning (HILML) to facilitate easier video description for novice describers and to improve the quantity and quality of video accessibility for blind and visually impaired users. While the algorithms applied in this paper are not new, this is the first time that they have been applied and tested in the field of automated video descriptions for blind and visually impaired users. We built a video description interface with automated scene segmentation, text generation of images, and text generation of text on screen using state-of-art deep learning Image Captioning [40] and Microsoft AI Video Insight (called Video Indexer) [7]. The human is then able to edit the generated output, allowing human-machine collaboration to produce high quality video descriptions while keeping a low barrier to entry for volunteer describers. We evaluated this interface on novice describers and compared it to a control condition with no machine learning assistance (Experiment 1). We then asked blind and visually impaired users to rate the quality and clarity of the video descriptions produced by the two conditions (Experiment 2). Thus, the contributions of this paper are as follows:

1. Design and develop a Human-in-the-Loop Machine Learning (HILML) system to generate video descriptions for visually impaired and blind users.
2. Demonstrate that the HILML system is faster and easier to use for novice video describers than a human only control condition.
3. Demonstrate that the quality of the video descriptions and topic understanding is significantly higher for the HILML system for blind and visually impaired raters compared to a human only control condition.

RELATED WORK

There have been a wide range of innovative works in the field of HCI and accessibility for blind and visually impaired users including visual graphics [48], data visualizations [70], maps [53, 56], programming [59], board games [18], video games [29, 62, 8], and appliance displays [26]. For the purposes of this paper, we focus on research in video description and divide the related work in this area into: 1) front-end user interface designs for increased video accessibility, and 2) tools developed for the automation of video description.

Video Description and User Interfaces

Front End User Interface Design for Video Accessibility

Various researches have investigated the front-end of the user interface for increased accessibility of videos and films for blind and visually impaired users. Chapdelaine and Gagnon [15] investigated on-demand video description with a Website platform that could provide users with extended descriptions that exceeded the natural pauses in the soundtrack by pausing the video. Results showed that selecting various video description levels was appreciated by around 92% of the users. Chapdelaine [14] demonstrated a DVD player with video description with three extra specialized functions: contextual functions, differing quantities of audio description, and recall assistance. Results showed that 94% of blind and visually impaired users appreciated having these extra functions.

Encelle et al. [22] investigated annotation-based video enrichment made up of both speech synthesis and 'earcons', i.e., non-verbal audio messages to provide information to the user. Results showed that earcons could be used together with speech synthesis to enhance understanding of videos when accompanied with explanations. Encelle et al. [21] investigated the use of artificial pauses in audio described videos in order to transmit more information in the video descriptions and suggest the placing of artificial pauses right after the soundtrack's natural gaps [21].

Tools Toward Automated Video Description

LiveDescribe was the first known computer-assisted video description [13]. It was designed for real-time descriptions of live television broadcasts. The two automated features of LiveDescribe were: 1) automatic detection of silent periods where a describer could choose to add a description, and 2) allow descriptions to go beyond the silent periods by automatically pausing the video to accommodate extended descriptions. Feedback from describers was positive but also reflected the high cognitive workload of live description [13].

The Collaborative Annotation for Video Accessibility (ACAV) project developed a prototype for an adaptive video enrichment system personalized to blind and visually impaired users based on decision trees [63]. After an initial learning phase with user feedback, the system predicted feedback using a decision-tree. The idea of personalizing video descriptions to individuals is laudable, however, as stated by the authors, the prototype needs to be evaluated with user studies to investigate the adaptive system further [63].

As part of the e-inclusion project, Gagnon et al. [28, 27] built a prototype software for professional describers that used computer vision to extract visual content from videos. They used computer vision techniques such as shot transition detection, key frame identification, face recognition, text spotting, visual

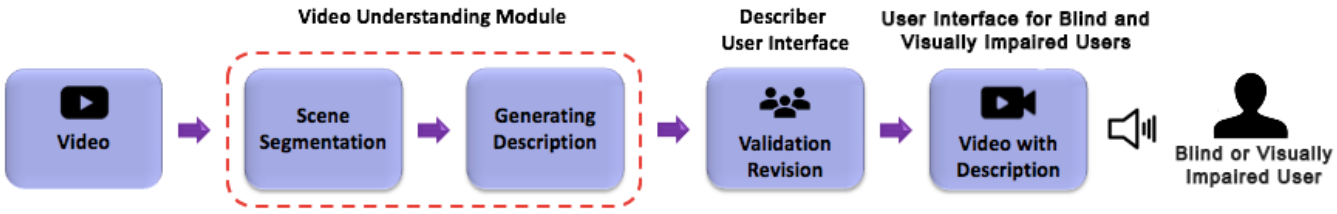

Figure 1. The workflow of the HILML system: the inputted video is segmented into scenes, automated text is generated for the images and text on screen, which can then be edited by the human. The descriptions (with the segmented timings) are synthesized to speech and outputted to play alongside the video for the blind or visually impaired user.

motion, gait/gesture characterization, key-place identification, key-object spotting, image categorization, and also detected silent areas in the soundtrack [28, 27]. Professional describers, already have expertise and training in how to sufficiently and succinctly use language for video description. Gagnon et al. [28, 27] displayed detected features on an interactive timeline which allowed professional describers to have an overview of the frequency and duration of features in the video, as well as the silent areas. This helped them to use their expertise in *placing* their text in the correct areas. Their work differs from ours in two key ways: 1) their computer vision techniques were employed to produce specific tasks such as key-face recognition, key-object spotting etc while our approach utilizes deep learning general intelligence of video description generation, and 2) their user interface was designed specifically for professional video describers while we aim to lower the barrier to entry for novice describers, with human-machine collaboration.

Kobayashi et al. created audio description synthesis by using external metadata of videos on an online platform [45, 46]. They provided a script editor to the describers, a prototype video player for the blind and visually impaired users, and metadata storage exchange between the components. However, they found there was a gap in the quality of audio descriptions produced by the software against professional narrated descriptions. Kobayashi et al. [46] tested their describer script editor, which has a visual interface to edit the video description sentences and to specify where on the video timeline to read each sentence aloud, on one novice and one professional describer. Ratings by blind and visually impaired users showed significantly lower scores for the novice describer [46]. The novice describer estimated that it would take three times longer without the script editor to produce video descriptions, but no control condition was carried out.

Machine Learning in Video Understanding

Classical video description approaches combined subject, object, and verb (SVO) detection from visual entities with template based language models to generate sentences [47]. However, the release of large datasets revealed that these methods cannot cope with the diversity in unconstrained open domain videos. The 'Inverse Hollywood Problem' approach tried to describe a series of actions into semantic tag summaries in order to develop a storyboard from instructional videos [12]. Many classical SVO approaches [32, 20, 9] were able to heuristically parse the videos into a series of key actions and generate a script that describes actions detected in the video. It also generated key frames depicting the detected causal events and defined the series of events into semantics representation, but the object and vocabulary/actions were very limited [3].

