Vision 2008
Montréal, Canada

Proceedings of the 9th International Conference on Low Vision,
Vision 2008, Montreal, Quebec, Canada
July 7 – 11, 2008

Automatic detection of visual elements in films and description with a synthetic voice - Application to video description

Langis Gagnon[a,*]

[a]*R&D Department, Computer Research Institute of Montreal (CRIM)*

Abstract. We present two prototype software tools which aim at increasing the e-accessibility of video content for people with vision loss. One tool is to assist humans in generating an offline video description rough for mass production; the other is to allow people with vision loss to select the type and level of video description during playback. The tools integrate various video processing technologies to (1) automatically detect and recognize pertinent visual contents (shot transitions, faces, places, texts, etc.), (2) associate and embed a textual description to them on the audio track and (3) render with a synthetic voice during playback. This long term project first necessitated to meet with producers to understand the industry practice and with end-users to identify their needs. The paper gives a rather non-technical and descriptive overview of the current development and functionality of the tools.

Keywords: e-accessibility, video description, audio description, computer vision

1. Introduction

The objective of this work is to integrate computer vision technologies to develop software tools to facilitate the production and play of video description (also called audio description). “Video description, or described video, consists of narrative descriptions of a program’s key visual elements so that people with vision loss are able to form a mental picture of what is occurring on the screen” [CRTC (2004)]. In the industry, video description is done manually by human describers and requires 20 to 30 hours of work for a one-hour film. As part of the production chain, the task involves a kind of film summarization process to identify visual content that could be relevant to understand the story flow, such as scenes transitions, whom the actors are and when they appear, where the action takes place, pertinent textual information in the scene, where the silent segments are in the audio track, etc.

Computer vision technologies can automatically detect and recognize such visual elements (e.g. objects, faces, text, etc.). This is used, for instance, to automatically index large databases of images and videos or for content-based video summarization [see for instance, Boujemaa (2007) and Gagnon (2006) and references therein]. The same technologies can be adapted for computer-assisted video description to provide computer-vision tools that can automatically detect generic visual content to reduce workload and help describers produce more quickly, especially for mass production; consequently increasing the accessibility of media documents for people with vision loss.

On the other hand, not all people with vision loss appraise video description in the same way; user needs are diverse [Mathieu (2007)]. Whatever the quality or quantity of the video description track, each individual has preferences depending on his/her level of vision, tastes, and experience. Thus, designing and implementing a video description player that is adaptive is an important consideration to take into account.

In this paper, we present two prototype software tools currently under development in our laboratory [Foucher (2007)]. The first tool, called for now Video description manager (VDM), is to assist humans in generating offline video description; the second, called Video description player (VDP) is to allow people with vision loss to select the type and level of video description during playback.

2. Methodology

Our work targets two user groups: video describers (post-production) and end-users (people with vision loss). There is little literature and no regulated standard of practice regarding the production and usability aspects of video description. Some guidelines are available [Turner (2004), CNICE (2005), Clark (2001)] as well as interesting related works by Canadian researchers exploring the feasibility of online video description [Branje (2006), Fels (2006)]. However, getting inputs from both user groups is an important element to take into account for the success of this kind of project. Their needs cover all the aspects for this type of application: efficiency of the describing process, as well as pertinence of the visual description.

* langis.gagnon@crim.ca. CRIM, 550 Sherbrooke West, Suite 100, Montreal, QC, CANADA, H3A 1B9.

Both groups are involved in our project methodology which can be summarized in two main elements:

1. Meetings with producers to understand the industry practice and compare with end-users needs,
2. Selection, design and development of computer-vision tools taking into account the describers/end-users needs and the technical feasibility.

Details regarding the setup and outcomes of the first element can be found in [Mathieu (2007)]. One can summarize them as follows:

1. Producers do many viewings to assess who is doing what and when, what is the general idea conveyed by the film and where the silences are, and how long they are. This work necessitates a lot of planning in order to get a general idea of how to create video description and where to place it.
2. Overall, most of the description currently given by producers is about action, movement of the characters, occupation/roles of the characters, decor, facial/corporal expressions, textual information included in the image, and attitude of characters.
3. For end-users, priority should be given to identifying the principal characters as soon as possible, then presenting the time, the place, and the action, all while avoiding the pitfalls of interpreting the action and adding information not present in the image.
4. Finally, the question of personalizing the level and type of description is central to them.

