Models for Video Enrichment

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

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

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

Olivier Aubert
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
oaubert@liris.cnrs.fr

ABSTRACT
Videos are commonly being augmented with additional content such as captions, images, audio, hyperlinks, etc., which are rendered while the video is being played. We call the result of this rendering "enriched videos". This article details an annotation-based approach for producing enriched videos: enrichment is mainly composed of textual annotations associated to temporal parts of the video that are rendered while playing it. The key notion of enriched video and associated concepts is first introduced and we second expose the models we have developed for annotating videos and for presenting annotations during the playing of the videos. Finally, an overview of a general workflow for producing/viewing enriched videos is presented. This workflow particularly illustrates the usage of the proposed models in order to improve the accessibility of videos for sensory disabled people.

Categories and Subject Descriptors
I.7.2 [**Document Preparation**]: Multi/mixed media.

General Terms
Design, Languages.

Keywords
Models for videos enrichment, video enrichment, hypervideo

1. INTRODUCTION
Videos are commonly being augmented with additional content such as captions, images, audio, hyperlinks, etc., which are rendered while the video is being played. We call the result of this rendering "enriched videos". The goal of video enrichment can be either to make parts of the video content available to people that cannot fully perceive its visual or audio content, or for complementing it with additional information so as to enhance the watching experience.

This article presents several contributions regarding the production of enriched videos. Two models are first detailed: the first one is for representing the content of enrichments as temporally situated structured annotations, and the other is for describing the presentation modalities of these annotation contents. These models are illustrated with an example corresponding to the ACAV (Collaborative Annotation for Video Accessibility) project general workflow. ACAV explores how enriching videos can improve their accessibility for sensory disabled people.

DocEng'11, September 19–22, 2011, Mountain View, California, USA.

We first introduce the key notion of video enrichment and associated concepts (section 2). We then present in section 3 the models we have developed for video enrichment: the first one is for annotating the video and the second is for rendering annotations during the playing of the video. Next, in order to illustrate a possible usage of these models, we present an overview of the general ACAV workflow for producing/viewing enriched videos. The related works (section 4) focuses on existing approaches for producing enriched videos before concluding and presenting future work.

2. VIDEOS ENRICHMENT USAGES
Utility of video enrichment is twofold. Video can be enriched either to *translate* parts of its content or to *complement* it with additional information in order to enhance the watching experience. Concerning *translation-based video enrichment*, the objective is that people who cannot fully understand the video visually or aurally can apprehend it. For instance, subtitling and superimposed dubbing have been two common means of enriching a video to translate its dialogs in a foreign language. For sensory impaired people, the objective is to present the key audio information or key visual information of the video using respectively either some visual presentation modalities or some audio, tactile (Braille) presentation modalities. For instance, the *audio description* of a video concerns visually impaired people and consists in adding verbal information to the audio track of the video in order to describe the visual content of the video [6]. For deaf and hearing-impaired users, *teletext* is for instance a digital service that allows a television channel to broadcast closed-captions that describe the audio track of a program (dialogs, sounds, music), as are subtitles for hearing-impaired on DVDs.

Complement-based video enrichment is different from *translation-based video enrichment*: the objective is not anymore to ensure that viewers will apprehend the video as intended by its creators, but to offer new experiences. For instance, chat messages rendered as subtitles can comment on a TV program, video meaning can be changed (*e.g.* from tragedy to comedy) by added sounds, added visual elements (e.g. arrows) can underline important elements in a scenery, *etc.* As another example of complement-based video enrichment, Díaz Cintas [4] emphasizes some new subtitling practices of professionals or hobbyist annotators. He stresses out subtitling activities that add precision to little-known terms (*e.g.* specialized vocabulary) by using explanations in brackets or texts placed at the top of the screen, which he calls *headnotes* or *topnotes*. One step further, videos can become part of hypermedia, as video enrichment paves the way for new interaction possibilities [7]. Indeed, a hyperlinked video, or hypervideo, is a hypermedia document into which video streams are enriched with embedded, clickable anchors. Clicking these anchors results in navigating to other places in the same video, or to other videos, or to others information elements. Such combination of video with non-linear information structure can be

used in various domains: storytelling (*HyperCafe* [9]), e-learning (adding slides, references, links, *etc.* to video content).