Classical approaches and other advancement such as statistical methods were soon replaced with deep learning, the current state of the art in video description. In particular, Convolutional Neural Networks (CNNs) [6] are the state of the art for modeling visual data recognition [6, 65, 67] and Long Short-Term Memory (LSTMs) [34] are now dominating the area of sequence modeling such as NLP [16, 31, 66]. The deep learning approaches to video description are also usually divided into two sequential stages, namely, visual content extraction and text generation. Visual features represented by fixed or dynamic real-valued vectors are produced instead. This is often referred to as the video encoding stage. CNN, Recurrent Neural Network (RNN), or LSTM are used in this encoding stage to learn these visual features, that are then used in the second stage for text generation, also known as the decoding stage. For decoding, different flavours of RNNs are used, such as deep RNN, Bi-directional RNN, LSTM, Gated Recurrent Units (GRU), Attention or Transformer [69]. The resulting description can be a single sentence or multiple sentences. Roemmele et. al used a recurrent-neural-network architecture to generate stories in a sequence-to-sequence manner [61]. Martin et. al proposed an event representation for neural-network-based story generation [51]. Fan et al. created a hierarchical model that automatically generates stories conditioning on the writing prompts [24]. Huang et. al introduced the visual storytelling task, in which the trained model takes a sequence of photos as input and generates a short story that narrates this photo sequence [38].

Our approach assists sighted users to create video descriptions by combining automation using deep learning with human input and editing. Such an approach can be considered Human-in-the-Loop Machine Learning, which we examine next.

Human-in-the-Loop Machine Learning

Human-in-the-Loop Machine Learning (HILML) has been defined as human and machine learning (ML) processes interacting to solve one or more of the following: 1) Making ML more accurate, 2) Getting ML to the desired accuracy faster, 3) Making humans more accurate, and 4) Making humans more efficient [54]. In this paper, we aim to use HILML to make the describers more accurate and efficient, however, in the long-term, the data collected can ML to become more accurate as well. HILML is an important field to both the HCI and

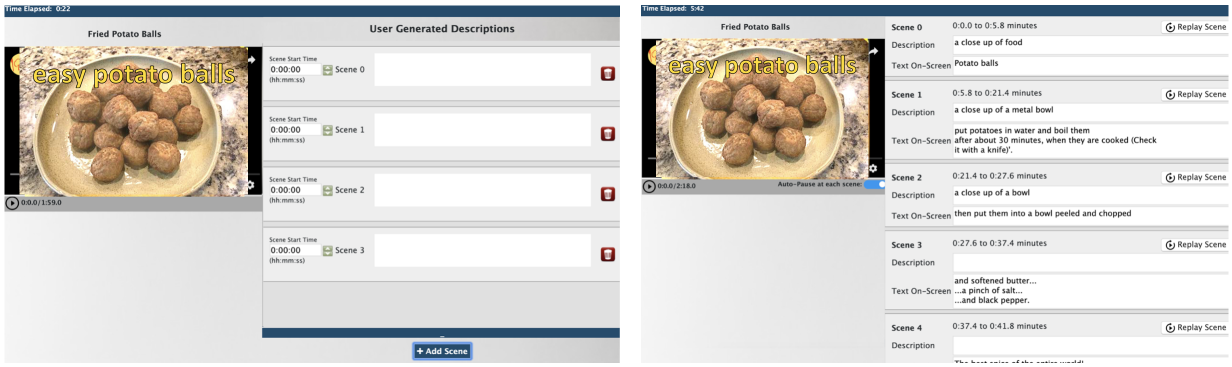

Figure 2. The Describer User Interface for the sighted volunteers in Experiment 1 across the Human-Only (Left Panel) and HILML (Right Panel) conditions. The video is played on the left-hand side and the edit/input interface is on the right. In the Human-Only condition, describers need to create their own scene segmentations and generate all of the descriptions from scratch. In the HILML condition, the scenes are automatically segmented for users. Automated description from images in the video and text on screen is generated for users in script editors, allowing users to edit the automated output. Figure 3 shows a close-up of this section.

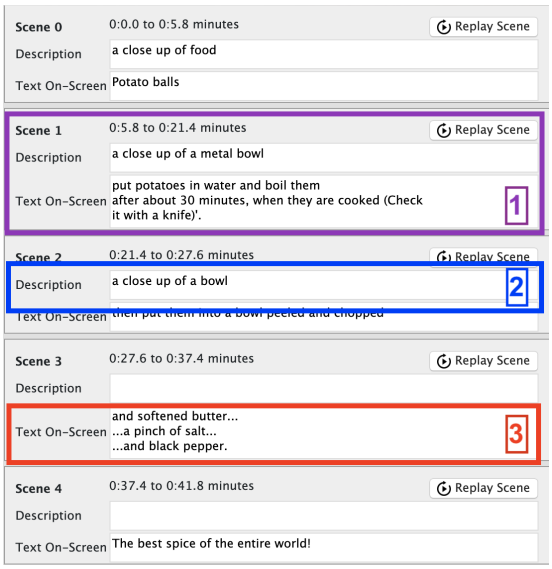

Figure 3. Close-up of the Describer User Interface and script editor of the HILML system displaying: 1) automated scene segmentation, 2) text generation for the description of images in the video, and 3) text generation for text that appears on the screen. This automated text can then be edited by the describer (the text shown here is the automated output with no human edits).

ML communities and can create important collaborations. For example, [54] points out that when machine translation was hitting difficulties in making human translation faster and more accurate, it was an award winning paper in human-computer interaction by Green et al. [64] that created a breakthrough with human post-editing for language translation.

HILML has been used in fields such as assistive robotics [30], data analytics [64], systems [35, 72], and audio labeling [43]. In the field of image and video classification, Wang et al. [71] used HILML to create a semi-automatic method to segment foreground moving objects in surveillance videos. They combined human outlining of a small number of moving objects with a convolutional neural network to reach similar levels of accuracy as a human with less manual work than human-only analysis [71]. Pirrung et al. [58] created a HILML image organization web application where humans can organize their images and the machine learns from these groupings. In return, the machine can reposition images or regroups them to reflect its assessment of the human's mental model, which may then be refined by the user.

In this paper, we leverage both human and machine intelligence to create a HILML system to aid sighted humans create video descriptions for blind or visually impaired users. We now describe the HILML system we created for video description generation.

SYSTEM

We built a system to assist human volunteers in producing video descriptions using a HILML approach. Video description including YouDescribe uses either in-line description or extended-description to add descriptions over existing video. In-line description is inserted to the original video if audio is quiet and is preferred. Extended-description is inserted by stopping video flow when the original video audio track is too busy with dialog or important sound effect. While both descriptions are helpful, descriptions are desired to be succinct and minimal while delivering important information of the scenes. With these required features, we have designed the system as below.