In this paper, we concentrate on the second element of the methodology. We adopt a non-technical and descriptive formulation in order that the paper being accessible to non-expert in computer vision.

3. Tools description

The production practices and end-user needs were very informative as to the type and relative importance of visual content for video description. Automating all the visual content extraction process, however, is far beyond the state of the art in computer-vision. We had to make a trade-off between what is technically feasible and what would be most useful to the producers and end-users. We have thus concentrated on the followings: non-speech segments detection, shot transition detection, key-frame identification, key-face recognition, key-text spotting and reading, key-place identification, camera motion and human action characterization.

3.1 Video description manager

The extraction of the audiovisual content is produced by a collection of specialized filters [Gagnon (2006), Foucher (2007)]. Each filter requires adjusting the parameters to achieve their goal, and some are dependent on the output of others. The overall processing is managed by the VDM (Figure 1) whose goals are to ensure synchronization between the filters; that is, the acquisition of parameters, the order of execution, the inter-dependencies among the modules, the integrity of the output data, and the generation of scripts for batch processing on a collection of video files.

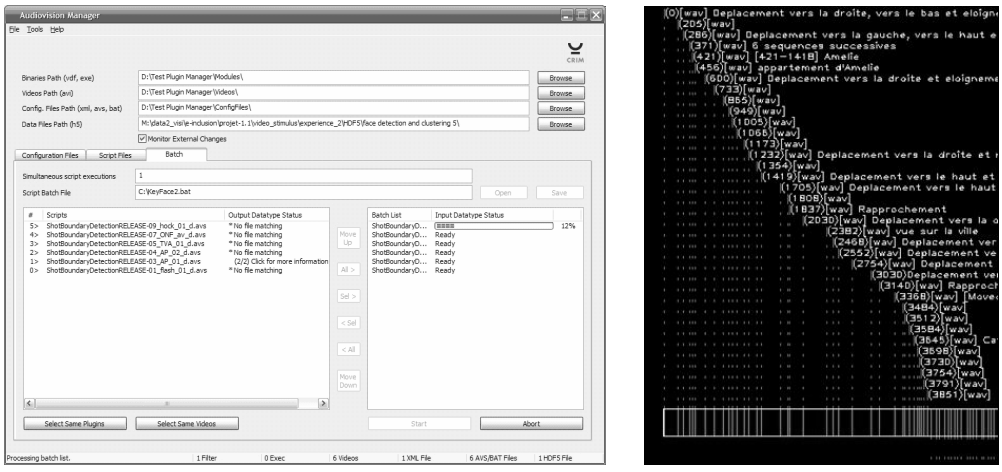

Fig. 1. Screen capture of the VDM prototype (left) and an example of output showing video description generated at various times (right).

Two main outcomes are generated at the end of the process:

1. A video file with a modified audio band containing video description rendered by a speech synthesizer.

2. A video description script composed of sentences along with time tags that can be passed to a speech synthesizer. It can then be edited by the producer or directly read by the VDP along with the video.

3.2 Video description player

The proof-of-concept version of the VDP explores the possibility of selecting and changing the level of video description at any time during video playback. The video itself does not contain any video description; the dynamically created video description recital is sent to the describer using an external speech synthesis program. The configuration dialog window of the VDP (Figure 2) shows the various controls available to the user during playback. In the upper part, the reading options determine the way video description is read. The reader voice can be changed and the reading speed can be increased or decreased. The lower part of the configuration dialog window shows the available video description layers that can be activated or not. The last two layers, the character identification and the location where the action is taking place, offer the additional possibility of being recalled at any time by clicking on the corresponding button.

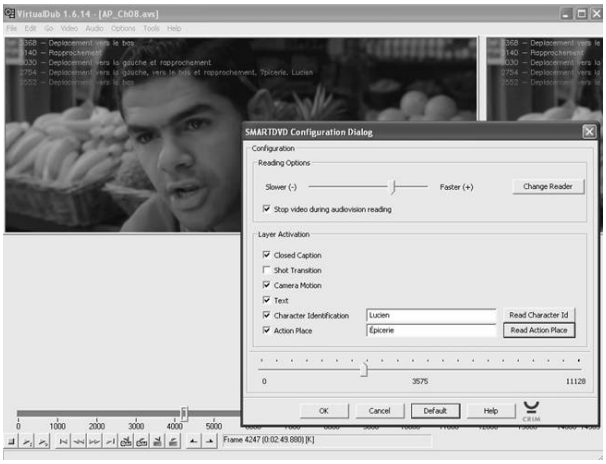

Fig. 2. Screen capture of the VDP prototype exploring the possibility to select the type of video description.

