Universal Access in the Information Society
https://doi.org/10.1007/s10209-018-0634-4

LONG PAPER

CineAD: a system for automated audio description script generation for the visually impaired

Virginia P. Campos[1] · Tiago M. U. de Araújo[2] · Guido L. de Souza Filho[2] · Luiz M. G. Gonçalves[1]

Abstract
Audio description (AD) is an assistive technology that allows visually impaired people to access cinema and understand the story of a movie. Basically, the visual content of the story is told by way of using a voice, narrated during the film gaps of silence. Nonetheless, this assistive technology is not widely used, due to several factors, among them the high cost and time involved in creating such audio descriptions. Towards solving this problem, this work proposes a solution that automatically generates AD scripts for recorded audiovisual content, named CineAD. This solution detects the breaks in the spoken lines in the video receiving the AD and generates these descriptions from the original script and subtitles. Alternatively, the solution can be incorporated into a speech synthesizer or used by an audio description narrator to generate the audio that contains the descriptions. To evaluate the proposed solution, qualitative tests with visually impaired users and audio description narrators are conducted. The results show that the proposed solution can generate descriptions of the most important events in the videos, and therefore, can help to reduce the barriers in accessing video faced by visually impaired, when the script and subtitles are available.

Keywords Audio description · Automated generation · Script · Video accessibility

1 Introduction

Visually impaired people could be excluded from a society that relies heavily on audiovisual content as a source of information and entertainment because of the barriers that they face when accessing content that is encoded in the form of visual information. To reduce these barriers, the audio description (AD) assistive technology is used. Audio description is an important technology, which has been developed to improve access for blind people or any person who does not have access to visual information for some reason. It allows these people to participate, intrinsically, in visual experiences such as cinema via audio descriptions of the images, characters, scenery, and the narration of the actions. To describe the films, the audio description scripts are produced manually by professionals, most often, which requires a large amount of time and money. According to Lakritz and Salway [20], approximately 60 h are required for a professional AD narrator to describe a 2-h movie.

According to Szarkowska [33], "a lengthy preparation process and high production costs are among the greatest obstacles to the wider availability of audio description". Therefore, given this scenario and due to several other impeding factors, AD is not offered in most videos, and therefore, blind people do not have access to this type of cultural information. In the USA, according to the American Council of the Blind (OCB), the audio description offered in DVDs is growing since 2010, though in 2017 only 30% of the DVDs identified had description tracks.[1] It is noticed that the actual percentage of described DVDs could be lower since some of the DVDs could not be located. With regard

✉ Tiago M. U. de Araújo
tiagomaritan@lavid.ufpb.br

Virginia P. Campos
vcampos@dca.ufrn.br

Guido L. de Souza Filho
guido@lavid.ufpb.br

Luiz M. G. Gonçalves
lmarcos@dca.ufrn.br

[1] Federal University of Rio Grande do Norte, 3000, Campus Universitario, Natal, RN, Brazil

[2] Federal University of Paraiba, R. dos Escoteiros, s/n-Mangabeira, João Pessoa, PB, Brazil

[1] http://www.acb.org/adp/dvdsoverview.html.

Published online: 31 August 2018

to the videos of TV series, it is also mentioned that none of them are described in DVD.

Notice that an automated solution for creating AD scripts could be used to facilitate this process helping to overcome the above-cited problems. This solution would make the audio description process more efficient decreasing the production time and related costs. Hence, automated generation is an important solution both for the creation of audio description tracks in videos and for aiding human narrators that manually produce descriptions.

In this direction, the objective of this study is to propose an automated system for generation of audio description scripts for recorded audiovisual content to make them accessible to blind or visually impaired individuals. The solution presents a method of alignment and synchronization between script and subtitles that relates actions to video time gaps. Besides that, it selects script actions using summarization heuristics based on weights to form the AD. The system generates a script that contains descriptions of the actions in the video, which, in its turn, can be used by a professional narrator or by a speech synthesizer software to create the audio description.

Considering the large number of videos that do not have AD when they are distributed, the focus of our work is to verify if the descriptions generated by the system have the potential to improve blind people's understanding of the recorded audiovisual content, especially when a human AD is not available, which occurs in most situations. In addition, we would like to assess whether this solution could also be used by film producers and professionals to reduce the cost and time associated with the process of creating the audio description tracks.

In the remainder of this article, Sect. 2 presents the conceptual basis of audio description in cinema and some studies performed in the topic. In Sect. 3, the proposed system and its components are described in detail. Section 4 contains the experiments performed and the results obtained for the validation of the proposal. Finally, the conclusions and suggestions for future work are presented in Sect. 5.

2 Audio descriptions in cinema

Audio description is an assistive technology for blind or visually impaired individuals to address visual contents in cinema. According to Benecke [3], AD is a technology connected to the visual product that translates the images, plot, characters, scenery, and action into narration that is inserted between the dialogues during the film evolution. In other words, it transforms the visual information into words, i.e., it audibly describes everything that is seen in the movie. This is done in such a way that it does not interfere with the original movie audio. Besides cinema, it is also possible to use AD in the theater, opera, museums, and television, among others.

According to Nunes et al. [25], the creation of audio descriptions requires the development of a script adapted to the visual content being described. The audio description writer is responsible for creating an accurate script with insertion time points and instructions for the readers that contain all of the relevant information for understanding the film, not including unnecessary information. It is important to use as few words as possible in the descriptions to avoid the inclusion of excessive information in the audio.

Currently, the creation of AD scripts requires a substantial amount of time particularly because the process is predominantly made manually. Notice that the use of current technology can significantly aid the insertion and creation of audio descriptions in videos. For this reason, several studies have been conducted to incorporate computational methods in the processes involved. In one of these studies, Encelle et al. [14] give a thorough report about the use of artificial pauses in videos with audio descriptions. The idea is to increase the time available for narration and consequently transmit more information via audio description.

Concerning narration, Kobayashi et al. [19] conducted a study about the use of speech synthesizers to provide the narration of the descriptions. Their study shows that the synthesized descriptions are acceptable independently of the language used. However, their work is best suited for instructional videos where understanding is critical. For entertainment videos, the human narration is better recommended because it allows a more natural and pleasing experience. Fernández-Torné [16] also presents a study on the use and acceptance of voice synthesis in audio description. Experiments showed that users with visual impairments preferred the human voice given the naturalness of voice, however, 94% of participants agree that text-to-speech (TTS) tools are acceptable alternative solutions.

A platform using speech synthesis to add audio description to on-line videos is introduced by Kobayashi et al. [18]. The platform includes a tool for creating AD scripts that provides a graphical interface for editing phrases and specifying the time line when each description should be read. This script editor allows to modify the voice parameters, such as gender and speed. Tests with users show that the synthesized speech is acceptable and could significantly improve the experience compared with videos without AD. The study, however, does not include experiments on the script editing tool.