The following describes the workflow (Figure 1) of the developed system:

1. **Input Data:** Videos for which descriptions have been requested are forwarded to the model for further processing.

2. **Scene Segmentation and Key Frame Extraction** [Video Understanding Module]: The video is segmented into a sequence of scenes of varying time spans, and key frames are sampled to maintain the appropriate granularity of the scene for generating the description. A new scene is created when either 1)the background changes, or 2)a new character in the foreground enters or exits the scene. A new description is generated and attached to the scene. We merge any scenes with continuous dialog as the description is read out using text-to-speech technology. Describers can also merge scenes

if they wish. Information extracted from the video is carried out by Microsoft Azure Video Indexer API [52]. Text scripts with the speaker identity and timestamp tags are provided by ListenByCode [49].

3. **Generating Video Description** [Video Understanding Module]: Sequences of key frames are processed by the model to generate descriptions that best explain the scene in the video. The description covers any text in the key frames, people with ID (to handle reappearances), gender, emotion, hair color, age, objects, and environment. People are recognizable if they are known celebrities. We utilize Pythia Deep Learning image description generation model [60] for each key frame image as there are only a few key frames per scene. We used BLEU [57] scores to drop similar sentences and combined sentences to build multi sentence baseline text. Some type of text on screen is very important and can often be forgotten by volunteer describers while other types of text appear on the screen such as street signs that should not be read out. We used a threshold for time spent on screen to determine whether the text should be described.

4. **Validating or Revising the Description** [Describer User Interface]: Through the describer interface (Figures 2 and 3), sighted users can view video scene segments, their associated descriptions, and can make their own edits and notes in the script editor. Volunteers improve this narration by revising, merging, or completing the machine-generated descriptions. Compared to free-form descriptions that volunteers create from scratch, the proposed process assists sighted volunteers to structure and complete the descriptions. Figure 3 shows the the scene segmentation, text generation for description of images in the video, and text generation for text that appears on the screen. When the video plays, each corresponding scene gets highlighted and automatically stopped at the end of each scene to assist sighted volunteers. Volunteers can replay the working scene with a button click.

5. **Playing Video with Description** [User Interface for Blind and Visually Impaired Users]: The video is played to the visually impaired or blind user with audio descriptions synchronized to the original video. Audio is generated using IBM Watson's text-to-speech API. Synthesized video descriptions have been shown to be acceptable to blind and visually impaired users [15, 46] and are used by professional video description production companies, such as 3play Media [2] and Automatic Sync Technologies [68]. ListenbyCode [49] to processes the audio track to return the script of the dialog, timestamp at each dialog, and the corresponding speakers. This information determines the scene segmentations and also detects the period where the text-to-speech converted descriptions can be inserted. If a suitable period is not found due to busy dialog, then the speech will be inserted as "extended description" that pauses the video.

Machine Learning Algorithm

When picking the machine learning algorithms described above, we tested several state-of-the-art image captioning deep learning technologies: Microsoft AI Video Insight, Pythia, the winning entry from Facebook AI Research (FAIR)'s A-STAR team to the VQA Challenge 2018 [40] and GLACNet, the winning champion from Visual Story Telling challenge for text generation [44]. While these state-of-the-art image captioning technologies can recognize objects successfully, their vocabulary and sentence generations are too limited by the data sets (e.g. Common Object in Context (COCO) data set [39]) to successfully describe diverse YouTube videos and actions are rarely captured. Although the machine-generated descriptions are currently not as good as human-generated ones [36], the latest research shows potential for rapid development in the near future. For the present time, it allows sighted volunteers some support in getting started with video descriptions rather than starting from scratch.

EXPERIMENT 1: DESCRIBING VIDEOS

Experiment 1 investigates the use of the HILML system for video description generation for novice describers. Participants described videos in two conditions: one with the HILML system with machine learning assistance for text generation and scene segmentation, and one with no machine learning assistance (Figure 2).

Experimental Design

In Experiment 1, twenty two participants (11 male, 11 female), aged 18 to 34 (mean age of 24.0, SD of 4.45) took part in a within-subject design and will be referred to as the *describers*. All describers first took a short tutorial to familiarize themselves with the video description software. Describers were then given two videos to describe, one with and one without the HILML system. In the *HILML condition*, automated text and scene division was provided which could then be edited by the describers if they wished. In the control condition, which we refer to as the *'Human-Only' condition*, describers typed out all of their video descriptions from scratch in the user interface. Figure 2 show examples of both conditions in the user interface.

Both videos were 'How-To' cooking videos of around 2 minutes length each. 'How-to' videos are requested by visually impaired users on YouDescribe. The 'How-To' video style requires a high degree of quality description which includes: 1) text that is on the screen to be described if it is present for greater than 1 or 2 seconds, and 2) the scene to be described every time it changes. The describers were divided into two groups: Group 1 described Video 1[2] with the HILML system and Video 2[3] without the HILML system. Group 2 described Video 1 without the HILML system and Video 2 with the HILML system. The order of the conditions and group allocations was alternated between participants. Both videos had music playing throughout with text on screen appearing with instructions and ingredients. At the conclusion of each video description, describers were asked to fill out a questionnaire on their describing experience including an unweighted NASA-TLX survey [33] (a subjective workload assessment tool) and were given a short interview.

[2]https://www.youtube.com/watch?v=cNj3aOTYdQQ
[3]https://www.youtube.com/watch?v=nqXz8hhAYGo

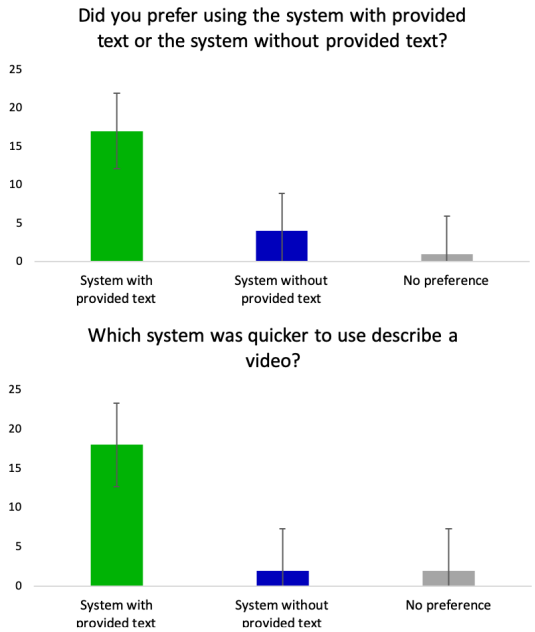

Figure 4. Describers' ratings of the HILML and Human-Only conditions. Describers showed (Top:) noticeable preference for the HILML system and (Bottom:) felt it was quicker to use.

