

Project: Video for End-users

Technology Scouting Video Transcription

User stories

Version: 1.0

Date: March 5, 2010

SURFnet / Kennisnet Innovatieprogramma 2010

Video Transcription: User stories

Introduction

With the growth of educational video content being available, the demand for more effective ways to increase the usability of online video content also rises. Usability will be increased when techniques like video transcription lead to improved metadata and search ability.

Transcription, a traditional definition:

Transcription is the conversion into written, typewritten or printed form, of a spoken-language source, as in the proceedings of a court hearing. It can also mean the conversion of a written source into another medium, as by scanning books and making digital versions.

In the context of the project 'Video for end-users' where technical aspects are investigated to enhance the video services that SURFnet and Kennisnet offer to the Dutch educational sector, Video Transcription can be explained as the automatic process of the conversion of speech and images into machine readable (meta)data. This includes optical character recognition (OCR) where printed text in images and in video is translated into machine-editable text. This machine-readable data can be used for captioning videos (subtitles) but also for extending the video metadata.

There is an important difference between transcription and captioning.

Transcript:

A transcript is a written text representation of spoken words not synchronized with the spoken words.

Captions:

Captions are the written text representation of spoken words synchronized with the spoken words.

This document describes the user-stories and uses cases to include such a Video Transcription service in adjunction with the video services SURFnet and Kennisnet already delivers to the Dutch educational sector, like SURFmediaⁱ and Teleblikⁱⁱ.

These end-user video services use the joint SURFnet and Kennisnet video back-end architecture VP-Core. VP-Core is the middleware video distribution platform that facilitates access to, and usage of (shared) storage capacity, metadata databases, transcoding- and streaming servers. VP-Core is empowered by MediaMosaⁱⁱⁱ software.

MediaMosa is an open source solution to build a full featured, webservice oriented, media management and distribution platform. MediaMosa is developed by Madcap commissioned by SURFnet and Kennisnet in an innovation and cooperation program.

Benefits

Besides the obvious benefits to viewers with hearing disabilities, transcription and captioning also offers a number of additional benefits to a much broader community of users that should not be overlooked:

- **Improved Accessibility:** Improved accessibility will make content more useful to a broader audience. Viewers with many types of Learning Disabilities will benefit from the increased comprehension and increased retention that captioning brings.
- **Indexing and Searching:** Because captioning involves the synchronization of text content with the audio/video material, it allows the content to become easily searchable with traditional text searches.
- **Flexibility:** With the increasing popularity of mobile devices, viewers may be in environments where access to the audio is limited. Captioning allows them to view your content whether they are in the library or on a noisy bus.
- **Localization** – adding translations to your captions, with support for multiple caption tracks, widens your potential audience massively.

Deliverables

The deliverable of this Technology Scouting project is a document, in the English language, with an overview of available transcription tools with their pro's and con's together with an advice how to integrate the best available technique into the current end-user video services, like SURFmedia and Teleblik, and how this integrates with the MediaMosa/VP-Core video backend.

The deadline to produce this document is May 1st 2010.

This report includes:

- An inventory of (open source) transcription and caption tools. Open source is definitely preferred, but the inventory should not be limited to open source only when there are good alternatives.
- Usability analysis of these tools in relation to SURFnets requirements and the video infrastructure.

This report will be used to form an advice on follow-up strategy that includes both front-end (SURFmedia/Teleblik) and back-end MediaMosa/VP-Core adjustments.

Requirements

This research is bound by the following requirements/conditions.

Subject of content to be transcribed/captioned is 'media' and includes:

- Video
- Audio
- Images

Topics included, but not limited to, are:

- Media Analysis tools to extract meaningful data from video and audio files.
- Video and Slide OCR that extracts time-synched metadata from slides.
- Captioning: Tools that enables closed captioning of media.
- Time-Synched Metadata.
- Serving Technology; how to use transcripts in combination with the player.

Preferred tools:

- Open Source
- PHP

User stories

- 1) A user uploads a video together with metadata. The video gets analyzed and processed. The result can be a combination of any:
 - The online video as is.
 - The online video with captions of the words spoken in the video.
 - The Metadata as is.
 - The Metadata extended with transcription data of the words spoken in the video.
 - A file with time-synched metadata containing a transcript of the words spoken in the video.
 - The uploaded video will have increased accessibility, will be easily searchable and will have improved flexibility.
- 2) A user uploads an audio fragment together with metadata. The audio gets analyzed and processed. The result can be a combination of any:
 - The online audio as is.
 - The Metadata as is.
 - The Metadata extended with transcription data of the words spoken in the audio fragment.

- A file with time-synced metadata containing a transcript of the words spoken in the audio fragment.
 - The uploaded audio fragment will have increased accessibility and will be easily searchable.
- 3) A user uploads an image together with metadata. The image gets analyzed and processed. The result can be a combination of any:
- The online image as is.
 - The Metadata as is.
 - The Metadata extended with transcription data of the words that have been OCR-ed from the image.
 - A file with time-synced metadata containing a transcript of the words OCR-ed from the image.
 - The uploaded image will have increased accessibility and will be easily searchable.

References

- Timed Text Markup Language (TTML) Version 1.0:
<http://www.w3.org/TR/2010/CR-ttbf1-dfxp-20100223/>
- OCRopus, document analysis and OCR system:
<http://code.google.com/p/ocropus>
- Sphinx-4, speech recognition engine:
<http://cmusphinx.sourceforge.net/sphinx4/>

ⁱ SURFmedia (Dutch): <http://www.surfmedia.nl/>

ⁱⁱ Teleblik (Dutch): <http://www.teleblik.nl/>

ⁱⁱⁱ MediaMosa: <http://mediamosa.org/>