Chapdelaine and Gagnon [6] perform an accessibility study on a website that provides videos with audio descriptions using an adapted video player called VDPlayer. An accessible site is developed to provide five short films, and the audio descriptions were provided in two presentation modes. In the standard mode, descriptions are inserted in

the dialogue gaps. The extended mode uses all of the audio descriptions produced even if they exceed the duration of the gaps. To automatically generate audio descriptions, two approaches are used: visual information analysis and script analysis.

Regarding the analysis of visual information, several studies use deep learning techniques for recognition and detection of visual elements or events. Convolutional neural network (CNNs) models are being used to recognize objects in images [9, 27, 28]. Other studies use recurrent neural networks' (RNNs) models to generate image textual descriptions [7, 11, 15]. Concerning videos, several works use combinations of the CNN and RNN models to describe videos, such as [11, 26, 29, 34, 35].

Due to the diversity of elements and events in a video, the audio description task for this kind of media becomes complex and, therefore, current solutions that extract information using only the video cannot satisfactorily reach the users' needs, mainly when used in videos of general domain.

Another approach is to use the original script as a source of information for the generation of the audio description. Some works already use the script for other purposes such as discovery of tasks and human action annotations through the alignment between the information in the script and the videos [12, 21, 22], and creation of data sets through alignment between screenplay and video [4, 8, 22, 30].

The creation of audio descriptions using scripts is addressed by Lakritz and Salway [20]. A semi-automatic system extracts pertinent information from the script using the most frequent words and converts it to a language that is more appropriate for audio description. The results show that this approach can extract 80% of the important information and converts 66% of the phrases into audio descriptions. The system is evaluated by professional narrators, who indicate that the solution does not improve the efficiency of the AD script creation process because the output does not contain timing information (or synchronization points) that links the AD to the video. According to the authors, the main problem is that the manual identification of the intervals where the AD should be added followed by a searching for the correct description for each time interval requires a substantial effort. However, there are tools with silence and sound-detection features that can support this task, such Audacity[2] and Cubase,[3] as well as works that could be applied to aid in the analysis and classification of audio, such as Giannakopoulos [17] and Wang et al. [36].

The script is previously elaborated to video recording and, because of this, some divergences between the story and the video itself eventually occur. This may give rise to imprecise descriptions. Despite of this issue, the script has a high potential to be used as a source of information because it contains the most important events in the movie. Considering situations where the AD is not present, the script appears as a good source of information. Scripts can be obtained with the media itself or with the video producers.

3 CineAD system

The CineAD proposal seeks to reduce the accessibility barriers that blind and visually impaired persons have when watching videos, thus being a solution for the automated generation of audio description scripts. CineAD creates audio descriptions by analyzing two major elements: the video script and the subtitles. A schematic view of the proposed solution is shown in Fig. 1. Basically, the original screenplay of the video is first analyzed, and its major elements are extracted, including scene titles, actions, and characters, among others. Next, the gap identification component detects the time intervals between lines of the video dialogue, which constitute the potential gaps without speech that are candidates for subsequent insertion of the audio descriptions. In the sequence, the summarization component extracts the most important sentences from the screenplay and thus summarizes the original script by removing secondary information that is less important for audio description. Finally, the AD script generation component creates the audio description script by placing the sentences extracted in the summarization step into the gaps detected by the gap identification step. Each component is described in more detail in Sects. 3.1–3.4.

3.1 Script parser

This component is responsible for reading and extracting the elements of the script, such as scene titles, actions, dialogue, and characters. To facilitate this task, we use a digital format, which allows the identification and annotation of the script elements. In the present study, the *.celtx* format, which is defined for the CELTX tool, is used. However, it is possible to extend this module to support other formats.[4] The CELTX program uses the *.celtx* format, which is a compressed folder with four files: two in the *rdf* format containing metadata and two in the *html* format containing the scripts and other information such as annotations and records. In the CELTX file, it is possible to write scripts and to identify each element with a specific annotation. The script parser component reads the files contained in the *.celtx* format and extracts the script elements, such as scene title, characters, dialogues and

[2] https://www.audacityteam.org/.

[3] https://www.steinberg.net/en/products/cubase/start.html.

[4] https://www.celtx.com.

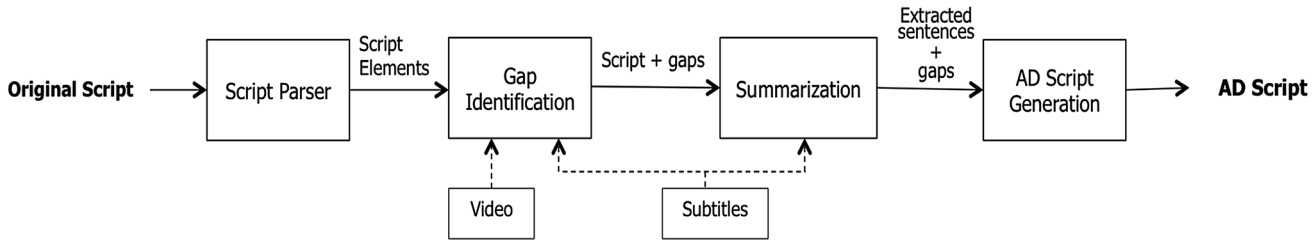

Fig. 1 Schematic view of CineAD

actions. The output of the reading and extraction process is a group of elements sorted chronologically in the sequence of the original script.

3.2 Gap identification

The gap identification component seeks to detect the time intervals in the video that do not contain personal dialogue. Gaps are spaces in the original audio that serve as candidates for audio description tracks because they do not interfere with the lines in the video. We identify the gaps by looking at the subtitles. Each identified gap has a start time and an end time associated with the intervals without words, which thus serve as a synchronization point for inserting the AD tracks, as shown in Fig. 2.

One difficulty with this approach is the detection of the end of the last gap because this information is not in the subtitles and the end of the subtitles does not necessarily coincide with the end of the video. Thus, the last gap, which is the interval between the last spoken line and the end of the video, is obtained by analyzing the video duration, which represents the end of the last gap.

3.3 Summarization component

Unlike the script, which contains all of the descriptions for the film, the audio description script does not contain all of the information present in the original script because it is limited to the spaces without dialogue (the gaps). Therefore, the information in the original script should be summarized to be included in the audio description. To perform this task, the summarization component extracts the most important sentences from the script and discards the less relevant ones.

One method for summarizing is to extract sentences, which involves concatenation of several sentences from the material to be summarized as they appear and without changes [24]. According to Edmundson [13], the most important information in a text can be identified by the frequency of the keywords contained in the text. Therefore, finding the most frequent words allows to detect the most important sentences in a document.