Results

Questionnaire Data

Questionnaire data shows that describers preferred the HILML system and felt that it was quicker than the Human-Only condition (Figure 4) even though both systems were generally easy to use for most participants (Figure 5). This is interesting as the accuracy of the provided text by the HILML system was reported as being somewhat accurate, with the mode response being 'neutral' (Figure 7). This suggests that even though the accuracy of the HILML system was not perfect, it was *still* helpful (as shown in Figure 6) than no machine learning assistance at all.

Time Taken to Complete Video Descriptions

We performed a t-test on the mean time (in seconds) that describers spent describing a video across the HILML and Human-Only conditions. Results showed that users described videos with the HILML condition (μ = 1285.41, σ = 659.51) significantly faster than the Human-Only condition (μ = 1825.45, σ = 658.51) ($t(21) = 2.83, p = 0.005, d = 0.60$). Figure 8 shows the time taken between the two conditions. These findings correspond with the questionnaire data by describers that the HILML system was quicker to use than the Human-Only condition (Figure 4).

	Human-Only	HILML	Wilcoxon Z	*p*	effect size
Mental demand	**64.1 (20.7)**	**43.4 (23.9)**	**-2.486572**	**0.011110**	**0.530139**
Physical demand	19.8 (20.5)	14.8 (17.6)	-0.813411	0.427694	0.173420
Temporal demand	40.5 (29.7)	28.0 (22.0)	-1.463076	0.148385	0.311929
Performance	31.1 (22.3)	26.8 (27.8)	-0.179176	0.866728	0.038200
Effort	**60.7 (18.5)**	**45.7 (23.3)**	**-2.199999**	**0.026089**	**0.469041**
Frustration	**43.4 (27.3)**	**23.6 (22.9)**	**-2.548356**	**0.008972**	**0.543311**

Table 1. Mean (st. dev.) NASA-TLX values (0=low, 100=high) and results from Wilcoxon Signed-ranked test. Rows in bold indicate significant results.

Describer Workload

We used the NASA-Task Load Index (TLX) to capture and analyze perceived operator workload for the two conditions of video description. We inputted the participants' ratings into the NASA-TLX app for iOS [4] which calculated a score for each of the six ratings scale results as shown in Table 1. Results of the NASA-TLX responses (Table 1) showed that describers found the HILML system to be significantly less mentally demanding, require less effort, and be less frustrating. This is consistent with the subjective questionnaire data that describers found the HILML system to be easier to use, less frustrating, and more enjoyable than the Human-Only condition (Figure 5).

Interview Data

All describers were asked the following three interview questions: 1) '*How helpful was the provided text?*', 2) '*Does using the provided text make describing the videos easier?*', and 3) '*Is there anything else you'd like to add about the systems you used to describe the videos?*'.

Sixteen out of 22 describers felt that the HILML system was helpful or easier to use because it provided guidelines:

"I think the provided text was a good set-up because it allows us to get a better idea of what the plot is versus misinterpreting it, and I think that is beneficial if you give it to anyone."

"[The provided text] provided a lot more help than just going in blank without some kind of subtext... it shows the bigger picture of what you are going for versus again describing something that may not be as relevant allowed me to save a lot more time on that."

"I feel like it gave me some type of guideline as to what I should put, and not too many details."

The HILML system also seemed to provide a guideline for scene segmentation, with four out of 22 describers making comments such as:

"More helpful part was that it was already booking up into the different scenes because the fact that it was automatically booking up meant that all I have to do is just create the script."

"Like with the second video [Human-Only], it made it a lot harder when I had to find out which time it started, which time it ended for each scene."

Four out of 22 describers felt that the descriptions were "*vague*" and "*simple*" (although two of these describers still felt the HILML system did provided them a starting point or a guideline) and made comments such as:

"It just wasn't that descriptive. It was kind of vague."

"It was helpful for starting it. But it wasn't too helpful in being more descriptive. It would be a very simple description like 'a bowl was present'. "

Describer interview data was consistent with questionnaire data on the neutral level of accuracy of the HILML text generation (Figure 7) with comments such as:

"I didn't think so. I didn't think it was very accurate sometimes, and if it was just something really simple and then I just thought I could describe it better."

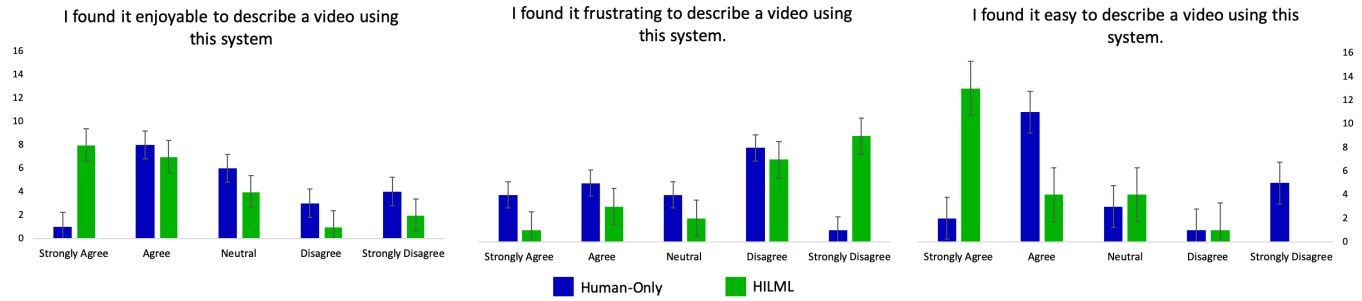

Figure 5. Describers' ratings of the HILML and Human-Only conditions for Left: enjoyment of system, Center: frustration caused by system, and Right: ease of use of system.

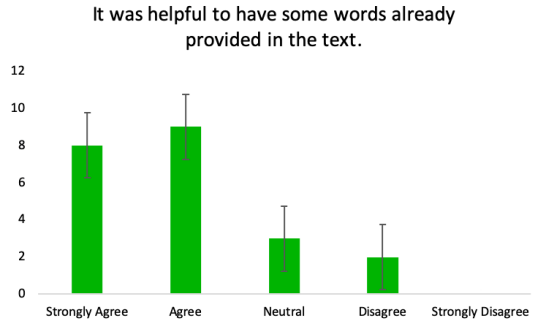

Figure 6. Describers' ratings of the helpfulness of the provided text generated by the HILML system. Most describers agreed or strongly agreed that it was helpful.

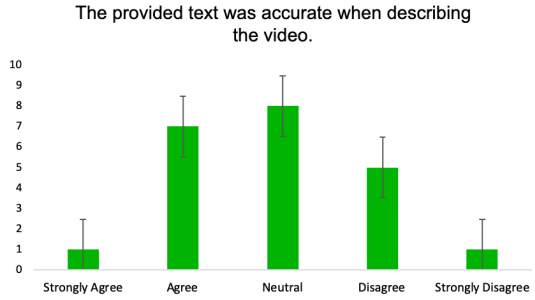

Figure 7. Describers' ratings of the accuracy of the provided text generated by the HILML system. The most highly rated level was 'neutral'.

