International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

The Current Landscape of Audio Description Research: A Survey

Labdhi Gada[1], Ankit Rajiv Jindal[2], Neelam Jain[3]

[1]Student, Department of Computer Science, SVKM's Mithibai College of Arts, Chauhan Institute of Science & Amrutben Jivanlal College of Commerce and Economics (Autonomous), Affiliated to University of Mumbai, Vile Parle - West Mumbai, India[1]
labdhigada12[at]gmail.com

[2]Chairperson, Friends for Inclusion and CEO, Incluistic Private Limited, Bangalore, India
ankitrajivjindal[at]gmail.com

[3]Head of Department, Department of Computer Science, SVKM's Mithibai College of Arts, Chauhan Institute of Science & Amrutben Jivanlal College of Commerce and Economics (Autonomous), Affiliated to University of Mumbai, Vile Parle - West Mumbai, India
neelam.jain[at]mithibai.ac.in

Abstract: ***Audio description (AD) enables those with visual difficulties to hear what cannot be seen on film, television, and real-time footage. Visually impaired persons could benefit from AD by receiving an audio rendition of the visual content. The task of automatically summarising or subtitling a natural language video is considered one of the most challenging computer vision issues to solve. However, the problem is not minor, especially when a film has several notable events, which often happen in real recordings. This research project aims at providing dense subtitles, which include both - the identification and description of events in a video.***

Keywords: Audio description, captioning, audio captioning, screen captioning, media accessibility

1. Introduction

The field of computer science has seen a growing interest in the development and use of audio description in recent years. Audio description (Fryer 2016)[7], also known as descriptive audio or video description, is a narration track that provides additional information about visual elements in a video, such as actions, characters, costumes, and settings. It is typically used to make media accessible to people who are visually impaired or blind.

One of the main challenges in the field is the lack of audio description in digital media. While audio description is widely used in traditional media, such as television and cinema, it is not yet widely available in digital media. Despite the advancements in technology, visually impaired individuals still face barriers when trying to access and engage with digital media, such as films, television shows, video games, and virtual reality experiences.

Therefore visually impaired individuals are unable to fully engage with digital media, and cut out from important cultural and societal experiences. Additionally, the process of adding audio description to digital media is time-consuming and expensive, which can be a barrier for content creators.

Automating the process of generating audio description is one potential solution, but current automatic generation methods often produce low-quality descriptions.

This paper explores the ongoing and recent works in the generation of audio description via various methods, and comments over the progress made.

2. Background

Audio description generation (Fryer 2016)[7] has grown in recent years as technology has advanced and the need for accessibility has become more pressing.

The use of audio description in media dates back to the 1970s, when it was first used in television and cinema to provide additional information for visually impaired individuals. However, the use of audio description in digital media, such as films, television shows, video games, and virtual reality experiences, is a relatively new area of research.

In the early days, the process of adding audio description to digital media was done manually, which was time-consuming and expensive. As a result, the availability of audio description in digital media was limited. With the advancement of technology, researchers began to explore ways to automate the process of generating audio description using natural language processing and machine learning techniques.

In recent years, the field of audio description in computer science has seen a growing interest in the use of audio description in various application areas such as education and virtual reality. Researchers are also exploring ways to improve the quality and effectiveness of audio description, and to evaluate the user satisfaction.

There is a growing interest in the use of audio description in various application areas and the development of efficient methods for automatically generating audio description.

3. Datasets

This section aims to identify and analyse relevant existing datasets as a prerequisite for training a model or conducting exploratory analysis before model synthesis.

Specifically, datasets that consist of video or image data that can be used to train and evaluate audio description models.

a) LSMDC
The Large Scale Movie Description Challenge (LSMDC) (Rohrbach et al. 2017)[17] dataset includes a parallel corpus of 128,118 sentences aligned to video clips from 200 movies, with a total of 118,000 sentence-clip pairs and 158 hours of video. Additionally, the dataset was updated in 2016 to include manual alignment of the training/validation sets, resulting in 101,046 training clips, 7,408 validation clips and a total of 128,000 clips.