This approach is used, for example, to extract sentences in a semi-automatic audio description method proposed by Lakritz and Salway [20]. This extraction method, which is based on word frequency, returned 80% of the important descriptions, demonstrating that the method can extract relevant information for audio description. Considering these results, the solution proposed in this study uses an extended version of the method by Lakritz and Salway [20] to extract the sentences.

Sentence extraction in CineAD is performed in several steps. First, three lists of high-frequency words are created: (1) the most frequent words in audio description scripts; (2) the most frequent words in the script that will be described; and (3) a list with the names of all of the characters.

Salway et al. [31] presented the thirty (30) most common words in audio description scripts, identified by analyzing the AD script of various film genres. The first list contains

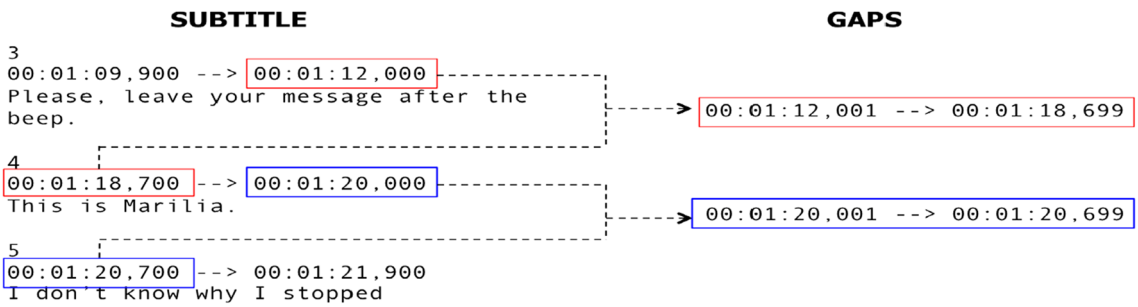

Fig. 2 Detecting intervals without dialogue (gaps) using subtitles

these most frequent words presented by Salway et al. [31], except for the words 'tom' and 'john', which were excluded because they are common names of characters among the scripts evaluated by the authors.

The second list contains the 30 most frequent words in the current script (i.e., the script that the AD is being created). Thus, we first need to remove stop words, such as articles, pronouns, interjections, adverbs, and prepositions. To create this list, we used a text search library, named Apache Lucene,[5] which returns the terms that occurred most often in the document. Lucene contains a specific analyzer for the Brazilian Portuguese language containing the stop words in the language.

Due to the variations of verbal timing and conjugations in the Portuguese language, Lucene only considers the root of the words to identify the most common words. This guarantees that the various forms of the word are considered. For example, for the verb *to run*, only the root *run* is considered; thus, any version of the word (*runner*, *running*, *ran*) are included. This behavior is also applied in the first list of frequent words.

Finally, the third list contains the names of all of the characters in the script of interest. This information is obtained by reading the *.celtx* file metadata, which contains the number of characters as well as the name and description of all the parts in the script. After creating the lists, all actions in the script are broken down into sentences. The selection of sentences for the audio description script is performed as follows. For each sentence, the presence of at least one of the words in the three lists is verified. If a sentence contains, at the minimum, one word from the frequent word lists, it is a candidate for the audio description script. Figure 3 shows an example of candidate sentence extraction.

Next, a weighted average is calculated for each sentence based on the number of words that it contains from each of the three lists. This average is a score that quantifies the importance of the sentence based on the number and type of frequently used words in the script that are contained in the sentence. Formally, this score is defined as:

$$\text{Score} = \frac{\left(\text{qty}_{\text{list1}} \times 2\right) + \left(\text{qty}_{\text{list2}} \times 1\right) + \left(\text{qty}_{\text{list3}} \times 3\right)}{\text{total\_number\_of\_words\_in\_the\_sentence}}, \quad (1)$$

where $\text{qty}_{\text{list1}}$, $\text{qty}_{\text{list2}}$ and $\text{qty}_{\text{list3}}$ represent the number of words in lists 1, 2 and 3 contained in the sentence, respectively.

According to Eq. 1, we can observe that each list has a different weight. In the third list, the character names receive the greatest weight (weight = 3) due to its importance in the video. This is because when a script sentence refers to a character it generally describes an action performed by or involving this character. The second list, which contains words commonly found in the video, receives the lowest weight (weight = 1) due to the high frequency of the word occurrences in the film. Finally, the list with common words in the AD receives a weight of 2.

The result of this process is a group of candidate sentences extracted from the script, where each sentence corresponds to a score. This score is used by the next component to generate the AD script file.

3.4 Creation of the audio description script

The script generation component creates the audio description script and its timestamps based on the information generated by the previous components. More specifically, the goal of this component is to place the sentences extracted in the summarization step into the intervals without words (gaps) identified by the gap identification component. This component is executed in four steps: (1) verification of the similarity between the subtitles and the script dialogue, (2) extraction of actions between the dialogues, (3) division of the gaps for action insertion and (4) generation of a file containing the AD script.

The schematic view for the AD script generation component is shown in Fig. 4, and the function of each of the steps is described in Sects. 3.4.1–3.4.4.

3.4.1 Verification of similarity

The first step shows the similarity of the subtitles with the dialogue in the script that links the sentences (or groups of sentences) in the subtitles to the dialogue of a character in the script. This verification is necessary because the script dialogue can be associated with several sentences (or groups of sentences) in the subtitles. An example is shown in Fig. 5.

To link the sentences of a subtitle to each line of the dialogue in the script, a similarity calculation is performed using the Apache Lucene library. The results of this step makes it possible to have pairs of information from subtitles and script dialogues linked to specific video times. Thus, the script dialogue gains time markings.

3.4.2 Selection of actions between dialogue

After mapping the dialogue, the following step uses the actions selected by the summarization component that are chronologically located between the dialogue in the script. Notice that several actions pertinent to the story can occur between two lines in the film, as shown in Fig. 6. Thus, in this step, we select the candidate sentences between two

[5] http://lucene.apache.org/.