"It was kind of neutral because I also needed to interpret what it was saying. Some of the text was helpful but the other things I deleted it completely and start from scratch."

We discuss the implications of these findings, along with those from Experiment 2, in the overall Discussion section.

EXPERIMENT 2: VIDEO DESCRIPTIONS RATED BY BLIND AND VISUALLY IMPAIRED USERS

In Experiment 2, twelve blind and visually impaired participants were asked to rate the videos created by the describers in Experiment 1 (akin to [46]). These participants will be referred to as the *raters*. The typical method reported by raters to access images were audio description (9 participants), magnification (2 participants), and tactile graphics (1 participant). The frequency of general use of video descriptions, were daily

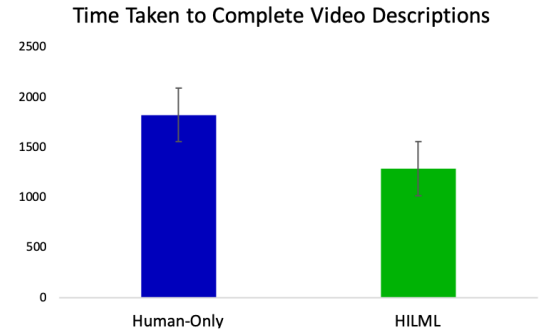

Figure 8. Mean and standard error of the time taken (in seconds) to complete the video descriptions between the HILML (green) and Human-Only (blue) conditions. The mean time taken was significantly less in the HIML condition than the Human-Only condition ($p < 0.01$).

or almost daily (5 participants), several times a week (2 participants), a few times a month (3 participants), a few times a year (1 participant), and never (1 participant).

Experimental Design

All raters worked remotely and were each given half of the 44 videos described (as each of the 22 describers had produced 2 videos). Six raters were given videos produced by the first 11 describers and six raters were given videos described by the last 11 describers. Therefore, half of the videos provided to each rater were described using the HILML system and half were described by the Human-Only condition. The videos were split evenly into Video 1 and Video 2. The order of the videos provided to raters were randomized. Raters did not know which videos had been produced by the HILML system or by the Human-Only condition. They were not even aware that some videos had been described using the assistance of machine learning technology, they were simply told that the videos had been described by 11 different describers. They were not told which pairs of videos were described by the same describer.

Raters could play each video through and hear the video descriptions being read out by text-to-speech synthesis. After watching each video, raters were asked to rate the quality of description on a scale of 1 to 5 ('Poor' to 'Excellent'), their understanding of the topic in the video, and any other comments they had about the video.

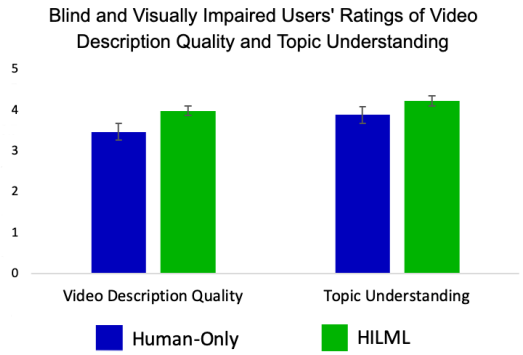

Figure 9. Mean and standard error of ratings (Likert Scale 1 (Poor) - 5 (Excellent)) made by blind and visually impaired users on the quality of video descriptions and understanding of topic in videos described by the HILML (green) and Human-Only (blue) conditions. Both video description quality ($p < 0.01$) and understanding of topic ($p = 0.01$) were rated as significantly higher in the HILML conditions.

Results

Video Description Quality and Topic Understanding

Results of the Shapiro-Wilk test demonstrated that the data was non-parametric. We therefore used Wilcoxon Signed-ranks tests on raters' evaluation on Video Description Quality and Topic Understanding. Results showed that both mean video description quality was rated significantly higher in videos created by the HILML system ($\mu = 3.9394$, $\sigma = 0.5276$) than the Human-Only condition ($\mu = 3.4303$, $\sigma = 0.7328$) ($Z = 2.5897$, $p = 0.0068$, $r = 0.7476$) and that mean understanding of video's topic was rated significantly higher in videos created by the HILML system ($\mu = 4.1720$, $\sigma = 0.6650$) than the Human-Only condition ($\mu = 3.8417$, $\sigma = 0.8026$) ($Z = 2.5001$, $p = 0.0117$, $r = 0.7217$) (Figure 9).

Comments By Raters

The comments provided by the raters were extremely interesting and revealing. Out of the 12 raters, 6 of them used the word "*favorite*", "*excellent*", or "*best*" in their comments of one video. Five out of the 6 were for the HILML system. Examples of such sentences include:

"Good cooking advice, actually describes what TO do and not just what is going on. Would say excellent." [HILML]

"Excellent! Simple, tactile. All around easy to understand! Best so far." [HILML]

"This video is probably the best described, gave info on how to cut the vegetables and exact measurements." [Human-Only]

The level of detail provided in descriptions was very important to raters. Novice describers are known to provide too much detail in video descriptions. Raters made comments such as:

"Too wordy and algorithmic! Would say it is horrible. The time stamp helps NO ONE and distracts from purpose of the video. Fair. "[Human-Only]

"It was not necessary to say [where] the words or items were on the screen and that the words faded in or out." [Human-Only]

"Too specified, too confusing to understand." [Human-Only]

"Unnecessary descriptions may confuse viewers 'cut pieces into even smaller pieces' does not help. Bad." [Human-Only].

"The details were very vague and no text shown on the screen was even mentioned." [Human-Only].

However the HILML system seemed to be better at guiding describers to descriptions that were better received by raters, with comments such as:

"Clear and specific." [HILML]

"This video is an example of the minimum info that should be included in all videos. It is not too overly descriptive with each step. It lays out what is happening and then reads the text for the ingredients." [HILML].

Another category highlighted by raters were the provision of scene identification or shifts by videos that were created using the HILML system. Closely related to this was the alignment of the audio and video, which the automated scene segmentations provided help with.

"Weird transitions, bad descriptions of bowls confused me. Lack of verbal instruction is an issue. Bad." [Human-Only]

"Once again, the pace of the description does not match the pace of the video. The description went faster than the video thus leaving a huge pause in which nothing was being described only music was playing. I have a feeling the video was still going. I think the description was done very poorly." [Human-Only]

"I believe that aligning an audio track with the description is much more beneficial in order to prevent any confusion." [Human-Only].

