Increasing Video Accessibility for Visually Impaired Users with Human-in-the-Loop Machine Learning

Beste F. Yuksel
Soo Jung Kim
Seung Jung Jin
Human-Computer Interaction Lab
Computer Science Department
University of San Francisco
San Francisco, CA, USA
byuksel@usfca.edu
skim104@usfca.edu
sjin14@usfca.edu

Pooyan Fazli
Umang Mathur
Vaishali Bisht
Ilmi Yoon
Computer Science Department
Yue-Ting Siu
Department of Special Education
San Francisco State University
San Francisco, CA, USA
pooyan@sfsu.edu
umathur@mail.sfsu.edu
vbisht1@mail.sfsu.edu
ilmi@sfsu.edu ysiu@sfsu.edu

Joshua Junhee Lee
Computer Science Department
University of Washington
Seattle, WA, USA
joshua.junhee@gmail.com

Joshua A Miele
The Blind Arduino Project
Smith-Kettlewell Eye Research Institute
San Francisco, CA, USA
joshua.a.miele@gmail.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

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 while allowing humans to edit the output. 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 people are blind [3]. Despite international guidelines and standards [19, 6], there is still a paucity of videos on the Internet that are made accessible through video description for blind or visually impaired (BVI) 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” [2]. A significant bottleneck in video accessibility for BVI users is the time and cost to produce video descriptions professionally. 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 BVI users. We built a video description interface with automated scene segmentation, text generation of images, and text generation of text on screen using machine learning. 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 BVI users to rate the quality and clarity of the video descriptions produced by the two conditions (Experiment 2).

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

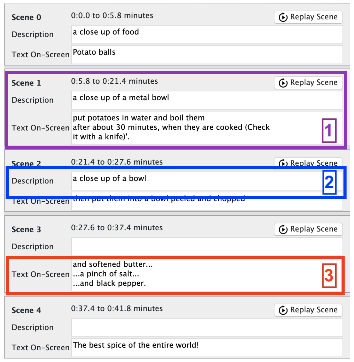
Related Work

Video Description Tools

LiveDescribe was designed for real-time descriptions of live television broadcasts with two automated features to work with silent periods. Feedback from describers was positive but also reflected the high cognitive workload of live description [4]. Villamizar et al. developed a prototype for an adaptive video enrichment system personalized to BVI users based on decision trees [22], however, as stated by the authors, the prototype needs to be evaluated with user studies to investigate the adaptive system further [22]. Gagnon et al. [9, 8] built a prototype software for professional describers that used computer vision to extract visual content from videos. They 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. Kobayashi et al. created audio description synthesis by using external metadata of videos on an online platform [14, 15]. Kobayashi et al. [15] developed a describer script editor with a visual interface to edit the video description sentences. They tested it on one novice and one professional describer. Ratings by blind and visually impaired users showed significantly lower scores for the novice describer [15]. 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 [16]. However, the release of large datasets revealed that these methods cannot cope with the diversity in unconstrained open domain videos and were replaced with deep learning. In particular, Convolutional Neural Networks (CNNs) [1] are the state of the art for modeling visual data recognition [1, 24, 26] and Long Short-Term Memory (LSTMs) [12] are now dominating the area of sequence modeling such as NLP [5, 10, 25]. CNN, Recurrent Neural Network (RNN), or LSTM are used in the encoding stage to learn visual features that are then used in the second stage for text generation (decoding stage). For decoding, different flavours of RNNs are used, such as deep RNN,



Scene 0 0:0.0 to 0:5.8 minutes
Description: a close up of food
Text On-Screen: Potato balls

Scene 1 0:5.8 to 0:21.4 minutes
Description: a close up of a metal bowl
Text On-Screen: put potatoes in water and boil them
after about 30 minutes, when they are cooked (Check
it with a knife)'.

Scene 2 0:21.4 to 0:27.6 minutes
Description: a close up of a bowl
Text On-Screen: then put them into a bowl peeled and chopped

Scene 3 0:27.6 to 0:37.4 minutes
Description:
Text On-Screen: and softened butter...
...a pinch of salt...
...and black pepper.

Scene 4 0:37.4 to 0:41.8 minutes
Description:
Text On-Screen: The best spice of the entire world!

Figure 1: 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).

Bi-directional RNN, LSTM, Gated Recurrent Units (GRU), Attention or Transformer [27]. The resulting description can be a single sentence or multiple sentences. Roemmele et. al used a RNN architecture to generate stories in a sequence-to-sequence manner [21]. Martin et. al proposed an event representation for neural-network-based story generation [17]. Fan et al. created a hierarchical model that automatically generates stories conditioning on the writing prompts [7]. 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 [13].

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 [18]. HILML is an important field to both the HCI and ML communities and can create important collaborations. In the field of image and video classification, Wang et al. [28] 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 CNN to reach similar levels of accuracy as a human with less manual work than human-only analysis [28]. Pirrung et al. [20] 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 BVI users.