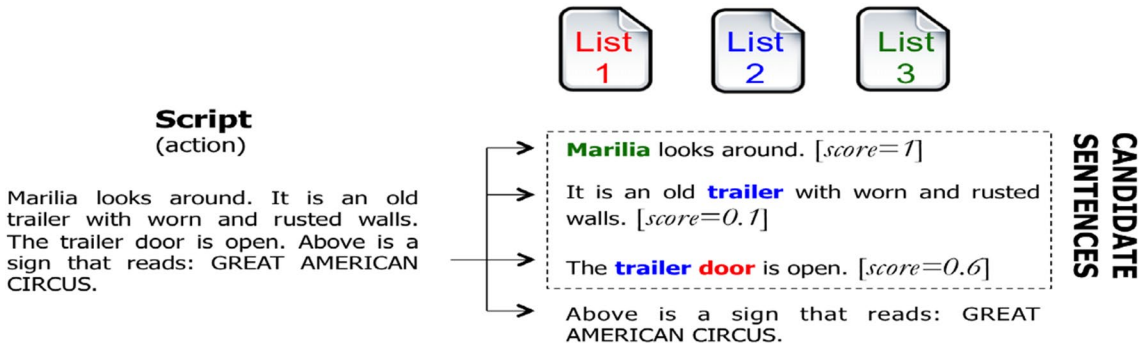

Fig. 3 Candidate sentence extraction process

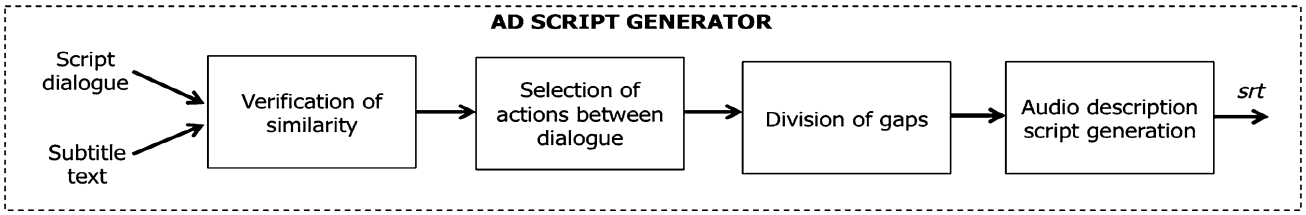

Fig. 4 AD script generation flowchart

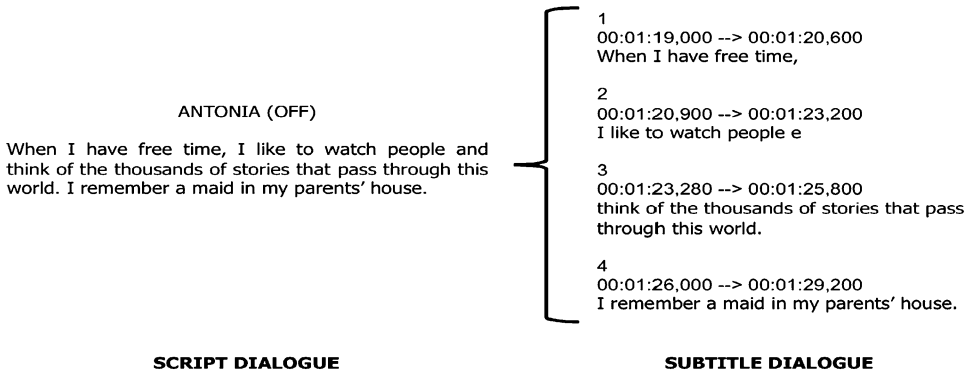

Fig. 5 Example of the similarity between the dialogue and subtitle script

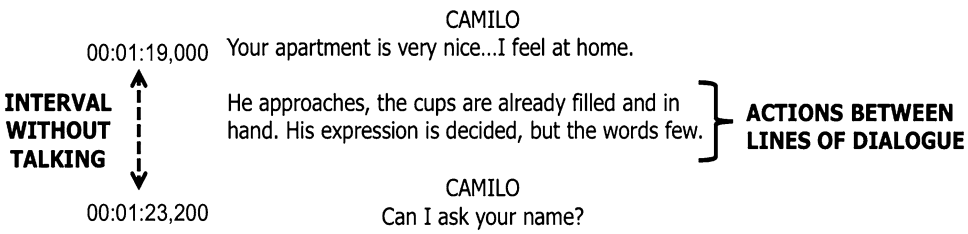

Fig. 6 Example of actions between lines of a script

lines of dialogue; i.e., there is a chronological association between the identified gaps and the candidate sentences found within the gap. This step is based on the information produced by the summarization component and the gaps without spoken lines identified by the gap identification component.

3.4.3 Division of gaps

Next, in the gap division step, the available time for audio description is divided between the candidate sentences. Eventually, not all of the candidate sentences selected by the summarizer will be part of the AD script because the gap may not fit all pre-selected sentences. In addition, Nunes

et al. [25] emphasize that it is important that only a few words are used in the audio description to avoid excessive audio information. Therefore, the duration of the gap defines which candidate sentences are to be part of the final script. To perform this task, the component identifies the duration of the interval to be used to insert the selected actions. The time interval is then divided based on the number of words that can be narrated in that interval. According to Araújo [2], people can read about 180 words per min, i.e., three words per second, on average in a subtitle. Thus, in the present study, we use this time parameter to divide the gaps.

With this, the number of words in each gap is defined by the length of the interval. For example, a 20-s gap can contain up to 60 words. Thus, the phrases narrated in this interval should contain, in total, less than 60 words. Any candidate sentence that is outside this limit can be excluded from the AD script. The candidate sentences that will be inserted into the gaps are determined from the score calculated by the summarization component. This score defines the level of importance that this sentence has within the script based on the most frequent words it contains. Thus, it is possible to rank the sentences by their importance in the script. The gaps are divided by inserting the most important phrases, i.e., those with the highest individual scores, that together contain the maximum number of words that can be narrated within the gap, as determined (empirically) by the limit of three words per second. In general, the sentence with the highest score has the preference, though if it is too long for the gap space, the others with the highest score will then be inserted into the AD.

As an example, consider two candidate sentences in the same gap with the following characteristics, respectively: 50 words—score 0.8 and 30 words—score 0.6. If we have a 20-s gap, whose total capacity is 60 words, only the first sentence (with 50 words) will be selected because it has a higher score. In this case, the second sentence (with 30 words) will be discarded because it does not fit the capacity of the gap. However, if this gap is 15 s (with 45 word capacity), the second sentence will be selected. This happens because, although the first sentence has a higher score, the amount of words is too high to be narrated in this range, then the sentence with the second highest score and the number of words compatible is selected, in this case, the second.

Consider another example in which we have three candidate sentences: (1) 40 words–score 2; (2) 30 words—score 1.2; and (3) 15 words—score 1. If we have a 20-s gap, the sentences 1 and 3 will be inserted. The first sentence has the highest score within the gap, so it will be inserted and occupies 40 words in space. The second sentence, despite having a score greater than the third, will be excluded because it has a greater number of words than the remaining space, which is 19 words. The third sentence is inserted by having the third best score and having a quantity of words that fits into space. This gap will present in total a narration of two sentences accumulating 55 words.

In this step, we insert the sentences in the gap based on their chronological order as found in the script. Therefore, ranking by the score allows the most important sentences to be chosen. Thus, the system synchronize the actions with the video, and thus, the description can be narrated in sync with the occurrence of the event. In addition, we do not overlap the dialogues of video with the descriptions.