Examples of positive comments by videos created by the HILML system include:

"There weren't any long pauses in the video and it felt as if the description was done more carefully and with precision to match the pace of the original video." [HILML]

"All the descriptions were there. The narrator announced scene shifts as well." [HILML]

The reading out of text on screen was very important to raters and was highlighted in a number of comments. The HILML system automatically generates text on screen for the human, therefore the lack of text on screen transcription was most common for video descriptions created by the Human-Only condition. Examples include comments such as:

"There were far to many gaps in this description and why tell me there is instruction on the screen but not read it that does not work at all." [Human-Only]

"No text was read and there were a large amount of grammatical errors." [Human-Only]

"The description of this video was done very poorly. The details were very vague and no text shown on the screen was even mentioned." [Human-Only]

"On-screen text is critical, this is a strong video description." [HILML]

'This first video was done very well. The person describing gave plenty of details to be able to understand what was happening. I really liked that he was reading the text that appeared on screen and read it alongside the descriptions." [HILML]

Many raters commented on the text-to-speech synthesis of the video descriptions across both conditions. Traditionally,

in the literature, synthesized video descriptions have been shown to be acceptable to blind and visually impaired users [15, 46]. However, the synthesized voice, while acceptable created mispronunciations of words such "*bowl*":

"Using an actual human as the describer would be better only because this speech synthesizer mispronounces the word 'bowl,' which may confuse some people." [HILML]

DISCUSSION

Results from Experiment 1 showed that novice describers found the HILML system to be easier and quicker to use even though it was only somewhat accurate. The time taken to describe videos was significantly less for the HILML system and describers reported the HILML system to be less mentally demanding, requiring less effort, and to be less frustrating using NASA-TLX measures. Interview data showed that describers found the HILML system to be useful in providing guidelines for the level of description detail and scene segmentation. Experiment 2 showed that blind and visually impaired users found the videos that were described by the HILML system to have a significantly higher quality of video description and understanding of the topic than the Human-Only control condition. The raters' comments reveal a multi-faceted picture as to why this could be.

Level of Description Detail

One of the main features of the HILML system was automatic text generation for video descriptions. We investigate this further by examining the level of detail provided in the video descriptions, which was an important factor for both describers and raters. One of the biggest issues encountered by novice describers, and for the blind and visually impaired users who watch their video descriptions, is that novice describers have no training or knowledge of the level of detail that is required for video descriptions. Video descriptions need to be succinct, clear, and to the point. This can vary based on the type of the video that is being described and can vary with individual preferences across blind and visually impaired users. While too little detail can be provided, usually, novice describers often err on the side of more descriptions.

The quality of the video description was rated significantly higher when created by the HILML system compared to the Human-Only control condition, and this was backed-up by comments from raters. That is not too say that the HILML system level of detail was perfect, but it does suggest that the HILML system is providing some type of guidance on the level of detail for describers through text automation. This was consistent with feedback from describers on the HILML system. On the topic of detail, the HILML system captured the subtitles/text on screen, and it is extremely worthwhile to note that some raters commented on the importance of this.

Scene Segmentation

Another important feature of the HILML system was scene segmentation of the videos, providing novice describers with a framework of where to put their video descriptions. Scene segmentation was also very important to blind and visually impaired raters. The HILML system also seemed to create a more cohesive "*flow*" for raters, as opposed to a speed that was "*rushed*" in the control condition. The alignment of the audio and video tracks and correct division of video segments was commented on raters as being crucial to the understanding of the topic and may well be part of the reason why the HILML system created videos that were rated with significantly higher understanding of the topic than the Human-Only condition.

Text to Speech

The videos were all played back to blind and visually impaired raters using synthesized voices with text-to-speech technology. Several raters commented that while speech synthesis is generally acceptable, there were mispronunciation of words such as "*bowl*" and that the voice did not sound natural, such as missing gaps between sentences. This suggests that the synthesized voice could take away from the clarity of the descriptions.

The raters also highlighted another problem with the text-to-speech technology which was that there was no human reading checking of the text for spelling mistakes and grammatical errors. There were several comments by raters across both videos created by both the HILML and Human-Only conditions of such errors that were being read out incorrectly by the text-to-speech converter. Of course, the describer of the 'Human-in-Loop' part of the system should have been checking for these but volunteer describers either created errors or let text with spelling mistakes and grammatical errors pass through to the blind and visually impaired users. An additional note on this by a blind user who gave informal feedback to the authors was that blind and visually impaired individuals themselves may wish to lend their own voices to pre-generated text for video descriptions.

Increased Speed and Ease of Use for Novice Describers

Describers gave feedback that the HILML system was easier and quicker to use than the Human-Only condition. This was backed up by statistical analysis on time taken to describe videos and NASA-TLX subjective workload assessments. These results are very important for the blind and visually impaired community just for the simple fact that systems that automate text generation and scene segmentation, even if they need human input and editing, still make the job much easier for volunteer describers. If such a system could lower the barrier to entry for volunteer describers, this could start increasing the number of video descriptions online and start decreasing the huge digital divide and accessibility between blind and sighted individuals. We had initially hoped that the HILML system would lower the barrier to entry for describers, which in itself, would have been progress. The fact that the HILML system also increased the video description quality is a step forward on two fronts.

Machine Learning Algorithm

In regards to the state-of-the-art, while the algorithms applied in this paper are not new, this is the first time that they have been applied and tested in the field of automated video descriptions for blind and visually impaired users. This paper demonstrates that such algorithms can be used to aid video description generation for sighted volunteers. The rapid development of such machine learning algorithms could lead to a potentially rich field for automated video descriptions.

We have referred to the technique used in this paper as the human-in-the-loop approach, which is an already established one. However, in this specific case, it is possible to start looking at it instead as a machine-learning-in-the-loop approach, where machine learning is being used as a guiding scaffolding for volunteers' work. This idea of machine learning being used as a supporting mechanism for people's work could become very important in the field of volunteer video description for blind and visually impaired users.

Limitations

It is important to acknowledge that the videos being described in this study had music in the background and text on the screen. There was no speech or dialogue occurring during the videos. Background speech requires a more complex placement of descriptions into the video. Descriptions need to be placed where there is no background speech or sometimes the video will need to be paused for the description to be inserted. This can be challenging for novice describers, especially as they have not been trained on how much detail they should provide. We address this in our future work section with an 'on-demand' HILML approach.

It is also important to acknowledge that the accuracy of the text generation was rated as being mainly neutral by describers. The text-generation machine learning technology is still a work-in-progress. However, despite the neutral accuracy of the machine learning, describers still found the system easier and quicker to use than no help at all, and blind and visually impaired users still found the video descriptions to have a higher quality. This suggests that the HILML system is helping by providing the bigger picture of the video with scene segmentation and the guidance on level of detail.

Microsoft Cognitive Services is not at the point where it can generate connected sentences that tell a story. The task of multi-image cued story generation, such as the visual storytelling dataset challenge, is to compose multiple coherent sentences from a given sequence of images [44]. Such sentences are currently not at the level needed to replace the 'human' of the Human-in-Loop [37], but would improve the speed and accuracy substantially more. We address this in our future work section with a 're-training' HILML approach.