The researchers gathered movie scripts used in previous studies and contrasted the descriptions from both sources. They merged the Montreal Video Annotation Dataset (M-VAD) and MPII-MD datasets, eliminating any duplicates, and excluded script-based movie alignments from the validation and test sets of MPII-MD. They utilised the most advantageous features of M-VAD and MPII-MD for the public and blind test sets. The manual alignment caused multiple sentence descriptions to be separated, and the more accurate alignment shortened the average length of the clips.

b) M-VAD
The dataset (Pini, Cornia, and Bolelli 2019)[18] features annotations of characters' physical appearances, including face bounding boxes, and connections with their textual references. The annotations were produced using a semi-automated method, which was applied to identify and annotate characters in each movie's video clip. To include more characters, some inaccuracies in the original M-VAD annotations were corrected. The dataset is limited to face tracks and the annotation process was improved for greater precision in face tracking and to handle characters who have multiple names in the movie.

c) MPII-MD
The MPII Movie Description dataset (MPII-MD) (Rohrbach et al. 2017)[17] contains a parallel corpus of over 68,000 sentences and video snippets from 94 HD movies. It provides transcribed and aligned AD and script data sentences. Despite the potential benefit of ADs for computer vision, they have not been widely used in the field apart from (Yao et al. 2015)[20], with only a few studies focusing on automating AD production (Gagnon et al. 2010)[21] (Lakritz and Salway 2006)[22].

d) TACOS MULTI-LEVEL
This dataset (Rohrbach et al. 2015)[19] provides a high-quality alignment of sentences with video segments to support the grounding of action descriptions in visual information. The dataset features paraphrased descriptions of the same scene, which can be utilised for various purposes, including generating text from videos. Additionally, it provides the alignment of textual activity descriptions with sequences of low-level activities, making it possible to examine the division of action verbs into fundamental activity predicates.

e) TRECVID MED12
The authors (Das et al. 2013)[24] used TRECVID Multimedia Event Detection (MED12) dataset for generating lingual descriptions of real-life videos. The training set includes 25 event categories, each with approximately 200 videos of positive, relevant examples of event descriptions. The descriptions in the training set are brief and general, averaging 10 words, including stopwords. Additionally, the authors utilised a separate dataset from the Multimedia Event Recounting (MER) task, which contains 6 test videos per event for 5 selected events from the 25 events in the MED12 dataset.

f) FLICKR-SOUNDNET
The authors (Wang et al. 2021)[1] (Arandjelovic and Zisserman 2017)[25] used a large unlabelled dataset of videos from Flickr which contains over 2 million videos. They use a random subset of 500k videos for the study, with 400k for training, 50k for validation and 50k for testing. The researchers limited their analysis to only the first 10 seconds of each video, and did not utilise any information outside of the videos themselves.

g) KINETICS SOUNDS
Kinetics sounds dataset (Wang et al. 2021)[1] (Arandjelovic and Zisserman 2017)[25] a subset of the Kinetics dataset, which contains videos manually annotated for human actions using Mechanical Turk and cropped to 10 seconds around the action. The subset consists of 19,000 video clips (15,000 for training, 1,900 for validation, and 1,900 for testing) for 34 human action categories that can be expressed both visually and audibly.

h) ACTIVITY NET
The ActivityNet Captions (Wang et al. 2021)[1](Krishna et al. 2017)[26] dataset contains 20k videos with an average of 3.65 temporally localised sentences per video, resulting in a total of 100k sentences. The dataset features a normally distributed number of sentences per video that grows with the video length. The average length of each sentence, which is also normally distributed, is 13.48 words. The dataset places emphasis on verbs and actions.

4. Live Captioning

Live captioning (Evans 2003)[34] is a technique used to provide real-time information for live events, such as speeches, conferences, webinars, and, as well as screen captioning.