3.4.4 Audio description script generation

After placing the actions into their respective gaps, the audio description script is finally created. The last step of the AD script creation component is the creation of an *srt*-formatted file, which is commonly used for subtitles. We choose this format due to its simplicity, and because it is widely used and allows that the descriptions and their timestamps can be represented. Thus, the AD script contains the audio description text and their timestamps (synchronization points). The *srt* file can serve as a input for human audio description narrators or for speech synthesis systems.

An example of an AD script generated by CineAD by combining elements from the subtitles and the original script is shown in Fig. 7. The column on the left shows the subtitle text and timestamps. The middle column shows the video script with the actions and dialogues, and the right column shows the automatically generated audio description script. We can observe that the subtitle lines are related to the dialogue in a manner similar to how the AD script descriptions are related to the original script.

Figure 7 also shows that the timestamps of the AD script are the inverse of the subtitle timestamps. In addition, the actions present in the script but not in the audio description were removed by the summarization component because they did not contain frequently used words or in the gap division step because they did not have sufficient time available for their narration.

An example of the AD generated using this implementation can be seen in a video posted at YouTube (https://youtu.be/k7cTWc5Ox2U).

4 Experiments

To evaluate the proposed solution, some tests were conducted with CineAD from two points of view: visually impaired users and professional audio description narrators. The first part consists of an evaluation of films containing audio descriptions generated by the proposed solution by visually impaired users to analyze their understanding of the video contents. Thus, it is possible to investigate whether these users could understand films containing audio

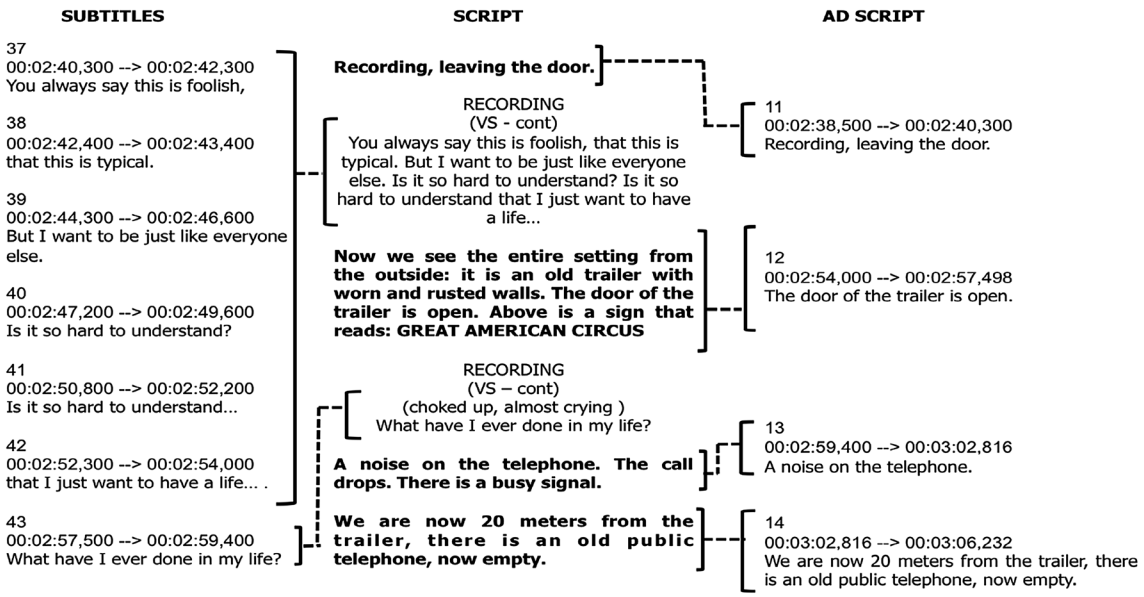

Fig. 7 Example of AD script generation (portion of the script from the movie Trs Minutos, 1999)

descriptions generated by the system, to overcome the barriers related to accessing the information. The second part, more qualitative, consists of an analysis of the audio description script generated by the system to identify strengths and weaknesses of the solution from the perspective of a professional in the field.

4.1 Analysis of the level of understanding

The first part of the experiments analyzes the level of understanding of visually impaired users. The tests were performed at the Paraiba Institute of the Blind and included 12 Brazilian visually impaired (partially or completely blind) participants in the following age groups: 25% below 18 years; 25% between 18 years and 30 years; 25% between 31 years and 40 years; and 25% above 40 years. Regarding the degree of visual impairment, 33.3% users have total blindness, while 66.7% have partial blindness.

The users were divided randomly into two groups each containing six participants. In each group, two participants were totally blind, while four had partial blindness. The first group watched the films without audio descriptions, which is the most common situation for the visually impaired in Brazil. The second group watched the films with audio descriptions generated using CineAD and narrated using a speech synthesis software. A third group was planned to attend films with AD performed by a human professional. However, due to the small number of blind volunteer users available, it was not possible to perform the test with the third group. However, conducting a deeper analysis including testing with AD performed by human professionals is a proposal for future work.

Table 1 Test videos and script features

Question	Video 1	Video 2	Video 3
Duration	05 min 29 s	01 min 43 s	05 min 43 s
Qty of scenes	6	1	1
Qty of actions	35	16	83
Qty of dialogs	22	2	11
% time with dialogs	23.8%	12.6%	32.7%
% time without dialogs	76.2%	87.4%	67.3%

The goal of these first experiments was to evaluate the users' level of comprehension of the content (videos) of groups without AD and with AD generated by CineAD. Each group analyzed videos from the fiction genre: (1) short film (Video 1); (2) A scene from a feature film (Video 2); and (3) short film (Video 3). Table 1 shows the characteristics of the used videos. We chose fiction films, because they are the most representative genre of Brazilian Cinema. According to the 2015 statistical yearbook of the Brazilian National Film Agency[6] (ANCINE) fiction movies correspond to 81.1% of the audience and 78% of the total number of films released in Brazil.

According to Table 1, Video 1 contains 6 scenes. In this script, there are 35 action phrases and 22 dialogues. The

[6] http://oca.ancine.gov.br/sites/default/files/publicacoes/pdf/anuario_2015.pdf.