System

We built a system to assist novice human volunteers in producing video descriptions using a HILML approach. Our framework uses publicly available APIs, such as Microsoft Azure Cognitive Services [23] and IBM Watson [29] for computer vision and language. These APIs provide capabilities for scene detection, generation of phrase-level captions on individual frames, face and emotion detection, text detection, speech synthesis, and sentiment analysis. We built a model based on this extracted knowledge to generate effective descriptions of video content. These descriptions are minimal and mainly focus on major scene or condition changes, primarily changes in foreground objects, people in the scene, and text.

The following describes the workflow of the system:
Input Data: Videos for which descriptions have been requested are forwarded to the model for processing.
Scene Segmentation and Key Frame Extraction: The video is segmented into a sequence of scenes of varying time spans. Key frames are sampled to maintain the appropriate granularity of the scene for generating the description.
Generating Video Description: 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.
Validating or Revising the Description: Through the describer interface (Figures 1 and 2), sighted users can view video scene segments, their associated descriptions, and can make their own edits and notes in the script editor. Vol-

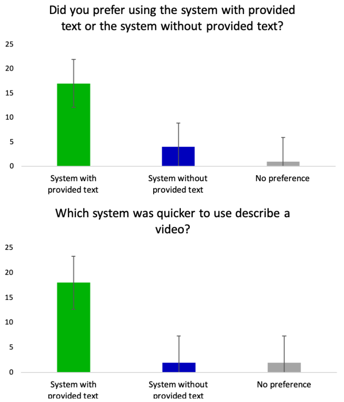
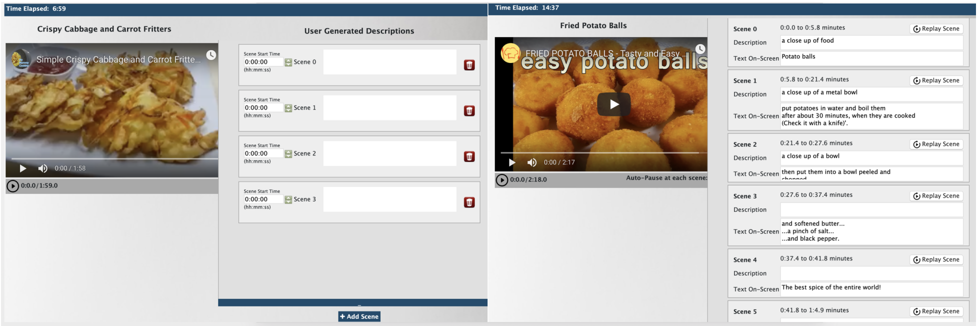


Figure 3: Describers' ratings of the HILML and Human-Only conditions. Describers showed (Top:) noticeable preference for the HILML system and (Bottom:) felt it was quicker to use.


Time Elapsed: 6:59
Crispy Cabbage and Carrot Fritters
Simple Crispy Cabbage and Carrot Fritte...
0:00 / 1:58
0:0.0/1:59.0
User Generated Descriptions
Scene Start Time
(hh:mm:ss)
+ Add Scene
Time Elapsed: 14:37
Fried Potato Balls
FRIED POTATO BALLS - Tasty and Easy ...
easy potato balls
0:00 / 2:17
0:0.0/2:18.0
Auto-Pause at each scene:
Replay Scene

Scene Start Time 0:00:00 (hh:mm:ss) Scene 0

Scene Start Time 0:00:00 (hh:mm:ss) Scene 1

Scene Start Time 0:00:00 (hh:mm:ss) Scene 2

Scene Start Time 0:00:00 (hh:mm:ss) Scene 3

Scene 0 0:0.0 to 0:5.8 minutes
Description: a close up of food
Text On-Screen: Potato balls

Scene 1 0:5.8 to 0:21.4 minutes
Description: a close up of a metal bowl
Text On-Screen: put potatoes in water and boil them
after about 30 minutes, when they are cooked
(Check it with a knife)'.

Scene 2 0:21.4 to 0:27.6 minutes
Description: a close up of a bowl
Text On-Screen: then put them into a bowl peeled and chopped

Scene 3 0:27.6 to 0:37.4 minutes
Description:
Text On-Screen: and softened butter...
...a pinch of salt...
...and black pepper.

Scene 4 0:37.4 to 0:41.8 minutes
Description:
Text On-Screen: The best spice of the entire world!

Scene 5 0:41.8 to 1:4.9 minutes

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.

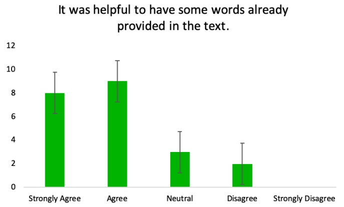


Figure 4: Describers' ratings of the helpfulness of the provided text generated by the HILML system. Most describers agreed or strongly agreed that it was helpful.

unteers 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 1 shows the the scene segmentation, text generation for description of images in the video, and text generation for text that appears on the screen.