("What is the difference between open and closed captioning? | DO-IT" 2021)[37] Captions can be categorised into the following types:

a) Open Captions
Open captions are a type of captioning that is always present and cannot be turned off. They are integrated into the media content and are always visible to the viewer. For instance, during a movie where characters speak in a language

different from the primary language of the movie, the subtitles may be directly incorporated into the video stream.

b) Closed Captions

Closed captions, on the other hand, offer the viewer the flexibility to turn them on or off as per their preference. These captions are not a part of the media content and can be enabled or disabled by the viewer. A common example of closed captions is the auto-generated speech-to-text feature found on YouTube.

The aforementioned examples are merely illustrations and captions can contain supplementary information that is beneficial in enhancing accessibility for individuals with disabilities.

Within the domain of computer science, research endeavours pertaining to live captioning have primarily centred on the advancement of automatic speech recognition (ASR) algorithms (Prazák et al. 2012) and natural language processing (NLP) methodologies (Mehta, Pai, and Singh 2020) with the aim of producing closed captions.

Additionally, research has also focused on developing techniques for real-time text-to-speech synthesis and speech-to-text conversion to improve the accessibility of live captioning for people who have visual or audio impairment.

This is an active field of research and new techniques and methods are being developed to improve the live captioning experience.

Live captioning can be further bifurcated into two main domains of research, with respect to audio and visual media:

a) Audio Captioning

Audio captioning refers to the process of converting audio content into written text, often in real-time, to provide a text representation of the audio for people who are deaf or hard-of-hearing, or for individuals who may benefit from reading along with audio content, such as language learners. Audio captioning can be achieved through automatic speech recognition (ASR) algorithms, which can transcribe spoken words into written text, or through human-generated captions, which may be more accurate but also more time-consuming. Audio captioning is typically used in a variety of settings, including educational lectures, live events, and television programs.

b) Visual or Screen Captioning

Screen captioning refers to the process of providing written text representations of content excluding audio in multimedia productions, such as movies, television shows, online videos, as well as input that may arise from sensors such as a camera. Screen captions are usually displayed on the screen and are synchronised with the audio content, allowing individuals who are deaf or hard-of-hearing to follow along with the audio and understand the content. Screen captioning can be either open or closed.

5. Common Patterns

This section explores the prevalent trends, potential difficulties encountered, and various opportunities for enhancement in the evaluation of studies that address Audio Description (AD) through computational methods or survey-based approaches.

a) Accessibility for Visual Media and Audio Description on the Web

The authors (Gleason et al. 2020)[2] (Gleason et al. 2019)[3] (Gleason et al. 2020)[4] (Salisbury, Kamar, and Morris 2017)[5] mention that traditionally, alternative text or "alt text" is used to add captions to images but has to be manually curated by authors with no automation assistance. However, recent research has been done on generating alt texts of images (Gleason et al. 2020)[4], memes (Gleason et al. 2019)[3] and GIFs (Gleason et al. 2020)[2], with the efforts on making GIFs accessible being the most relevant to the current research.

The purpose of the research in those papers is to examine longer videos with audio content and develop an automated workflow for generating and evaluating audio descriptions. The authors highlight that audio descriptions enhance accessibility by providing verbal explanations of visual media and note that many studies have been conducted to assess the benefits of audio descriptions on various video types, audiences, and description styles.

However, they also mention that the creation of audio descriptions raises new questions, particularly regarding the massive quantity and types of videos and the editing methods used. The authors believe that a fully automated audio description generation tool can help blind and visually impaired individuals to access online videos more easily, and their goal is to create an early prototype of such a tool and evaluate its advantages and disadvantages, and to provide suggestions for future advancement.

b) Synchronisation between Audio and Visual Elements

This research (Gaver and William 1993)[6] explores the relationship between audio and visual elements in media and the importance of training the listening skills of visually impaired individuals to enhance their understanding of their surroundings. The authors (Gaver and William 1993)[6] mention that when observing a large number of visual-audio combinations, people learn the correlations between visual and audio unconsciously.