Considering the technological point of view on video enrichment, our approach considers *annotation-based video enrichment*. A video annotation is here defined as *any information associated to a fragment of a video* (e.g a textual transcription of a dialog associated to a temporal fragment, defined by two timecodes) [2]. Annotation data can be *rendered* so as to enrich a video – i.e. presenting its content using an adequate modality (*e.g.* visual enrichment with textual captions, images, video fragments, *etc.* or auditory enrichment with voice, music or sounds). As a result, the general process for enriching videos is made up of two main steps: an annotation step and a rendering step (*cf.* Figure 1).

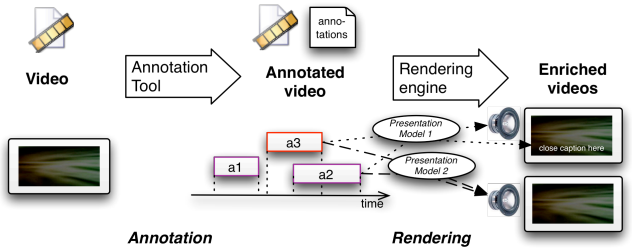

Fig. 1. The general process of annotation-based video enrichment.

In our opinion, this separation –similar to the structure/presentation separation in document engineering– has good properties: annotations and their renderings are independently defined. This can foster innovation by allowing different people to create content and content rendering, for example in a collaborative process. It also allows performing "live" video enrichment according to end-user preferences that can change during the rendering itself, paving the way to real-time adapted enrichments. The following section deals with models we have developed first for annotating videos and second for specifying presentation intents of the content of annotations.

3. MODELS FOR VIDEO ENRICHMENT

3.1 Annotation Model

We have previously proposed [1] a general model for video annotation. This model has been implemented in the Advene application[1], and experimented within different contexts, including multimodal presentations of annotated videos. We borrow from that general model the main elements of our **annotation model** (cf. Figure 3, annotation package):

- *Annotations* are the main elements of our model. Basically, an annotation has a unique id, a content and is associated to a temporal fragment (two timecodes addressing the original video).
- *Annotation Types* are a way to structure annotations as every annotation has exactly one type (e.g. annotations of type *Character*, of type *Setting*, etc.). They define the semantics of annotations and constrain their content.
- *Annotation Tags* are a more flexible way to categorize annotations. Every annotation can be associated to one or more tags.
- An *Annotation Schema* embodies a particular annotation practice as a set of annotation types. For example, one could define a schema for describing the dialogues of a video, another schema for the musical part, etc.

As an example related to improving the accessibility of a movie for blind people, a schema called "VisualBase" that could contain textual annotations of type "*Character*", "*Action*", "*Setting*" can be created to describe key visual elements of the movie (*Character* annotations for describing characters appearance/role and their interrelations, *Settings* annotations for describing the different settings of the movie, etc.)[2].

3.2 Annotation Presentation Model

This section exposes the main elements of our annotation presentation model (*cf.* Figure 3, presentation package) as a specialization of the notion of views defined in [1].

Presentation rules. A *Presentation Rule* R specifies the presentation of a subset of annotations according to one or several presentation modalities (e.g. a text-to-speech engine, a subtitle displayer, etc.). A presentation rule is composed of an annotation selector S associated to a set of presentation actions A_i.

$$R = < S, \{A_1,\ldots, A_q\} > \textit{with } q > 0$$

An *Annotation Selector* S selects a subset of annotations from an annotation set according to a set of constraints. Constraint types can be:

- Structural: upon structural elements of our annotation model (schemas, types, tags);
- Intrinsic: upon annotation id or content.

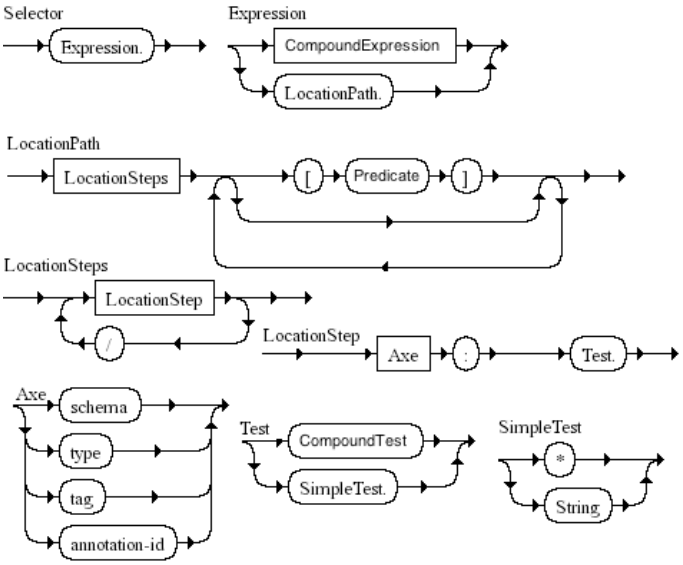

Fig. 2. Partial syntax diagram of an annotation selector.

