Towards Collaborative Annotation for Video Accessibility

Pierre-Antoine Champin
Université Lyon 1, LIRIS
UMR5205, F-69622, France
pchampin@liris.cnrs.fr

Benoît Encelle
Université Lyon 1, LIRIS
UMR5205, F-69622, France
bencelle@liris.cnrs.fr

Nicholas W. D. Evans
EURECOM
Sophia Antipolis, France
nick.evans@eurecom.fr

Magali O.-Beldame
Université Lyon 1, LIRIS
UMR5205, F-69622, France
mbeldame@liris.cnrs.fr

Yannick Prié
Université Lyon 1, LIRIS
UMR5205, F-69622, France
yprie@liris.cnrs.fr

Raphaël Troncy
EURECOM
Sophia Antipolis, France
raphael.troncy@eurecom.fr

ABSTRACT

The ACAV project aims to explore how the accessibility of web videos can be improved by providing rich descriptions of video content in order to personalize the rendering of the content according to user sensory deficiencies. We present a motivating scenario, the results of a preliminary study as well as the different technologies that will be developed.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information System**]: Audio, Video and Hypertext Interactive Systems; K.4.2 [**Social Issues**]: Assistive technologies for people with disabilities

General Terms

Languages, Standardization

Keywords

Video Accessibility, Media Fragments, Media Annotations

1. INTRODUCTION

While video consumption on the web is continuously increasing, a large part of this content is not accessible to various categories of users. For example, blind and deaf users have little access to this enormous amount of content while digital technologies could, in theory, greatly improve the accessibility of rich media. Governments are supporting more and more actions to provide equal access to digital information on the web. In this context, improving the accessibility of multimedia content to disabled users is both a great challenge and an opportunity.

The ACAV project (`http://www.acavideo.fr/`) aims to explore how the accessibility of web videos can be greatly improved. The participants of this project are a large video sharing web site (Dailymotion), two research groups with expertise in disabilities and video annotation (LIRIS) and semantic web technologies and audio processing (EURECOM). Furthermore, several associations and specialists involved with the education of young disabled people are involved in the project. The research questions tackled by ACAV are: *i)* what is required to make a video accessible on the web and how can it be achieved?; *ii)* how to increase the number of accessible videos on the web?

W4A2010 - Communications, April 26-27, 2010, Raleigh, USA. Co-Located with the 19th International World Wide Web Conference.

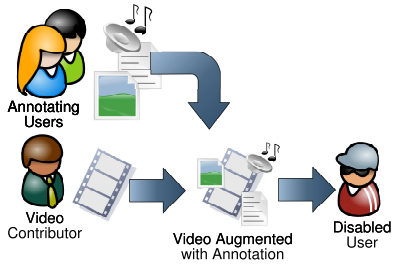

Figure 1: Motivating Scenarios Illustration

Our approach is to provide rich descriptions of video content in order to personalize the rendering of the content according to user sensory deficiencies. We advocate the use of speech processing technologies in order to provide an initial transcription of the audio content. We are developing tools to facilitate the manual correction of automatic transcriptions as well as the semantic annotation of the visual scene. We propose to add a social networking component in order to enable collaborative annotation and best practice sharing within communities. We are investigating how accessibility developed for the television can be adapted for the Web and we are designing novel interfaces for annotating and rendering video content.

In the next section, we present a typical scenario covered by the ACAV project. In Section 3, we describe a preliminary study involving blind users which aims to test various approaches to video rendering within the Advene platform. In Section 4, we present the various technological components required for the ACAV project. In Section 5, we discuss related initiatives for making video accessible on the Web. Finally, we give our conclusions and outline future work in Section 6.

2. MOTIVATING SCENARIOS

Luke has a deaf son, Brad. Luke is a member of an association of parents of disabled children in which he has heard about the Dailymotion tool for making videos accessible. He finds an interesting video uploaded by a video contributor

and decides to make it accessible to deaf people by adding textual descriptions (i.e. annotations) to some audio elements of the video. Luke first uses a speech-to-text module in which dialogues are automatically transcribed and speakers identified. Luke then corrects the initial transcription and adds further annotations corresponding to non-speech events (e.g. a car horn). Brad can then watch the video with these annotations presented as captions. After viewing his work, Luke decides to share his annotations with the other members of his association.