Computer vision researchers are examining various methods for attaining audio-visual comprehension, including identifying audio-visual events, enhancing the consistency of human speech, and creating sound that is aware of the scene. The authors have devised an algorithm that identifies disparities between audio and visual components in media and the identified disparities indicate the points in time where audio description results should be incorporated.

c) Automatic Generation of Audio Description for Videos

This research (Wang et al. 2021)[1] (Zhou et al. 2018)[9] (Li et al. 2018)[12] is focused on the field of video description

generation, which is the automatic generation of natural language sentences that describe the visual content of a video. Similar to image captioning, it requires recognizing salient objects, understanding actions and interactions in the video.

The current leading approaches to this issue address it through dense video captioning, which aims to identify all events in time and add captions for each event. These methods break down the problem into two separate tasks: event detection and caption generation.

(Zhou et al. 2018)[9] The authors intend to evaluate the existing algorithms in video description generation and identify the shortcomings, to pave the way for future avenues for audio description research. They propose to use similarity-based modelling to optimise the output from dense video captioning engines, by measuring semantic textual similarity, and to enhance user experiences of video with description by considering factors such as relevance, non-redundancy and non-confusing.

6. Related Work

(Zhou et al. 2018)[9] in their work have proposed a Deep Learning based algorithmic framework which is demonstrated by generating end-to-end text descriptions for proposals. Their end-to-end model is composed of 3 parts:

a) **A Video Encoder:** here, they have used a feed forward neural network to encode each frame of a video into a continuous representation by passing it through encoding layers. The self-attention mechanism allows each output time step to encode all context information. Unlike recurrent models, the path between time steps is only one, making the gradient updates independent of their position in time and making it easier to learn dependencies between distant frames.
b) **Proposal Decoder:** The proposal decoder takes the visual features from the encoder and outputs event proposals. The event proposal decoder is based on ProcNets (Zhou, Xu, and Corso 2018)[27], chosen for its superior performance on dense, long event proposals. The design includes the same anchor-offset mechanism as ProcNets and a set of N explicit anchors for event proposals.
c) **Captioning Decoder:** The Masked Transformer captioning decoder uses information from both the visual encoder and the proposal decoder as inputs. The mask prediction network creates a differentiable mask for a selected event proposal based on the proposal output. To enable the decoder to caption the current proposal, this mask is then applied through element-wise multiplication with the input visual embedding and the outputs from the proposal encoder.

(Krishna et al. 2017)[26] discusses an architecture that jointly localises temporal proposals of interest and then describes each with natural language. They have used LSTMs where the system takes a sequence of video frames as input and outputs a set of sentences, each consisting of start and end times and a set of words with varying lengths. The video frames are first processed by a proposal module to generate proposals, with only those with a high score being passed to the language model for captioning. The language model uses the context from other proposals and the hidden representation of the event proposal module as inputs to generate descriptions for each event.

(Krause et al. 2017)[28] have developed a model using hierarchical recurrent neural networks and generate a natural language paragraph description. It is designed to utilise the compositionality of both images and paragraphs. The image is analysed by identifying objects and areas of interest, then combining the features from these regions to create a combined representation that conveys the meaning of the image. This feature vector serves as input for a hierarchical recurrent neural network composed of a sentence RNN and a word RNN. The sentence RNN determines the number of sentences to generate and produces a topic vector for each sentence, which the word RNN then uses to generate the words for a single sentence.

(Karpathy et al. 2014)[29] explores the use of Convolutional Neural Networks (CNNs) to integrate temporal information in large-scale videos. They examine three ways of combining time information: Early Fusion, Late Fusion, and Slow Fusion. With Early Fusion, time information is integrated at the pixel level within a given time window, allowing the network to detect local motion direction and speed. Late Fusion separates single-frame networks and combines their outputs later in the process, which allows the network to calculate global motion characteristics by comparing the outputs of both networks. Slow Fusion blends the two methods by gradually integrating temporal information throughout the network, allowing higher layers to access more global information in both the spatial and temporal dimensions. This is done by extending the connections of all convolutional layers in time and computing activations using both temporal and spatial convolutions.

(Karpathy and Fei-Fei 2015)[30] The model presented generates natural language descriptions of images and their regions by learning the correspondences between language and visual data. The alignment model uses a combination of Convolutional Neural Networks on image regions, bidirectional Recurrent Neural Networks on sentences, and a structured objective that aligns the two modalities through a multimodal embedding. The model is then based on a Multimodal Recurrent Neural Network architecture, using the inferred alignments to generate novel descriptions of image regions.