A syntax similar to XPath [12] was used to specify selectors (cf. Figure 2). In fact we reuse the XPath concepts of "*Location Path*" and "*Location Step*" for selecting a subset of annotations using "constraints". A *location step* is first expressed according to an *axe* that indicates the nature of the constraint and then according to a *test* that filters the annotations subset according to the axe. For instance, the location step "*schema*:*VisualBase*" filters a set of annotations selecting only annotations associated to the *schema* "*VisualBase*". Others possible axes are: *type, tag, annotation-id*.

[1] http://www.advene.org

[2] This way of describing a movie actually corresponds to an existing description practice that is called "audio-description".

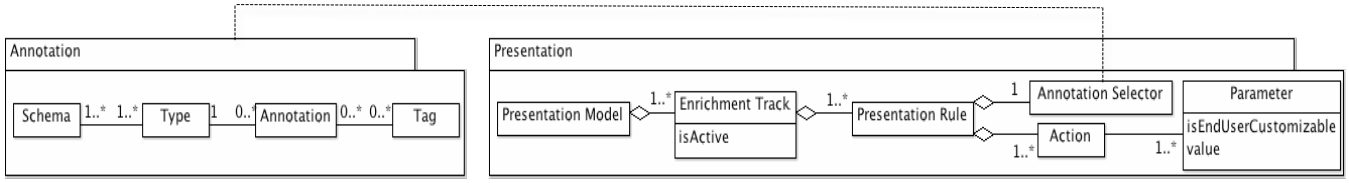

Fig. 3. Main elements of the models for video enrichment

Annotation selection also makes use of *predicates* that add intrinsic constraints by filtering the previously selected annotation subset (e.g. with the location steps) mainly by using conditions on annotation content. A predicate can be modeled as a regular expression. For instance the predicate "[cC]at | [dD]og" selects the annotations in the subset that have their content equal to "cat", "Cat", "dog" or "Dog".

Actions. An *action* A corresponds to an *action type* T and specifies the presentation of annotation content. Suggested action types are: speech synthesizing, Braille displaying, subtitling, close-captioning, audio icon playing, etc.

The author of a presentation model can parameterize actions. Some actions might also be customizable by the end-users watching the enriched videos. Most parameters and their possible values are taken from CSS [10].

$$A = < T, \{P_1, \ldots, P_r\} > \textit{with } r \geq 0$$

$$P_i = < \text{parameterName, value, isUserModifiable} >$$

For each parameter, the author of the model can give to the end-user the permission for changing its value (boolean *isUserModifiable*). For instance, some parameters associated to an action of type "speech synthesizing" are: *voice-family* (with possible values: male, female, child), *defaultPlaybackRate* (with possible values x0.5, x1, x1.5, x2), *volume* (percentage), etc.

We have proposed in [5] several families of parameters depending on properties associated to each action type. For instance, temporal actions (i.e. actions that present messages that generally evolve during time) parameters include *defaultPlayBackRate*, *minPlaybackRate*, *maxPlaybackRate*. Parameters related to audio actions include *volume*, *panning*.

In our example, the list of parameters of the action "speech synthesizing" would be completed with those parameters of both temporal actions and audio actions.

Enrichment tracks. An *Enrichment Track* T is made up of one or several presentation rules R_i and can be activated/deactivated during the playing of the video.

$$T = \{R_1, \ldots, R_n\} \textit{ with } n > 0$$

Presentation model. Finally, a *Presentation Model* M, aims to specify the presentation of a considered annotation set. It is made up of m enrichment tracks T_i,.

$$M = \{T_1, \ldots, T_m\} \textit{ with } m > 0$$

Some integrity constraints apply to our model (we only describe them informally because of space limitations):

- an action A cannot contain two parameters with the same *parameterName*,
- a rule R cannot have two actions with the same action type,
- in a track T, two actions with the same action type (hence in two different rules) cannot have different values for the same *parameterName.*

Going on with our example of annotations of types "*Character*", "*Action*", "*Setting*" corresponding to the "*VisualBase"* schema, we propose an example presentation model for producing audio enrichments in order to have an audio description of a movie. This presentation model is composed of two enrichment tracks, each one containing one presentation rule.