Jude is also a member of this association and has a blind child, Joe. Jude heard about the Dailymotion tool and decides to make the same video accessible to Joe. He thus employs the tool and adds new annotations to those produced by Luke, for describing some visual elements (e.g. characters, actions, etc.). Joe has a Braille display and can benefit from a multimodal presentation of annotations using the Braille display and a vocal synthesis (audio cues). Figure 1 illustrates this scenario.

While this scenario targets the "general public", we also consider other scenarios in different contexts: scenarios with educational video content (e.g. in a classroom with an instructional video described by a teacher to disabled pupils) and scenarios with copyrighted content uploaded by legal claimants with whom we already have agreements.

3. PRELIMINARY STUDY

In this section, we describe a preliminary study we conducted with blind users in order to tackle multimodality issues and in particular how to render videos that have been enriched with annotations. We extracted requirements for developing a system that will improve video accessibility.

3.1 Setup and Requirements

We conducted semi-structured interviews with two blind participants in order to capture requirements concerning the description of video. The first question (*Q1*) related to the participant's habits concerning the watching of programs with or without audio descriptions (e.g. TV programs, movies, theater). The second question (*Q2*) dealt with the advantages and drawbacks of the current French audio description process. Finally, possibilities given by multimodal presentations of descriptions (e.g. audio and tactile presentation) were discussed with the participants (*Q3*).

Regarding *Q1*, participants watch many programs without audio descriptions and often ask a nearby sighted person such as their husband, wife or friends, to give additional oral descriptions of the program. This process is only possible in specific situations that suppose the presence of a sighted relative or friend and assuming that these oral descriptions will not disturb others viewers.

Requirements 1a and 1b: As a result, on the one hand it seems to be important to develop solutions that suggest additional descriptions (1a). On the other hand, suggested solutions should provide unobtrusive access to descriptions (e.g. a tactile access for blind Braille readers)(1b).

Moreover, the analysis of the participants' comments about current descriptions (*Q2*) highlights the following problem: descriptions, depending on their types (e.g. places, character information) and on participants' preferences, are sometimes too verbose and too long: an appropriate balance between video story understanding (i.e. providing enough descriptions) and watching pleasure (i.e. providing just enough descriptions) has to be found.

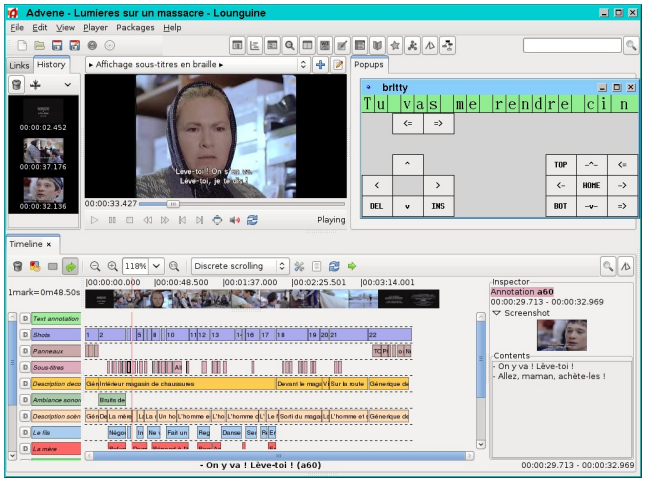

Figure 2: Annotations and Braille display emulation

Requirement 2: As a consequence, the possibility of suggesting descriptions with several levels of verbosity needs to be investigated.

In addition, according to participants' comments regarding *Q3*, the possibility of simultaneously providing two or more descriptions, using a system's multimodal output capacities (e.g. using a speech synthesis and a Braille display), seems to be promising (**Requirement 3**).