(Johnson, Karpathy, and Fei-Fei 2014)[31] discusses an architecture for localising regions of interest in images and describing each region in natural language. The main challenge is to create a model that supports end-to-end training and efficient and effective inference. The architecture is called the "dense captioning task" which combines object detection and image captioning by localising and describing salient regions in images. To tackle this task, the authors propose a Fully Convolutional Localization Network (FCLN) architecture that can process an image in a single efficient pass, requires no external region proposals, and can be trained end-to-end with a single

optimization step. The architecture consists of a Convolutional Network, a dense localization layer, and a Recurrent Neural Network language model to generate label sequences.

(Venugopalan et al. 2015)[32] This framework focuses on creating deep image description models to generate sentences that describe events in videos. It works by first transforming an image into a fixed dimensional vector representation and then using a Recurrent Neural Network (RNN) called Long Short-Term Memory (LSTM) to decode the vector into a sentence. The framework is based on the idea of converting a visual vector into an English sentence and is applicable to both static images and dynamic videos. They use a two layer LSTM model as the sequence model as it has shown exceptional performance in various tasks including speech recognition, machine translation and generating sentence descriptions of images. The first step of the process is to create a fixed-length visual input that summarises a video, which is then decoded by the LSTM to generate textual output.

(Xu et al. 2015)[33] aim to overcome the limitation of previous Convolutional Neural Networks (CNNs) that only perform single-label centred-object classification. Their solution integrates a base CNN into multiple Fully Convolutional Neural Networks (FCNs) to create a multi-scale network capable of detecting multiple objects of varying sizes and locations. The FCNs, which have been used in image segmentation, generate class heat-maps and handle multiple scales. The authors also incorporate a Multiple Instance Learning mechanism to process objects in different positions and scales. The result is a unique end-to-end trainable architecture that combines the multi-scale multi-instance approach with a sequence-to-sequence recurrent neural network to generate sentence descriptions based on visual information.

Table I: Summary of different methods for Description Generation

Author	Purpose	Methodology
Zhou et al. 2018 [9]	To produce end-to-end descriptions for proposals	• Deep Learning oriented algorithmic Framework
Krishna et al. 2017 [26]	To identify describing events and produce captions depending on past events	• Natural Language Processing • Long Short Term Memory Architecture
Krause et al. 2017 [28]	To generate Natural Language paragraph description	• Hierarchical Recurrent Neural Networks • Natural Language Generation
Karpathy et al. 2014 [29]	To produce time stamps for large scale video classifications	• Deep Learning • Convolutional Neural Networks
Karpathy and Fei-Fei 2015 [30]	To generate descriptions of image regions	• Bidirectional Recurrent Neural Networks • Convolutional Neural Networks • Multimodal Recurrent Neural Networks
Johnson, Karpathy, and Fei-Fei 2014 [31]	To localise the regions of interest in image and produce image captioning	• Fully Convolutional Localization Network • Convolutional Networks • Recurrent Neural Networks
Venugopalan et al. 2015 [32]	To produce Natural Language sentences from videos	• Recurrent Neural Network • Long Short-Term Memory
Xu et al. 2015 [33]	To generate Natural Language Description for videos	• Fully convolutional neural networks • Multiple Instance Learning mechanism

7. Future Enhancements

Given the limitations seen in the current body of research this paper suggests the following approaches to be considered when developing future models solving for audio description.

a) Reducing Training Time
This involves having trained over a variety of data, or to develop a model that functions on the zero, one or few-short principle. One can also use pre-trained models conjoined with models that can be trained and the entire model fine-tuned for this task.

b) Accuracy
In a few-short learning paradigm it becomes necessary to pick models that are capable of state-of-the-art generalisations.