Lastly, we do not know how this approach may affect professional describers. It would certainly be interesting to investigate. There is a phenomenon in learning called *the expertise reversal effect* whereby instructional techniques that are beneficial to beginners can have the reverse effects on more experienced learners [41, 42]. It may be that such as system would have to redesigned for the level of expertise of the describer.

FUTURE WORK

We are currently working on improving the accuracy and quality of the machine-generated descriptions. To this end, the discrepancy between the machine-generated and revised narrations are recorded and the revised versions are being used as input to retrain and improve the accuracy of the machine learning model. We also plan to train a visual storytelling model [38] to generate a more comprehensive narrative of the scene with spatiotemporal coherency.

Informal feedback from visually impaired and blind users identified that 'on-demand descriptions' would be helpful when the user pauses the video to ask questions using natural language. For on-demand descriptions, we will train visual question answering [5] and visual dialog [19] models to hold a meaningful dialog with blind and visually impaired users in natural language about the content of the video.

Furthermore, we plan to integrate the HILML system into the YouDescribe website. As Experiment 1 with describers showed, the HILML system was easier and quicker to use than the Human-Only condition. We will thus investigate whether our HILML system will reduce the barrier-to-entry for volunteer describers on YouDescribe and increase the number of video descriptions online.

Finally, we will compile and release a standardized, well-structured video description data set under a free, open-source license to help the machine learning and vision communities advance the state of the art in video understanding.

CONCLUSION

We designed and developed a Human-in-the-Loop Machine Learning (HILML) approach to automatically generate video descriptions and scene segmentation for online videos. Human describers could then edit the automated text generated by machine learning. Our aim was to improve video accessibility by increasing video description ease and speed for novice describers as well as increasing video description quality for visually impaired and blind users. We compared the HILML system to a Human-Only control condition with no machine learning assistance with novice describers. Results showed that the HILML system was significantly faster and easier to use, with comments from describers pointing to guidance provided by the HILML system on level of text detail and scene segmentation. Blind and visually impaired users rated the videos described the HILML and Human-Only conditions (with no knowledge of the conditions). Results showed that they rated the videos created by the HILML system as having significantly higher quality of video description as well as significantly higher understanding of the topic presented in the video than the Human-Only condition. Extensive comments provided by blind and visually impaired users indicate the importance of several factors discussed in this paper.

These promising results, along with the rapid development of deep learning techniques, suggest that such technology can soon be used by many sighted volunteers around the world in the creation of video descriptions for blind and visually impaired users. Our hope is that this will start decreasing the great digital divide between blind and sighted users.

ACKNOWLEDGMENTS

We thank the LightHouse for the Blind and Visually Impaired and their blind and visually impaired volunteers for their invaluable discussion and feedback. We gratefully acknowledge Craig Newmark Philanthropies, and San Francisco State University Center for Computing for Life Sciences for funding this work.

REFERENCES

[1] MAGpie 2.0. Accessed date 2019-07-17. (Accessed date 2019-07-17). http://main.wgbh.org/wgbh/pages/ncam_old/webaccess/magpie/

[2] 3play Media. Accessed Date 2019-09-15. (Accessed Date 2019-09-15). https://www.3playmedia.com/

[3] Nayyer Aafaq, Syed Zulqarnain Gilani, Wei Liu, Ajmal Mian, and Mubarak Shah. 2018. Video Description: A Survey of Methods, Datasets and Evaluation Metrics. *CoRR* abs/1806.00186 (2018).

[4] National Aeronautics and Space Administration (NASA). Accessed date 2019-07-17. NASA-TLX. (Accessed date 2019-07-17). https://humansystems.arc.nasa.gov/groups/TLX/tlxapp.php

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV*. 2425–2433.

[6] Grigorios Antonellis, Andreas G. Gavras, Marios Panagiotou, Bruce L. Kutter, Gabriele Guerrini, Andrew C. Sander, and Patrick J. Fox. 2015. Shake Table Test of Large-Scale Bridge Columns Supported on Rocking Shallow Foundations. *Journal of Geotechnical and Geoenvironmental Engineering* 141, 5 (2015), 04015009.

[7] Microsoft Azure. Accessed date 2019-07-17. (Accessed date 2019-07-17). https://azure.microsoft.com/en-us/services/media-services/video-indexer/

[8] David Bar-El, Thomas Large, Lydia Davison, and Marcelo Worsley. 2018. Tangicraft: A Multimodal Interface for Minecraft. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 456–458.

[9] Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell Waggoner, Song Wang, Jinlian Wei, Yifan Yin, and Zhiqi Zhang. 2012. Video In Sentences Out. (2012).

[10] United States Access Board. 2015. Proposed Information and Communication Technology (ICT) Standards and Guidelines. (2015). https://www.access-board.gov/guidelines-and-standards/communications-and-it/about-the-ict-refresh/proposed-rule

[11] Rupert RA Bourne, Seth R Flaxman, Tasanee Braithwaite, Maria V Cicinelli, Aditi Das, Jost B Jonas, Jill Keeffe, John H Kempen, Janet Leasher, Hans Limburg, and others. 2017. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *The Lancet Global Health* 5, 9 (2017), e888–e897.

[12] Matthew Brand. 1997. The" Inverse hollywood problem": from video to scripts and storyboards via causal analysis. In *AAAI/IAAI*. Citeseer, 132–137.

[13] Carmen Branje, Susan Marshall, Ashley Tyndall, and Deborah Fels. 2006. LiveDescribe. *AMCIS 2006 Proceedings* (2006), 368.

[14] Claude Chapdelaine. 2012. Specialized DVD player to render audio description and its usability performance. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 203–204.

[15] Claude Chapdelaine and Langis Gagnon. 2009. Accessible videodescription on-demand. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 221–222.

[16] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *In Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST-8* (2014).

[17] World Wide Web Consortium. 2018. Web Content Accessibility Guidelines 2.1. (2018). https://www.w3.org/TR/WCAG21/

[18] Frederico da Rocha Tomé Filho, Pejman Mirza-Babaei, Bill Kapralos, and Glaudiney Moreira Mendonça Junior. 2019. Let's Play Together: Adaptation Guidelines of Board Games for Players with Visual Impairment. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI (CHI '19)*. 631:1–631:15.

[19] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

[20] Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. 2013. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

[21] Benoît Encelle, Magali Ollagnier Beldame, and Yannick Prié. 2013. Towards the usage of pauses in audio-described videos. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. 31.

[22] 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 *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 123–130.

[23] Be My Eyes. Accessed date 2019-07-17. (Accessed date 2019-07-17). https://www.bemyeyes.com/

[24] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

[25] Diamantino Freitas and Georgios Kouroupetroglou. 2008. Speech technologies for blind and low vision persons. *Technology and Disability* 20, 2 (2008), 135–156.

[26] Giovanni Fusco, Ender Tekin, Richard E Ladner, and James M Coughlan. 2014. Using computer vision to access appliance displays. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 281–282.