3.2 Discussion

Based on these requirements, we propose the following two features for improving video accessibility and have informally tested results using the Advene platform.

Characterization of the descriptions (R1a, R2): Five general *types* of information have to be described for blind people (in order of importance): character information and relationships, actions, places, time/periods and visual scenes [11]. Three *levels* of verbosity have also been drawn up: minimal, normal, complete. As a result, each description has to be characterized according to these types and levels and be transmitted according to viewer preferences.

Presentation modes for the descriptions (R1b, R3): Several monomodal and multimodal modes for presenting descriptions have been suggested: a mode is defined according to the modalities used (e.g. vocal synthesis, Braille display (Grade 1 or contracted Braille) and for each modality, the associated description types and verbosity levels.

Evaluation: The Advene tool (`http://www.advene.org`) has been used for adding and characterizing descriptions (i.e. annotations) to videos and for developing different multimodal presentations of the annotations (Figure 2). We conducted semi-structured interviews with participants who have experienced watching an annotated video using several presentation modes. Participants were asked to talk freely about their feeling and understanding of specific video excerpts after a particular rendering mode was selected. Results tend to confirm the relevance of the suggested description characterizations (types and verbosity levels) and their usage during description presentation. Concerning presentation modes, the tactile modality was greatly appreciated but the selection of descriptions transmitted using this modality has to be well defined: the description maximum length should match each blind person's reading speed.

4. TECHNOLOGY

In this section, we present the different technological blocks that will be developed in the course of the ACAV project.

4.1 General Architecture

The general architecture and workflow is depicted in Figure 3 where black blocks represent the technological components that will be developed. The Dailymotion server currently has a video database containing videos uploaded by a variety of contributors.

We will complement the video database with an annotation database containing all the additional information required to make videos accessible. This information will comply with a dedicated metadata model (Section 4.2), and will be created by a community of annotating users using a specific GUI to help them in the task. Since video transcriptions will obviously be an important part of the annotations, annotating users will also be assisted by a built-in speech-to-text module (Section 4.3).

The video and its annotations will then be combined to provide disabled users with adapted visualizations. On the server side, this implies dynamic access to different parts of the video using the forthcoming W3C Media Fragment URI recommendation (Section 4.4). On the client side, it implies a specific visualization GUI, developed using standard technologies available in modern browsers (Section 4.5). An open-source browser plugin for driving braille devices will also be developed. This will allow our GUI to make use of such devices and, beyond that, foster standardized accessibility to Web applications for blind people.

4.2 Metadata Model

We have proposed in [1] a general model for video annotation. This model has been implemented in the Advene application, and experimented within different contexts, including multimodal presentations of annotated video (see Section 2). We are therefore confident that this model can be adapted to the particular needs of the ACAV project. The main strength of this model is a clean separation between three parts: annotations, schemas and views.

Annotations are pieces of information attached to fragments of the video. Unlike other video annotation models, annotations in ACAV will not be intended for a specific rendering modality. For example, the same annotation can be displayed as a subtitle, sent to a braille device or to a speech synthesis system, depending on the user's disability, preferences or context.

Schemas are a way to categorize and constrain the structure of annotations. They embody a particular annotation practice, and allow to define the semantics of annotations. For example, one could define a schema for describing the dialogues of a video, another schema for the musical part, and yet another schema for the scenes.

Finally, *views* specify how annotations can be rendered. A view can combine annotations from several schemas, and several views can be designed for the same schema. One of the challenges of ACAV will be to enable annotating users to define the most appropriate views, but also to allow disabled users to chose and customize views to suit their specific needs and preferences.

4.3 Speaker Diarization, Speech Transcription

Two speech processing modules will be developed within ACAV, namely those of speaker diarization [7] and automatic speech recognition (ASR) [9]. Speaker diarization is used to automatically detect the different speakers in a multimedia document and to identify intervals during which each speaker is active. Not only can it be used to enrich a text transcription with different speaker identities, i.e. by using a different colour for the text transcription of each speaker, speaker diarization can also be used to improve ASR performance through speaker adaptation, i.e. through speaker-attributed speech-to-text. In either speaker-dependent or speaker-independent mode, ASR can be used to provide an initial transcription of the spoken words. In addition, ACAV will provide a module for the manual and collaborative editing of automatically generated transcriptions.