3.3 Beta version

We are now engaged in the second phase of the project which targets a pre-commercial tool with more user interactivity and versatility (Figure 3).

Fig. 3. Screen capture of the VDM beta version under development showing improved graphical user interaction.

The usability is improved by including many standard features one can find on commercial video manipulation/editing tools. The current features of the VDM include: project creation can contain one or more productions of various formats (avi, mpeg, etc.), each production can be processed independently, project and

filter preferences can be save, the tool loads each filter and sorts them automatically according to their dependence, filters/functions and scripts can be run as many times as needed, feedback is provided to the user to monitor the process, batch and interactive processing modes, etc. Also, the VDP is currently being adapted for video description streaming on the Web, compatible with specialized text-to-speech software used by people with vision loss based on keyboard shortcuts.

4. Conclusions

In this paper, we have given a brief description of two prototype software tools currently under development in our laboratory. One tool is to assist humans in generating offline video description; the other is to allow people with vision loss to select the type and level of video description during playback. At this time, we have completed the implementation of a proof-of-concept that:

1. Automatically processes a film, extract audio-visual information like non-speech segments, shot transitions, key-texts, key-places and key-faces,
2. Assigns a text to each visual elements that are added to the audio band, and
3. Optionally read it using a voice synthesizer.

The extraction of the audio-visual content is produced by a collection of specialized filters that extract high-level descriptions. The overall processing is managed by a software tool which primary goals are to ensure the synchronization between filters, the acquisition of the parameters and the generation of the scripts for batch processing. A rough video description is produced by a video description generator that can then be edited by a human describer or read by a specialized player. User consultations allowed us to identify the important information for the development of such tools. The feedback we have received so far is encouraging and we are now engaged in the development of a pre-commercial tool with more interactive features and versatility. We will report in more details in the near future.

Acknowledgements

This work is supported in part by the Department of Canadian Heritage (http://www.pch.gc.ca) through Canadian Culture Online, and the Ministère du développement économique, de l'innovation et de l'exportation (MDEIE) of the Gouvernement du Québec (http://www.mdeie.gouv.qc.ca).

References

-CRTC (Canadian Radio-television and Telecommunication Commission) (2004): http://www.crtc.gc.ca/archive/eng/Decisions/2004/db2004-21.htm.

-Boujemaa, N. et al. (2007), State-of-the-art on Multimedia Search Engines, Technical report D2.1, Chorus Project Consortium

-Gagnon, L., Foucher, S., Laliberté, F., Lalonde, M., Beaulieu, M. (2006). Towards an Application of Content-Based Video Indexing to Computer-Assisted Descriptive Video, *Proc. of Computer and Robot Vision*, (on CD-ROM), 8 pages.

-Mathieu, S., Turner, J.M. (2007). Audiovision interactive et adaptable, Technical Report, http://hdl.handle.net/1866/1307

-Foucher, S., Héritier, M., Lalonde, M., Byrns, D., Chapdelaine, C., Gagnon, L. (2007). Proof-of-concept software tools for video content extraction applied to computer-assisted descriptive video, and results of consultations with producers, Technical Report, CRIM-07/04-07, 68 pages

-Turner, J.M., Colinet, E. (2004). Using audio description for indexing moving images, *Knowledge organization* (Vol. 31, no. 4), 222-230.

-CNICE (Canadian Network for Inclusive Cultural Exchange) (2005). Online video description guidelines, http://cnice.utoronto.ca/guidelines/video.php

-Clark, J. (2001). Guidelines for video description: www.joeclark.org/access/description/ad-principles.html

-Branje, C., Marshall, S., Tyndall, A., Fels, D.I. (2006). LiveDescribe. *Proc. Americas Conference on Information Systems,* 3035-3041

-Fels, D.I., Udo, J.P., Diamond, J.E., Diamond, J.I. (2006). A first person narrative approach to video description for animated comedy, *Journal of Visual Impairment and Blindness*, (Vol. 100, no. 5), 295-305