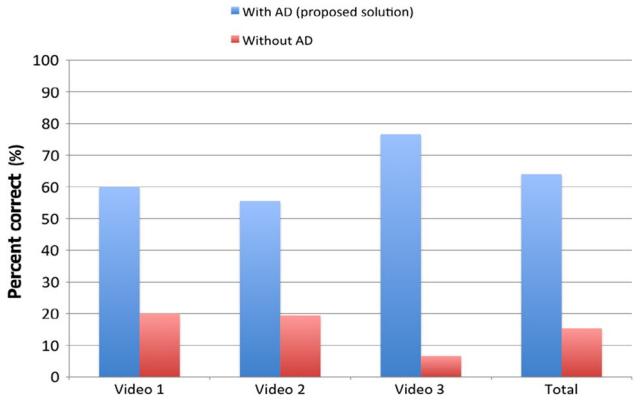

Fig. 8 Results for the comprehension tests

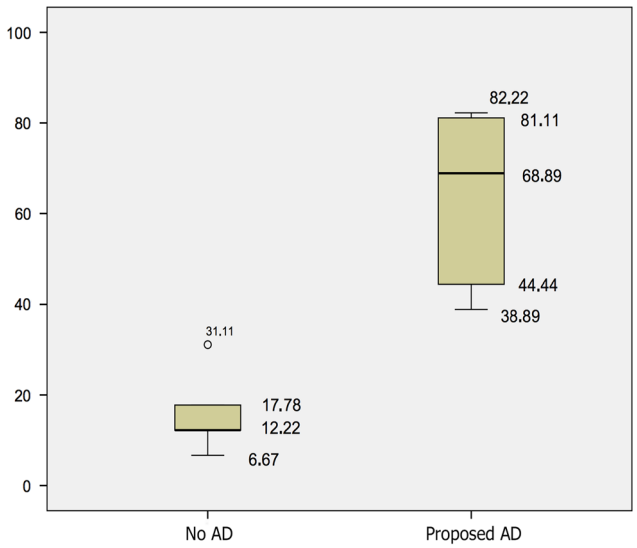

Fig. 9 Results for the comprehension tests (box plot)

time of the video occupied by dialogues corresponds to 23.8%, which implies that 76.2% of the film does not present speeches. Video 2, on the other hand, contains a single scene, 16 action phrases and 2 dialogues in the script, which occupy 12.6% of the time of the video. Finally, Video 3 contains 1 scene, 83 action phrases and 11 dialogues in the script. The duration of the dialogues corresponds to 32.7% of the time of the video, the remaining 67.3% of the time without speech. It should be noted that all videos have more visual information than dialogues.

The original scripts used in the experiment were provided by the producer. After each film was screened, the users were invited to respond to a questionnaire containing sixteen questions related to the visual content presented to assess their level of understanding. In these questions, users had to select which of the four alternatives (A, B, C or D) is related to the content presented, where only one of the alternatives is correct. For all questions, the fourth alternative (D) represented a "I do not know" option, which was included to prevent users from randomly choosing one of the alternatives when they did not know the correct answer. These questions were elaborated from randomly selected gaps of the film, five gaps for each video. For each randomly selected gap, we asked a generic question about aspects of the story that could only be accessed visually, such as: "What is being displayed on TV?", "What is the first thing the character does after waking up?", "How does a man dirty his clothes?".[7]

Figure 8 shows the results of the understanding level tests for each video/film and the overall results. The results for the level of understanding of the films showed that the users that watched the video with audio descriptions generated using CineAD, as expected, had a higher score on average than users in the group that watched without audio descriptions. However, besides observing whether we would have an improvement in the understanding, we also wanted to assess which would be the degree of the improvement.

As shown in Fig. 8, for all films, the average number of questions answered correctly when the film was watched without audio description is 15.37% with a standard deviation of 8.47%, whereas the group that watched these movies with audio descriptions generated by CineAD received an average score of 64.07% with a standard deviation of 18.37%. The audio descriptions generated by the proposed solution led to higher scores for all three films.

Additionally, to analyze the dispersion of the results, Fig. 9 shows a box plot of the results. As shown in Fig. 9, the users that watched the films with audio descriptions generated by the proposed solution performed better. For the group that watched the films without audio descriptions, the median and the first and third quartiles were 12.22%, 12.22% and 17.78%, respectively. A positive outlier was identified in the sample, which represents a user that obtained a score of 31.11%. This data point was not removed from the analysis because the personal experience of each visually impaired individual directly affects their perception level and is, therefore, considered valid for analysis.

For the group that watched the films with audio descriptions generated by CineAD, the median, first quartile and third quartile scores were 68.89%, 44.44% and 81.11%, respectively. 50% of the users scored between 81.11% and 44.44% with a minimum score of 38.89% and a high of 82.22%. Additionally, no outliers were identified, which means that all of the users obtained scores between 38.89% and 82.22%.

[7] The questionnaire can be accessed at this link: https://www.dropbox.com/s/sn2iejtpzapqass/Questionnaire%3AComprehensionTests.pdf?dl=0.

The Student's t test was also used to evaluate whether the performance of the two groups was significantly different. For this analysis, a 95% confidence interval and 10 degrees of freedom (total number of users: 12; degrees of freedom: $12 - 2 = 10$) were used. The t value obtained for this test was 5.896. The critical value for a 95% confidence interval and 10 degrees of freedom is 2.228 (t0,095,10). Because the t value obtained is greater than the relevant critical value, the null hypothesis (that the content with CineAD generated AD is understood at levels similar to the content without AD) can be rejected. Thus, we can conclude that the films with audio descriptions from the proposed solution were better understood by the users than the films without audio descriptions.

In addition to the level of understanding, other aspects of the proposed solution were also analyzed in this part of the experiment. At the end of the questionnaire, the users scored the following questions on a scale of 1 (confusion) to 6 (clear): (1) ease of understanding the film and (2) help provided by the audio descriptions in understanding the films.[8] This second question was not answered by the group that watched the films without audio descriptions.

The group with audio descriptions generated by the proposed solution scored the ease of understanding the films an average of 3.67 with a standard deviation of 0.81, whereas the group without audio descriptions gave an average score of 2.3 with a standard deviation of 1.5. Regarding the help provided by the audio description for understanding the films, the group that watched the films with CineAD-generated audio descriptions gave an average score of 4 with a standard deviation of 0.89.

The difference between the average scores for the ease of understanding the films between the two groups was only 1.37 despite the large difference between the scores of the analyzed groups (15.37% for the group without audio descriptions and 64.07% for the group with audio descriptions generated by the system). To analyze the consistency of these data, the correlation between the following variables was calculated: (a) accuracy rate in content comprehension tests and (b) degree of satisfaction (ease in understanding the films). The Pearson's correlation coefficient values and the Spearman's correlation coefficient values for the two variables were 0.353 and 0.265, respectively, which indicates a weak correlation between the two variables. As a result, we conclude that, in the present experiment, there was a weak correlation between satisfaction level and comprehension tests.

[8] A 1–6 scale was chosen because according to Morrissey [23], even scales encourage users to make positive or negative evaluations, avoiding neutral evaluations. In addition, this scale was also used in other works which also involve evaluation of solutions for people with disabilities (e.g., [10, 32]).

Table 2 Results of the specialists' evaluations

Question	Avg.	SD
Quality of the generated descriptions	2.8	1.3
AD script contains relevant and accurate information	3.2	1.79
Synchronization between events in the films and the descriptions	3.2	1.3
Gaps between dialogue are satisfactory	4.6	1.67

According to Wohlin et al. [5], one possible explanations for this result is that humans fear being evaluated, and when they undergo an evaluation process, they try to appear better than they really are, which could affect the outcomes of the experiment. Another explanation is that certain users are hesitant in using the assistive technologies and speech synthesis technologies that were used to narrate the generated descriptions in this experiment. However, further analysis must be performed to evaluate these aspects, which is one of our proposals for future research.