ASR is a mature technology and several toolkits exist. HTK (`http://htk.eng.cam.ac.uk/`) is perhaps the best known but its use is subject to various license restrictions. The CMU Sphinx toolkit (`http://cmusphinx.sourceforge.net/`) is an open-source alternative and is an ideal candidate for use in the ACAV project. In contrast, speaker diarization is a relatively new field of speech research. Systems based on the open-source ALIZE toolkit for speaker recognition (`http://alize.univ-avignon.fr/`) will be used for all work in speaker diarization.

4.4 Media Fragment URI

The current Web architecture provides a means for uniquely identifying sub-parts of resources using URI fragment identifiers (e.g. for referring to a part of an HTML or XML document). However, for almost any other media types, the semantics of the fragment identifier has either not been specified or is not commonly accepted. Providing an agreed upon way to localize sub-parts of multimedia objects (e.g. specific tracks, sub-regions of images, temporal sequences of videos or tracking moving objects in space and in time) is fundamental [5].

Specific media servers are generally required to provide for server-side features such as direct access to time offsets into a video without the need to retrieve the entire resource. Support for such media fragment access varies between different media formats and inhibits standard means of dealing with such content on the Web. We are working within the W3C Media Fragments Working Group (`http://www.w3.org/2008/WebVideo/Fragments/`) on the specification of a media-format independent way of addressing media fragments on the Web using Uniform Resource Identifiers (URI). In particular, media fragments are regarded along three different dimensions: temporal, spatial, and tracks. Further, a fragment can be marked with a name and then addressed through a URI using that name. The specified addressing schemes apply mainly to audio and video resources - the spatial fragment addressing may also be applied to images [5].

4.5 Interfaces

Various interfaces will be developed: authoring interfaces for annotating users and accessible visualization interfaces for disable users. Those interfaces will be based as much as possible on standard web technologies. This is made possible by the ongoing efforts in the development of HTML5, which is already largely supported by most modern browsers, and by video websites such as Dailymotion `http://openvideo.dailymotion.com/`.

Providing smooth interfaces for disabled users is never-

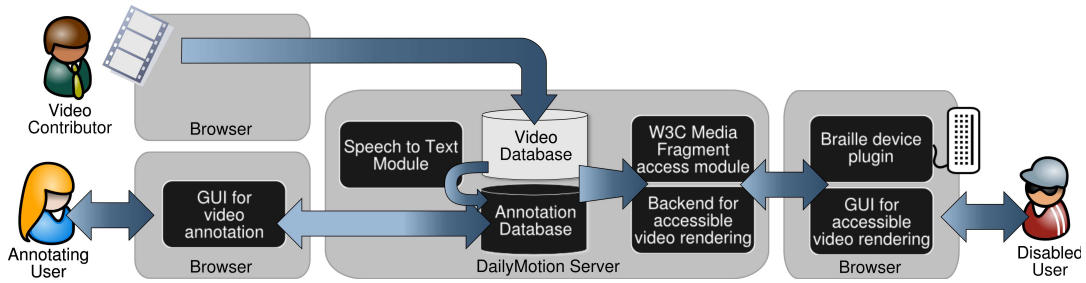

Figure 3: The architecture of ACAV

theless a challenging task. From a technical point of view, this will be distributed on several components: a server-side back-end will prepare the data so that it can be consumed by a client-side GUI based on standard technologies. The client-side GUI will also be able to drive a Braille display thanks to an open-source plugin that we will release as a result of this project.