In a data-intensive model however, a variety of data must be provided reducing the training process to a sort of classification and prediction process. However, many techniques, such as augmentation, shuffling, etc can be carefully applied even with limited data to enable the model in having better generalisation abilities.

c) Video Processing
In video processing, skip convolutions (Habibian et al. 2021)[8] could be used in the context of spatiotemporal modelling, where both spatial and temporal information are important for understanding the content of a video. The idea behind skip convolutions is to allow information to "skip" over one or more layers of the network, so that the network can learn to preserve and use both low-level and high-level features.

Skip connections can be implemented in several ways in video processing: one way is to concatenate the output of an earlier layer with the output of a later layer, effectively allowing the network to make use of information from multiple scales of the input.

Skip-Convolutions for video processing that leverage the redundancies in video streams to save computational resources. The method represents videos as a series of changes across frames and network activations, called "residuals". They reformulate standard convolution to be more efficient when computed on residual frames by adding a binary gate that decides whether a residual is important for the model's prediction or can be skipped.

Other video processing techniques such as compression, frame skipping could also be done when feeding input to the model.

d) Facial Expressions

Humans being social creatures with a strong visual cortex are sensitive to facial attributes or expressions. Therefore incorporating facial expressions in order to nuance the audio description would greatly enhance the user experience.

Facial expressions can be correlated with emotion recognition or personality trait extraction to provide a higher level of precision in captioning content that is more representative of human experience.

e) Named Entity Recognition

This approach requires the implementation of an AD system capable of recognizing entities present in the world through either audio or video input. To segment entities, existing NLP frameworks can be utilised or, alternatively, knowledge graphs can be employed to facilitate a deeper understanding and generate captions.

8. Conclusion

The main challenge identified is the lack of audio description in digital media, which makes it difficult for visually impaired individuals to fully engage with this type of media and participate in important cultural and societal experiences.

The research suggests that automating the process of generating audio description is a potential solution, but current methods often produce low-quality descriptions. Therefore, there is a need for more research and development in the field to improve the quality and effectiveness of audio description, and to make it more widely available in digital media.

Additionally, the field of audio description in computer science has a growing interest in the use of audio description in various application areas such as education and virtual reality.

Given the modern state of the art techniques, computing power and optimizations it is inevitable that better models for automatic Audio Description generation are on the horizon.

References

[1] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 277, 1–12. https://doi.org/10.1145/3411764.3445347

[2] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20). Association for Computing Machinery, New York, NY, USA, Article 24, 1–10. https://doi.org/10.1145/3373625.3417027

[3] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B Chilton, and Jeffrey P Bigham. 2019. Making Memes Accessible. In The 21st International ACM SIGACCESS Conference on Computers and Accessibility. 367–376. https://doi.org/10.1145/3308561.3353792

[4] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376728

[5] E. Salisbury, E. Kamar, and M. Morris, “Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind”, *HCOMP*, vol. 5, no. 1, pp. 147-156, Sep. 2017. https://doi.org/10.1609/hcomp.v5i1.13301

[6] Gaver, William. (1993). What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. Ecological Psychology. 5. 1-29. https://doi.org/10.1207/s15326969eco0501_1

[7] Fryer, L. (2016). An Introduction to Audio Description: A practical guide (1st ed.). Routledge. https://doi.org/10.4324/9781315707228

[8] Habibian, A., Abati, D., Cohen, T., & Bejnordi, B. (2021). Skip-Convolutions for Efficient Video Processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2695-2704). https://doi.org/10.48550/arXiv.2104.11487

[9] Zhou, L., Zhou, Y., Corso, J., Socher, R., & Xiong, C. (2018). End-to-End Dense Video Captioning With Masked Transformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.48550/arXiv.1804.00819

[10] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. ACM Computing Surveys (CSUR) 52, 6 (2019), 1–37. https://doi.org/10.1145/3355390

[11] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In Proceedings of the IEEE International Conference on Computer Vision. 4634–4643. https://doi.org/10.48550/arXiv.1908.06954

[12] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing

events for dense video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7492–7500. https://doi.org/10.48550/arXiv.1804.08274

[13] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. In Proceedings of the IEEE conference on computer vision and pattern recognition. 6504–6512. https://doi.org/10.48550/arXiv.1611.07675