4.2 Evaluation by experts

The second part of the experiment involved evaluations by experts (audio description narrators). This analysis was performed online and included the participation of five volunteer narrators. All participants reported that they had experience in audio description of movies: one expert with 1 year of experience; two experts with 5 years of experience; and two with 10 years of experience.

The evaluators had access to one short film video, the original script of this film, a version of the same video containing the AD generated by CineAD using a speech synthesizer and the AD script in textual format.

The participants were invited to analyze all the original materials of the film and those generated by CineAD and then to evaluate through a questionnaire, on a score of 1 (poor) to 6 (very good) [8], the following questions.

- Does the AD script generated by the system increase the efficiency of the description process?
- Concerning synchronization, how concurrent were the events in the films and the descriptions?
- Were the gaps between dialogue identified by the system satisfactory?
- How do you rate the quality of the descriptions generated?
- Does the audio description script contain relevant and accurate information?

In addition, the questionnaire also provided space for the participants to comment on each of the aspects evaluated. Table 2 shows the results of this evaluation.

According to Table 2, the average score for the quality of the generated descriptions was 2.8 with a standard deviation of 1.3. The narrators considered the quality to be highly variable with several good stretches and others that included unnecessary descriptions related to audible sounds in the original film.

On average, the audio description narrators scored the relevance and accuracy of the information a 3.2 with a standard deviation of 1.79. The comments generally indicated that the proposed solution creates a relevant audio description script because it uses the original script for the film. However, according to the evaluators, several important elements were ignored, requiring rework by the audio description narrator to edit and modify the script before generating the final versions of the audio description.

The average score for synchronization between events in the films and the descriptions was 3.2 with a standard deviation of 1.3. Only one narrator commented on this topic, which indicates that in certain places, the audio description was misaligned.

The identification of gaps in the dialogue received an average score of 4.6 with a standard deviation of 1.67, which is the highest score in the evaluation. According to evaluator comments, gap identification is the major contribution of the proposed method.

In addition to the aspects shown in Table 2, the participants also responded to other questions regarding the utility of the system within an audio description creation process, and the following results were obtained: all of the narrators thought that the generated descriptions serve as a good reference during the audio description creation process; however, several script modifications would be required. Three of the narrators stated that they would use the system to facilitate their job and would recommend it to others. Two narrators would not use the system because it represents a mechanization of the process and because it has certain limitations. For this reason, they would not recommend it to other professionals.

Finally, the audio description narrators mentioned how the system could be improved and highlighted several functionalities that they would like to see in the solution, which included the following: exclude information regarding obvious sound effects already present in the film audio, improve the choice of descriptions used in the script, use the present tense, adjust the synchronization between the events in the film and the moment when the description is presented, adjust the speed of the narration to avoid abrupt changes, allow the script generated by the system to be edited to adjust the description text, timing and speed and allow a narrator to record the narration audio (human voice) during script editing.

5 Final remarks

A solution has been proposed for the automatic generation of audio description scripts for recorded audiovisual content, the CineAD. The used approach involves an analysis of the original script and of the film subtitles, which returns a script with timestamps and descriptions of the actions. The solution consists of a set of software components that:

- calculate the similarity between subtitles and portions of the script;
- identify the intervals between the dialogues (gaps) in the video. These gaps are candidates to receive audio descriptions;
- select important actions between the dialogues.

To validate the solution, some tests were conducted to obtain evaluations from the point of view of blind or visually impaired individuals and audio description narrators. The goal is to perform an analysis of the perception of the film by end users and professionals that work to create this assistive technology.

Results of the experiments show that this solution is an important assistive technology, providing descriptions of visual content for the blind and visually impaired. The users that watched the films with CineAD-generated audio descriptions answered 64.07% of the comprehension questions correctly, whereas users that watched them without audio description answered 15.37% of the questions correctly. Thus, from these results, we can conclude that the solution has the potential to reduce the barriers in accessing culture and information faced by people with visual impairments. In addition, this solution increases comprehension of the film, especially in scenarios where human audio descriptors are not available or are not feasible.

The analysis by audio descriptive narrators shows that the audio description script generated by this system contains relevant and accurate information. The narrators generally agreed that the ability to identify gaps in the dialogue is the major functionality of the system and thought that the generated descriptions serve as a reference though would still require editing before generating the audio descriptions.

The CineAD system was capable of generating descriptions of the most important events in the film. The solution can be incorporated into a speech synthesis tool or used by an audio description narrator to create the description audio. The system has the potential to be used by film

producers and audio description professionals to reduce the cost and time associated with the process of creating scripts and audio description tracks. Ancine, (*Agência Nacional de Cinema do Governo Brasileiro*) The Brazilian Nations Cinema Agency cited CineAD as one of five technological solutions in the country that promotes accessibility to cinema in a regulatory notice promoting visual and audible accessibility in movie theaters [1].

A proposal for future research is to carry out a more in-depth analysis to adjust the parameters used in the solution such as the number of most frequent words, list weights and narration speed, to refine the solution and improve the quality of the AD script. In addition, we also plan to perform a deeper analysis with blind users to compare the AD generated by the proposed solution with the AD produced by human professionals. We also plan to perform some tests with human professionals to evaluate the cost of adapting the AD script to the different versions of the film and compare it with the cost to produce an AD manually.

With regard to script processing, another future improvement is the incorporation of a way to shorten sentences in the summarization steps to reduce the size of the lines to facilitate insertion of the descriptions and allow greater adjustment of the audio description narration timing. In addition, the creation of rules or heuristics to remove script elements related to obvious sounds in the film would help optimize the description because audio is perceived by the visual impaired and, thus, audio largely does not need to be described.

Another proposal for future work is to incorporate techniques of video analysis to aid in the automatic audio description process and reduce the dependency of the script. Moreover, future work, in the face of various information extracted from the script and the film, could involve generating several versions of AD script with adaptations in the parameter values of the solution and its content, according to user preferences. Currently, the AD is generated only in the Brazilian Portuguese language, however, this is not a limitation of the system so that the system can be extended to other idioms, also as further work.