5. RELATED WORK

Classical accessibility techniques for video include: *audio description* (adding a voice to the audio stream that describes the visual content during non-dialog moments), the use of a supplementary video stream with *sign language information* and *captioning and subtitling*. Only the latter technique, advocated in the Web Content Accessibility Guidelines (WACG) (`http://www.w3.org/TR/WCAG20/`), is commonly used on the web, thanks to the easy rendering of subtitles within videos and the availability of annotation tools such as MAGPie (`http://ncam.wgbh.org/invent_build/web_multimedia/tools-guidelines/magpie`), Nico Nico Douga (`http://www.nicovideo.jp/`) and YouTube subtitler (`http://yt-subs.appspot.com/`). Dedicated tools for video accessibility tailored to blind people have been proposed such as the aiBrowser [6]. The Canadian project E-inclusion [2, 4] is an ambitious initiative whose goal is to define automatic tools that analyze content in order to generate video metadata that can be used for accessible adaptable rendering. Even if this project goes further than ACAV on automatic processing, it does not focus on collaborative manual annotation nor multimodal rendering. Furthermore, we aim at full multimodality rendering (e.g. develop an open source Braille plugin for browsers) while e-inclusion consider only the audio modality. *Social accessibility* or the collaborative annotation of media for accessibility has been considered in [3, 10]. Finally, video on the web is gaining more and more importance through recent initiatives such as the open video conference (`http://openvideoalliance.org/open-video-conference/`); accessibility is now carefully considered. The Mozilla Foundation has recently reported a study on video accessibility [8] while an informal meeting held in Stanford (`http://www.w3.org/2009/11/01-media-minutes`) that gathered 25 experts led to the creation of an HTML accessibility task force within the W3C HTML5 Working Group in which we plan to participate actively.

6. CONCLUSION AND FUTURE WORK

This paper presents the general approach of the ACAV project, its technical components, and a preliminary study conducted towards annotation-based video enrichment for accessibility. Future work includes user studies on precise video enrichment questions, iterative design and development of the ACAV platform in close collaboration with partner associations, as well as various evaluation steps for the validation of our approach and developed technologies.

Acknowledgements

This paper was supported by the French Ministry of Industry (*Innovative Web* call) under contract 09.2.93.0966, "Collaborative Annotation for Video Accessibility" (ACAV).

7. REFERENCES

[1] O. Aubert, P.-A. Champin, Y. Prié, and B. Richard. Canonical Processes in Active Reading and Hypervideo Production. *Multimedia Systems*, 14(6):427–433, December 2008.

[2] C. Chapdelaine and L. Gagnon. Accessible videodescription On-Demand. In *ASSETS'09*, pages 221–222, 2009.

[3] S. Ferretti, S. Mirri, M. Roccetti, and P. Salomoni. Notes for a Collaboration: On the Design of a Wiki-type Educational Video Lecture Annotation System. In *Int. Conf. on Semantic Computing (ICSC'07)*, pages 651–656, Irvine, USA, 2007.

[4] L. Gagnon and *al.* Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Universal Access in the Information Society*, 8(3):199–218, 2009.

[5] M. Hausenblas, R. Troncy, Y. Raimond, and T. Bürger. Interlinking Multimedia: How to Apply Linked Data Principles to Multimedia Fragments. In *2^{nd} Workshop on Linked Data on the Web (LDOW'09)*, Madrid, Spain, 2009.

[6] H. Miyashita, D. Sato, H. Takagi, and C. Asakawa. aiBrowser for Multimedia: Introducing Multimedia Content Accessibility for Visually Impaired Users. In *ASSETS'07*, pages 91–98, 2007.

[7] NIST. The NIST Rich Transcription Evaluation, 2009.

[8] S. Pfeiffer and C. Parker. Accessibility for the HTML5 `<video>` element. In *6^{th} Int. cross-disciplinary conference on Web accessibility (W4A'09)*, pages 98–100, Madrid, Spain, 2009.

[9] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice Hall, 1993.

[10] H. Takagi, S. Kawanaka, M. Kobayashi, D. Sato, and C. Asakawa. Collaborative Web Accessibility Improvement: Challenges and Possibilities. In *ASSETS'09*, pages 195–202, 2009.

[11] J. Turner and E. Colinet. Using audio description for indexing moving images. *Knowledge organization*, 31(4):222–230, 2004.