The presentation rule of the first enrichment track selects annotations of types *Character* and *Action* and presents the content of these annotations using a speech synthesizer with a male voice:

R_1 = <"type:Character or type:Action", {<speechSynthesizing, {voice-family, male, false}>}>

The presentation rule of the second enrichment track selects *Setting* annotations and presents each annotation by first playing a short sound (that indicates a set change) and second speech synthesizing the annotation content (i.e. the set description) with a female voice:

R_2 = <"type:Setting", {<soundPlaying, {<file, bip.mp3, false>}>, <speechSynthesizing, {voice-family, female, false}>}>

Note that, according to the third integrity constraint, those two rules have to belong to different tracks, as they use two different values for the *voice-family* parameter of *speechSynthesis*.

3.3 An Example Workflow For Producing Enriched Videos Integrating Proposed Models

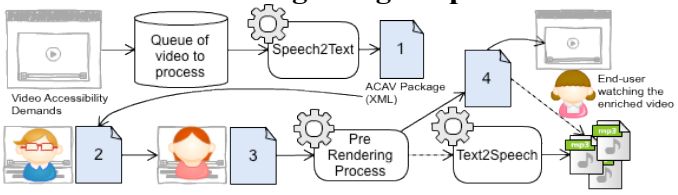

Fig. 4. ACAV general architecture for producing enriched videos

The ACAV project general workflow for producing enriched videos (cf. Figure 4) illustrates a usage of the suggested video enrichment models. The format of documents 1 to 3, and associated workflow steps (e.g. annotation and presentation model authoring steps) are a specialization of the document format and workflow steps used in the Advene project [1]. According to the ACAV workflow, the first output document mainly contains annotations that come from a *Speech2Text* process used for transcribing the dialogs of the video. The second document is made up of the content of the first one and of supplementary annotations added by "annotator" users (e.g. for describing key visual elements of the video). These two documents represent, in an XML syntax, elements of the annotation model (i.e. schemas, types, tags, annotations) created by annotator users and by the Speech2Text process. The third document expands the second one with the XML description of the presentation model. According to these specifications and some annotation contents, enriching contents can be created, such as *mp3* files containing generated audio description. Finally, the sensory disabled end-user watches the resulting enriched video and may customize it in adjusting some track action parameters (e.g. volume, defaultPlayBackRate for speechSynthesizing): document 4 that contains annotations and information about their rendering (e.g. links to *mp3* files) is

interpreted during the playing of the video, producing an accessible "audio described" video.

An important feature of that workflow, illustrated by Figure 4, is that different users can contribute to different elements of the packages (e.g. schemas, annotations, presentation models), as those can be defined independently. Furthermore, generic schemas or presentation models can be pre-defined, shared and reused with multiple videos.

4. RELATED WORKS

Concerning approaches for video enrichment, some work has been done for annotating multimedia content (i.e. including video content) [2, 3]. In comparison with our approach, the main difference in our opinion is that these approaches do not separate the annotation structure (content) from its presentation. To our mind, they do not foster, as much as our approach does, reusability and sharing. Indeed, with our approach, the same annotation package can be reused for making several different kinds of enrichments in using different presentation models (e.g. audio enrichments, visual enrichments, audio + tactile enrichments, etc.). In the same way, a successful presentation model can be applied on several videos through different annotation packages. Moreover, our workflow for producing enriched videos is collaborative by essence: it allows the involvement of several people for annotating videos and/or for designing presentation models, paving the way to the emergence of video enrichment practices.