References

1. ANCINE: Brazilian nations cinema agency (ancine)– Regulatory News: accessibility (2015). http://www.ancine.gov.br/sites/default/files/consultas-publicas/Not%C3%ADcia%20Regulat%C3%B3ria%20-%20acessibilidade%20exibicao.pdf. Accessed Dec 2015
2. Araujo, V.L.S.: O processo de legendagem no Brasil (the subtitling process in Brazil). Revista do GELNE (GELNE Magazine), Fortaleza **1/2**, 156–159 (2006)
3. Benecke, B.: Audio-description. Meta Transl. J. **49**(1), 78–80 (2004)
4. Bojanowski, P., Lajugie, R., Bach, F.R., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Weakly supervised action labeling in videos under ordering constraints. European Conference on Computer Vision - ECCV (2014), Zurich, Switzerland. Springer, 8693 (Part V), pp. 628–643 (2014)
5. Wohlin, C., Runeson, P., Host, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering: An Introduction. Kluwer Academic Publisher, Norwell, MA, USA (2000)
6. Chapdelaine, C., Gagnon, L.: Accessible videodescription on-demand. In: Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility, Assets '09, pp. 221–222. ACM, New York, NY, USA (2009). https://doi.org/10.1145/1639642.1639685
7. Chen, X., Zitnick, C.L.: Mind's eye: a recurrent visual representation for image caption generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2422–2431. IEEE, Boston, MA (2015)
8. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20–25 June 2009, Miami, Florida, USA, pp. 919–926 (2009). https://doi.org/10.1109/CVPRW.2009.5206667
9. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. NIPS'16 Proceedings of the 30th international conference on neural information processing systems - Barcelona, Spain, pp. 379–387 (2016)
10. De Araújo, T.M.U., Ferreira, F.L.S., Silva, D.A.N.S., Oliveira, L.D., Falcão, E.L., Domingues, L.A., Martins, V.F., Portela, I.A.C., Nóbrega, Y.S., Lima, H.R.G., Souza Filho, G.L., Tavares, T.A., Duarte, A.N.: An approach to generate and embed sign language video tracks into multimedia contents. Inf. Sci. **281**, 762–780 (2014). https://doi.org/10.1016/j.ins.2014.04.008
11. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine intelligence (CVPR 15), vol. 39, no. 4, pp. 677–691. IEEE, Washington, DC, USA (2017). https://doi.org/10.1109/TPAMI.2016.2599174
12. Duchenne, O., Laptev, I., Sivic, J., Bach, F.R., Ponce, J.: Automatic annotation of human actions in video. In: 2009 IEEE 12th International Conference on Computer Vision (2009)
13. Edmundson, H.P.: New methods in automatic extracting. J. ACM **16**(2), 264–285 (1969). https://doi.org/10.1145/321510.321519
14. Encelle, B., Beldame, M.O., Prié, Y.: Towards the usage of pauses in audio-described videos. In: Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, W4A '13, pp. 31:1–31:4. ACM, New York, NY, USA (2013). https://doi.org/10.1145/2461121.2461130
15. Fang, H., Gupta, S., Iandola, F.N., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G.: From captions to visual concepts and back. (2014) CoRR http://arxiv.org/abs/abs/1411.4952arXiv:abs/1411.4952
16. Fernández-Torné, A.: Audio description and technologies: study on the semi-automatisation of the translation and voicing of audio descriptions. Ph.D. thesis, Universitat Autnoma de Barcelona, Barcelona, Spain (2016)
17. Giannakopoulos, T.: pyAudioAnalysis: an open-source python library for audio signal analysis. PloS One **10**(12):e0144610 (2015). https://doi.org/10.1371/journal.pone.0144610
18. Kobayashi, M., Nagano, T., Fukuda, K., Takagi, H.: Describing online videos with text-to-speech narration. In: Proceedings of the 2010 International Cross Disciplinary Conference on Web

Accessibility (W4A), W4A '10, pp. 29:1–29:2. ACM, New York, NY, USA (2010). https://doi.org/10.1145/1805986.1806025
19. Kobayashi, M., O'Connell, T., Gould, B., Takagi, H., Asakawa, C.: Are synthesized video descriptions acceptable? In: Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '10, pp. 163–170. ACM, New York, NY, USA (2010). https://doi.org/10.1145/1878803.1878833
20. Lakritz, J., Salway, A.: The semi-automatic generation of audio description from screenplays. Technical report CS-06-05, Dept. Of Computing, University of Surrey (2002)
21. Laptev, I., Marszaek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE, Anchorage, AK (2008). https://doi.org/10.1109/CVPR.2008.4587756
22. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2929–2936. IEEE, Miami, FL (2009). https://doi.org/10.1109/CVPR.2009.5206557
23. Morrissey, S.: Data-driven machine translation for sign languages. Ph.D. thesis, Dublin City University, Dublin, Ireland (2008)
24. Nenkova, A., Maskey, S., Liu, Y.: Automatic summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011, HLT-11, pp. 3:1–3:86. Association for Computational Linguistics, Stroudsburg, PA, USA, Article 3, 86 pp (2011)
25. Nunes, E.V., Machado, F.O., Vanzin, T.: Audiodescricao como Tecnologia Assistiva para o Acesso ao Conhecimento por Pessoas Cegas. (Audio description as assistive technology for access to knowledge for the blind). In: Ulbricht, V.R., Vanzin, T., Villarouco, V. (eds.) Ambiente Virtual de Aprendizagem Inclusivo (Inclusive Virtual Learning Environment), p. 352. Pandion, Florianopolis (2011)
26. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4594–4602. IEEE, Las Vegas, NV (2016)
27. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517–6525. IEEE, Honolulu, HI (2017)
28. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149 (2017)
29. Rohrbach, A., Rohrbach, M., Schiele, B.: The long-short story of movie description. In: Gall J., Gehler P., Leibe B. (eds.) Pattern recognition. DAGM 2015. Lecture Notes in Computer Science, vol. 9358. Springer, Cham (2015)
30. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. Int. J. Comput. Vis. **123**, 94–120 (2017). https://doi.org/10.1007/s11263-016-0987-1
31. Salway, A., Vassiliou, A., Ahmad, K.: Whats happens in films? In: Proceedings of the IEEE International Conference on Multimedia an Expo, ICME (2005)
32. San-Segundo, R., Montero, J., Córdoba, R., Sama, V., Fernndez, F., Dharo, L., López-Ludeña, V., Sánchez, D., García, A.: Design, development and field evaluation of a Spanish into sign language translation system. Pattern Anal. Appl. **15**, 203–224 (2012)
33. Szarkowska, A.: Text-to-speech audio description: towards wider availability of AD. J. Spec. Transl. **15**, 142–162 (2011)
34. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R.J., Darrell, T., Saenko, K.: Sequence to sequence–video to text. (2015) ICCV '15 Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4534–4542 (2015)
35. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R.J., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Denver, Colorado, USA, pp. 1494–1504, May 31–June 5 (2015)
36. Wang, K.C., Yang, Y.M., Yang, Y.R.: Speech/music discrimination using hybrid-based feature extraction for audio data indexing. In: 2017 International Conference on System Science and Engineering (ICSSE), pp. 515–519 (2017). https://doi.org/10.1109/ICSSE.2017.8030927