[14] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. " Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–13. https://doi.org/10.1145/3313831.3376404

[15] Agnieszka Walczak and Louise Fryer. 2017. Creative description: The impact of audio description style on presence in visually impaired audiences. British Journal of Visual Impairment 35, 1 (2017), 6–17. https://doi.org/10.1177/0264619616661603

[16] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7190–7198. https://doi.org/10.48550/arXiv.1804.00100

[17] Movie Description, Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, Bernt Schiele, IJCV 2017 https://doi.org/10.48550/arXiv.1605.03705

[18] Pini, S., Cornia, M., Bolelli, F. et al. M-VAD names: a dataset for video captioning with naming. Multimed Tools Appl 78, 14007–14027 (2019). https://doi.org/10.1007/s11042-018-7040-z

[19] A. Rohrbach, M. Rohrbach, N. Tandon and B. Schiele, "A dataset for Movie Description," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3202-3212, https://doi.org/10.1109/CVPR.2015.7298940

[20] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Video description generation incorporating spatio-temporal features and a soft-attention mechanism. arXiv:1502.08029v3, 2015 https://doi.org/10.48550/arXiv.1502.08029

[21] L. Gagnon, C. Chapdelaine, D. Byrns, S. Foucher, M. Heritier, and V. Gupta. A computer-vision-assisted system for videodescription scripting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2010. https://doi.org/10.1109/CVPRW.2010.5543575

[22] Lakritz and Salway. The semi-automatic generation of audio description from screenplays. Technical report, Dept. of Computing Technical Report, University of Surrey, 2006.

[23] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, Manfred Pinkal; Grounding Action Descriptions in Videos. Transactions of the Association for Computational Linguistics 2013; 1 25–36. https://doi.org/10.1162/tacl_a_00207

[24] P. Das, C. Xu, R. F. Doell and J. J. Corso, "A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching," 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 2634-2641,

[25] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In ICCV. 609–617. https://doi.org/10.48550/arXiv.1705.08168

[26] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision. 706–715. https://doi.org/10.48550/arXiv.1705.00754

[27] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. AAAI, 2018. https://doi.org/10.48550/arXiv.1703.09788

[28] Krause, Jonathan & Johnson, Justin & Krishna, Ranjay & Fei-Fei, Li. (2017). A Hierarchical Approach for Generating Descriptive Image Paragraphs. 3337-3345. https://doi.org/10.1109/CVPR.2017.356

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1725-1732 https://doi.org/10.1109/CVPR.2014.223

[30] Karpathy, Andrej & Li, Fei. (2015). Deep visual-semantic alignments for generating image descriptions. 3128-3137. https://doi.org/10.1109/CVPR.2015.7298932

[31] Johnson, Justin & Karpathy, Andrej & Li, Fei-Fei. (2014). DenseCap: Fully Convolutional Localization Networks for Dense Captioning. https://doi.org/10.48550/arXiv.1511.07571

[32] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1494–1504, Denver, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1173

[33] Xu, Huijuan & Venugopalan, Subhashini & Ramanishka, Vasili & Rohrbach, Marcus & Saenko, Kate. (2015). A Multi-scale Multiple Instance Video Description Network. https://doi.org/10.48550/arXiv.1505.05914

[34] Evans, M. J., “Speech Recognition in Assisted and Live Subtitling for Television”, R&D White Paper WHP 065, BBC Research & Development, 2003 https://doi.org/10.1007/978-3-642-32790-2_62

[35] Prazák, A., Loose, Z., Trmal, J., Psutka, J. V., & Psutka, J. (2012, September). Novel Approach to Live Captioning Through Re-speaking: Tailoring Speech Recognition to Re-speaker's Needs. In INTERSPEECH (pp. 1372-1375). https://doi.org/10.1007/s11042-019-

[36] Mehta, N., Pai, S. & Singh, S. Automated 3D sign language caption generation for video. Univ Access Inf Soc 19, 725–738 (2020). https://doi.org/10.1007/s10209-019-00668-9

[37] What is the difference between open and closed captioning? | DO-IT (washington.edu)