Concerning technical solutions for publishing enriched videos, SMIL (Synchronized Multimedia Integration Language) can be used for synchronizing different multimedia contents (e.g. a video synchronized with an audio file containing an audio-description and with a subtitle file). SMIL can be also used to annotate some SMIL content [2]. Several technical recommendations, initiatives and formats have emerged from the community working on accessibility. The Web Accessibility Initiative (WAI) advocates in its recommendation entitled "Web Content Accessibility Guidelines (WCAG)" [11] the development of different versions of a given temporal content (audio and visual versions for sensory disabled people). Concerning formats, the Mozilla Foundation [8] advocates the usage of the Ogg format with multiplexed specialized tracks for video accessibility. In the same way, the HTML accessibility task force suggests adding several tracks to a video content to improve its accessibility: e.g. a subtitle track, an audio-description track, etc. These "enrichment" tracks would be represented as HTML 5 *Track* elements inside a *Media* element (i.e. *Video* or *Audio* element). The notion of enrichment track in our model is very closed from this HTML Track element. In comparison with the HTML 5 Track notion, our concept of enrichment track permits a deeper end-user customization of enrichment, by adjusting some parameters of actions (cf. 3.2).

5. CONCLUSION AND FUTURE WORK

We have proposed an annotation-based approach to produce enriched videos. Two models are presented: the annotation model permits the association of typed annotations to fragments of a video; the annotation presentation model allows to describe how the video is to be enriched with the rendering of annotation contents using various modalities (*e.g.* textual captions, images, video fragments, spoken texts, music or sounds). Concerning video enrichment for improving video accessibility, several experiments involving people with disabilities are currently being conducted in order to evaluate the consistency of these models. The annotation presentation model is inspired from XPath for defining annotation selectors, and from CSS for specifying presentation intents for selected annotations. The selector specification model should probably be extended in order to support the definition of temporal constraints upon annotations, e.g. for selecting annotations that start before the beginning of an annotation X and that end before the beginning of an annotation Y. We will also study mechanisms for checking the "consistency" of annotation presentation, as a presentation model can indicate that several time-overlapping annotations have to be presented using the same presentation action, resulting in hardly perceptible information for the end-user (e.g. two or more overlapping subtitles, etc.).

6. ACKNOWLEDGEMENTS

This paper was partly 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 and Y. Prié. 2007. Advene: an open-source framework for integrating and visualising audiovisual metadata. In *15th international conference on Multimedia* (MULTIMEDIA '07). ACM, 1005-1008.

[2] D.C.A. Bulterman. 2003. Using SMIL to encode interactive, peer-level multimedia annotations. In *2003 ACM symposium on Document engineering* (DocEng '03). ACM, 32-41.

[3] R.G. Cattelan, C. Teixeira, R. Goularte, and Maria Da Graça C. Pimentel. 2008. Watch-and-comment as a paradigm toward ubiquitous interactive video editing. ACM Trans. Multimedia Comput. Commun. Appl. 4, 4.

[4] J. Díaz Cintas. 2005. Back to the Future in Subtitling. In Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation, University of Saarland, 2005

[5] B. Encelle. Modèle pour la spécification de modèles de présentation d'annotations associées à des videos. Online : http://liris.cnrs.fr/Documents/Liris-5147.pdf. Accessed 06/14/11.

[6] L. Gagnon, S. Foucher, M. Heritier, M. Lalonde, D. Byrns, C. Chapdelaine, J. Turner, S. Mathieu, D. Laurendeau, N. T. Nguyen, and D. Ouellet. 2009. Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Univers. Access Inf. Soc.* 8, 3 (July 2009), 199-218.

[7] J. Geißler. Surfing the movie space: advanced navigation in movie-only hypermedia. In *3th international conference on Multimedia* (MULTIMEDIA '95). ACM, 391-400.

[8] S. Pfeiffer and C. Parker. Accessibility for the HTML5 <video> element. In *6th International Cross-Disciplinary Conference on Web Accessibililty* (W4A '09). ACM, 98-100.

[9] N. Sawhney, D. Balcom, and I. Smith. 1996. HyperCafe: narrative and aesthetic properties of hypervideo. In *7th ACM conference on Hypertext* (HYPERTEXT '96). ACM, 1-10.

[10] W3C, Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification. Online : http://www.w3.org/TR/CSS2/. Accessed 06/14/11.

[11] W3C, Web Content Accessibility Guidelines (WCAG) 2.0. Online : http://www.w3.org/TR/WCAG20/. Accessed 06/14/11.

[12] W3C, XML Path Language (XPath) 2.0. Online : http://www.w3.org/TR/xpath20/. Accessed 06/14/11.