[27] Langis Gagnon, Claude Chapdelaine, David Byrns, Samuel Foucher, Maguelonne Heritier, and Vishwa Gupta. 2010. A Computer-Vision-Assisted System for Videodescription Scripting. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. pp. 41–48.

[28] Langis Gagnon, Samuel Foucher, Maguelonne Heritier, Marc Lalonde, David Byrns, Claude Chapdelaine, James Turner, Suzanne Mathieu, Denis Laurendeau, Nath Tan Nguyen, and others. 2009. Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Universal Access in the Information Society* 8, 3 (2009), 199–218.

[29] Eitan Glinert and Lonce Wyse. 2007. AudiOdyssey: An Accessible Video Game for both Sighted and Non-sighted Gamers. In *Proceedings of the Conference on Future Play*. 251–252.

[30] Deepak Gopinath, Siddarth Jain, and Brenna D Argall. 2016. Human-in-the-loop Optimization of Shared Autonomy in Assistive Robotics. *IEEE Robotics and Automation Letters* 2, 1 (2016), 247–254.

[31] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

[32] Patrick Hanckmann, Klamer Schutte, and Gertjan J. Burghouts. 2012. Automated Textual Descriptions for a Wide Range of Video Events with 48 Human Actions. *Lecture Notes in Computer Science* (2012), 372–380.

[33] SG Hart and others. 1988. Development of NASA-TLX: Results of empirical and theoretical research." inP. A. Hancock and N. Meshkati (eds.), Human Mental Workload. (1988).

[34] Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[35] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M Pintea, and Vasile Palade. 2016. Towards interactive Machine Learning (iML): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In *International Conference on Availability, Reliability, and Security*. Springer, 81–95.

[36] Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao Huang. 2019a. Visual Story Post-Editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6581–6586. DOI: http://dx.doi.org/10.18653/v1/P19-1658

[37] Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao 'Kenneth' Huang. 2019b. Visual Story Post-Editing. (2019).

[38] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual Storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 1233–1239. DOI: http://dx.doi.org/10.18653/v1/N16-1147

[39] Common Objects in Context (COCO). Accessed date 2019-07-17. (Accessed date 2019-07-17). http://cocodataset.org/#home

[40] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0.1: the Winning Entry to the VQA Challenge 2018. (2018).

[41] Slava Kalyuga. 2003. The expertise reversal effect. *Educational psychologist* 38, 1 (2003), 23–31.

[42] Slava Kalyuga. 2007. Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review* 19, 4 (2007), 509–539.

[43] Bongjun Kim and Bryan Pardo. 2017. I-SED: An Interactive Sound Event Detector. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, USA, 553–557. DOI: http://dx.doi.org/10.1145/3025171.3025231

[44] Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation. (2018).

[45] Masatomo Kobayashi, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa. 2009. Providing synthesized audio description for online videos. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 249–250.

[46] Masatomo Kobayashi, Trisha O'Connell, Bryan Gould, Hironobu Takagi, and Chieko Asakawa. 2010. Are Synthesized Video Descriptions Acceptable?. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 163–170.

[47] Dieter Koller, N Heinze, and Hans-Hellmut Nagel. 1991. Algorithmic characterization of vehicle trajectories from image sequences by motion verbs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 90–95.

[48] Richard E Ladner, Melody Y Ivory, Rajesh Rao, Sheryl Burgstahler, Dan Comden, Sangyun Hahn, Matthew Renzelmann, Satria Krisnandi, Mahalakshmi Ramasamy, Beverly Slabosky, and others. 2005. Automating Tactile Graphics Translation. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 150–157.

[49] ListenByCode. Accessed date 2019-07-17. (Accessed date 2019-07-17). https://www.listenbycode.com/

[50] LiveDescribe. Accessed date 2019-07-17. (Accessed date 2019-07-17).

[51] Lara J Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[52] Microsoft. Accessed date 2019-07-17. (Accessed date 2019-07-17). https://docs.microsoft.com/en-us/azure/media-services/video-indexer/video-indexer-use-apis

[53] Joshua A Miele, Steven Landau, and Deborah Gilden. 2006. Talking TMAP: Automated Generation of Audio-Tactile Maps using Smith-Kettlewell's TMAP Software. *British Journal of Visual Impairment* 24, 2 (2006), 93–100.

[54] Robert Munro. 2020. *Human-in-the-Loop Machine Learning*. MEAP.

[55] United Nations. 2006. Convention on the Rights of Persons with Disabilities and Optional Protocol. (2006). https://www.un.org/disabilities/documents/convention/convoptprot-e.pdf

[56] Devi Archana Paladugu, Zheshen Wang, and Baoxin Li. 2010. On Presenting Audio-tactile Maps to Visually Impaired Users for Getting Directions. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 3955–3960. DOI: http://dx.doi.org/10.1145/1753846.1754085

[57] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[58] Meg Pirrung, Nathan Hilliard, Artëm Yankov, Nancy O'Brien, Paul Weidert, Courtney D Corley, and Nathan O Hodas. 2018. Sharkzor: Interactive Deep Learning for Image Triage, Sort and Summary. (2018).

[59] Venkatesh Potluri, Priyan Vaithilingam, Suresh Iyengar, Y Vidya, Manohar Swaminathan, and Gopal Srinivasa. 2018. CodeTalk: Improving Programming Environment Accessibility for Visually Impaired Developers. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI*. 618.

[60] Pythia. Accessed date 2019-07-17. (Accessed date 2019-07-17). https://learnpythia.readthedocs.io/en/latest/

[61] Melissa Roemmele. 2016. Writing Stories with Help from Recurrent Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*.

[62] Jaime Sánchez and Matías Espinoza. 2011. Audio haptic videogaming for navigation skills in learners who are blind. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. 227–228.

[63] José Francisco Saray Villamizar, Benoît Encelle, Yannick Prié, and Pierre-Antoine Champin. 2011. An Adaptive videos enrichment system based on decision trees for people with sensory disabilities. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*.

[64] Jessica Zeitz Self, Radha Krishnan Vinayagam, JT Fry, and Chris North. 2016. Bridging the gap between user intention and model parameters for human-in-the-loop data analytics. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. ACM, 3.

[65] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[66] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[67] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

[68] Automatic Sync Technologies. Accessed Date 2019-09-15. (Accessed Date 2019-09-15). https://www.automaticsync.com/

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[70] Steven Wall and Stephen Brewster. 2006. Feeling what you hear: tactile feedback for navigation of audio graphs. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 1123–1132.

[71] Yi Wang, Zhiming Luo, and Pierre-Marc Jodoin. 2017. Interactive deep learning method for segmenting moving objects. *Pattern Recognition Letters* 96 (2017), 66–75.

[72] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. 2018. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*.

[73] YouDescribe. Accessed date 2019-07-17. (Accessed date 2019-07-17). `https://youdescribe.org/`