Playing Video with Description: 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.

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

Describer Quotes on HILML system from Experiment 1:

“It was super helpful. It was a lot easier to describe what I was looking at just because it already had pre-written text. It helps get it started.”

“I think that one was pretty much set and it again gave us all the details we needed and if we need to fix it we could so I think it made the process a lot easier and quicker.”

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

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

“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.’ ”

“[The accuracy] 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 started from scratch.”

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[1] with the HILML system and Video 2[2] 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 [11] (a subjective workload assessment tool) and were given a short interview.

Results and Discussion

Questionnaire data: Questionnaire data shows that describers preferred the HILML system and felt that it was quicker than the Human-Only condition (Figure 3) even though both systems were generally easy to use for most participants. 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 of ‘neutral’ accuracy. This suggests that even though the accuracy of the HILML system was not perfect, it was *still* helpful (Figure 4) 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) $(t(21) = 2.83, p = 0.005, d = 0.60)$. These findings correspond with the questionnaire data by describers that the HILML system was quicker to use than the Human-Only condition (Figure 3).

	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: 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.

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?*’. [Representative quotes are provided in the margin.] The two main features provided by the HILML system for describers was automatic text generation for video descriptions and scene segmentation. Sixteen out of 22 describers commented on how these features helped provide guidelines for them as novices, particularly on the level of detail required in descriptions. However, four out of 22 describers felt that the descriptions were “*vague*” and “*simple*” (although two of these describers still felt the HILML system did provide them with a starting point or a guideline). Interview data was consistent with questionnaire data on the text generation’s neutral level of accuracy, yet the HILML was still helped the majority of novice describers.

[1] https://www.youtube.com/watch?v=cNj3aOTYdQQ

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

Rater Quotes from Experiment 2 on HILML system and Human-Only Control Condition:

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

“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 System]

“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 System]

“Too specified, too confusing to understand” [Human-Only condition]

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

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

Experiment 2: Video Descriptions Rated By Blind or Visually Impaired Users

Twelve BVI participants were asked to rate the videos created by describers in Experiment 1 (akin to [15]). These participants will be referred to as *raters*. Methods reported by raters to access images were audio description (9 raters), magnification (2 raters), and tactile graphics (1 rater).

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 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. 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.

Results and Discussion

Video Description Quality and Topic Understanding: Wilcoxon Signed-ranks tests showed that mean video description quality was rated significantly higher in videos created by the HILML system ($\mu = 3.9393$, $\sigma = 0.5276$) 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 ($\mu = 3.8417$, $\sigma = 0.8026$) ($Z = 2.5001$, $p = 0.0117$, $r = 0.7217$).

Comments by Raters: The comments provided by the raters were extremely interesting [please see margin for representative quotes]. The level of detail provided in descriptions was very important to raters, with the HILML system providing more concise and succinct descriptions. Novice describers often provided too much detail in the Human-Only condition. Raters highlighted many cases in the Human-Only condition where there was a lack of reading out text on screen. The HILML system automatically generates text on screen, guiding the novice describer in that respect. Correct scene segmentation and alignment of audio and video was also very important to BVI raters with the HILML system creating a more cohesive “*flow*” for raters, as opposed to a speed that was “*rushed*” in the Human-Only condition.

Conclusion

This work demonstrates that the HILML system 1) helps novice describers generate video descriptions for BVI users significantly faster and with increased ease, and 2) generates significantly higher quality video descriptions and understanding than human-only descriptions as rated by BVI users. This provides a hugely needed low barrier to entry method for online video descriptions in order to decrease the massive digital divide between BVI and sighted users. As part of future work, we are developing ‘on-demand descriptions’ to allow BVI users to pause the video and ask questions using natural language.

Acknowledgements

We thank LightHouse for the Blind and Visually Impaired, Craig Newmark Philanthropies, and San Francisco State University Center for Computing for Life Science.

REFERENCES

[1] 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.

[2] 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

[3] 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.

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

[5] 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).

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

[7] 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.

[8] 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.

[9] 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.

[10] 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*.

[11] 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).

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

[13] 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

[14] 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.

[15] 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.

[16] 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.

[17] 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*.

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

[19] 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

[20] 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).

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

[22] 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*.

[23] Microsoft Azure Cognitive Services. Accessed date 2019-09-15. (Accessed date 2019-09-15). https://azure.microsoft.com/en-us/services/cognitive-services

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

[25] 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.

[26] 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*.

[27] 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.

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

[29] IBM Watson. Accessed Date 2019-09-15. (Accessed Date 2019-09-15). https://www.ibm.com/watson