Natural-Language-based Conversion of Images to Mobile Multimedia Experiences

Bernhard Reiterer[1], Cyril Concolato[2], and Hermann Hellwagner[1]

[1] Klagenfurt University, Universitaetsstr. 65-67, 9020 Klagenfurt, Austria,
firstname.lastname@uni-klu.ac.at,
WWW home page: http://www.uni-klu.ac.at
[2] TELECOM ParisTech, 46, Rue Barrault, 75013 Paris, France,
cyril.concolato@telecom-paristech.fr,
WWW home page: http://www.telecom-paristech.fr

Abstract. We describe an approach for viewing any large, detail-rich picture on a small display by generating a video from an the image, as taken by a virtual camera moving across it at varying distance. Our main innovation is the ability to build the virtual camera's motion from a textual description of a picture, e.g., a museum caption, so that relevance and ordering of image regions are determined by co-analyzing image annotations and natural language text. Furthermore, our system arranges the resulting presentation such that it is synchronized with an audio track generated from the text by use of a text-to-speech system.

Key words: image adaptation, text analysis, image annotation, digital cultural heritage, computer animation

1 Introduction

Images on the Web, from personal photography or any other source very often have resolutions higher than the displays they are to be shown on. For mobile devices, we commonly find image width and height each about ten times larger than the display. Such cases are usually dealt with in one of two disadvantageous ways: creating a smaller image, either by scaling, cropping or seam carving [1], which is a more recent technique, all of which discard a lot of pixel data, or navigating the image manually, e.g., as provided by the iPhone[1] or certain websites (see, e.g., [2]), which tends to be tedious for images at high scaling factors.

An emerging alternative family of approaches for viewing images on mobile devices is the automatic transmoding of images to videos, resulting from moving a virtual camera over the image at varying zoom levels, so that the most relevant regions are clearly visible one after the other. Such approaches, using different detection techniques for finding interesting regions, are shown in [3] and [4].

We argue that the steering of the virtual camera could consider more information than what can be extracted from the image. The main idea underlying

[1] http://www.apple.com/iphone/, accessed on 15 July 2009

our approach is that a natural language description of an image, such as a painting's caption in a museum, can be a valuable source for determining the relevance of image regions. Furthermore, if presented to the user in an appropriate way, image and text can augment each other: the text explains the image, while the image illustrates the text. So, by transforming the image to a video and presenting the text in a synchronized way, either spoken or as subtitles, we seek to enhance the user experience in comparison to alternative methods.

2 Natural-Language-based Image Transmoding

2.1 System overview

Figure 1 shows a high-level overview of our Natural-Language-based Image Transmoding Engine. The core step, called *Virtual Camera Control* (VCC), is responsible for generating the script for the virtual camera, instructing the camera when to show which rectangle of the image. This step takes image annotations, text and various constraints as its input. Internally, VCC creates *Sync Points*, tuples that link positions to be synchronized across different media, from the input data and refines them until all relevant aspects are considered.

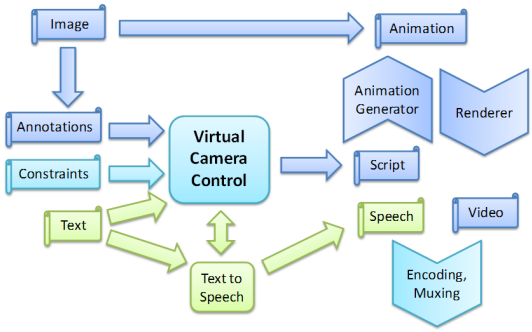

Fig. 1. Overview of the Natural-Language-based Image Transmoding Engine

Image annotations, either created manually or from computer vision systems, consist of shapes of regions of interest (ROIs) annotated with keywords. Constraints to be considered include the display size, the maximum video duration, and limits ensuring a pleasurable camera motion. A text-to-speech system, currently a MARY[2] server, produces audio and timing data from text.

From the script, an SVG [5] animation is generated. In the finalization steps, which, unless configured otherwise, delegate to software of GPAC [6] and FFmpeg[3], the animation is rendered to a video, encoded (by far the biggest

[2] `http://mary.dfki.de/`, accessed on 15 July 2009
[3] `http://ffmpeg.org`, accessed on 15 July 2009

part of the execution time) and multiplexed with the encoded audio track just generated, yielding an audiovisual file ready for playback by the user.

We provide a growing set of example inputs and results on our project website[4], illustrating the functionality and rich configurability of the system so far.

2.2 Matching and synchronization

The initial implementation of VCC is realized by looking up keywords from ROI annotations in the text. For each match between a ROI's keyword and the text, a Sync Point is generated. This will be enhanced by more powerful text analysis in order to allow for annotations on a lower level and thus to allow for more automatic analysis, e.g., by interpreting descriptions of visual features (mainly shapes and colours) or spatial relations (e.g., "left of", "in the center").

For making the text perceivable to the user, options are displaying it on the screen, either as subtitles or as running text, or transforming it to synthetic speech. The latter has the advantage of not using any display space, the resource whose limitation gives the main motivation for our work. The others are fallback solutions for cases in which the user cannot or prefers not to play audio, or for supporting languages for which the system is not able to generate speech.

We use text-to-speech preprocessing results for assigning a time value to each position in the text, thus enriching the previously generated Sync Points by precise timing. Unpleasantly short or long durations for motions or stays are resolved by slightly delaying Sync Points, discarding less relevant ROIs (e.g., repeated ones) or by adding unvisited ROIs, respectively.

3 Typical Application

As one possible field of application, our system could be incorporated into the website of a museum for helping users decide which parts of the museum to see: The museum staff annotates an image once, and different texts for multiple target groups (language, age, expert level or interests, e.g., artistic style, historical background) are provided per image. On the website, users give the criteria for selecting among those texts, along with their choice of potentially interesting parts of the museum and the time they want to spend for watching the video. With the users' approval, the server automatically receives information about the terminal's relevant properties such as display resolution and capabilities for decoding and playback. The users then receive individually generated videos presenting the highlight paintings of the selected museum wings.

Other potential applications include maps and directions, social Websites, slide shows for collections of personal photography, which are currently often generated manually, and also image viewing as a side activity, where a user's visual attention and navigation capabilities are hampered.

[4] `http://www.itec.uni-klu.ac.at/~reiterer/dawnlite/showcase.php`, accessed on 15 July 2009

4 Conclusion and Future Work

We presented our approach for turning a high-resolution image to audiovisual content for constrained devices by generating a video synchronized with synthesized speech after matching a descriptive text with image annotations. This way, we hope to preserve and even enhance the user's experience.

To get the most benefit from our approach, we will put the main focus of our future work on its most innovative component, the matching of text to image regions and the underlying text analysis, weakening the necessity of manual image annotations. However, any image analysis system (e.g., detection of persons, faces or certain objects) could be used with our system already, leading to functionality analogue to related work.

Along with linguistic improvements, the system should be enabled to rearrange the speech (e.g., inserting breaks, moving or erasing phrases) in order to even the temporal distribution of ROIs. Furthermore, if available, pre-existing audio should be usable, either by speech recognition or via textual transcripts.

The application of the text-based virtual camera control to 3D content instead of images seems feasible with manageable effort, since the software used for rendering now is already capable of handling 3D worlds. Example use cases would be illustrating text from natural sciences and 3D navigation systems.

Acknowledgements

This work is supported by the NoE INTERMEDIA funded by the European Commission (NoE 038419).

References

1. Ariel Shamir and Shai Avidan. Seam Carving for Media Retargeting. *Commun. ACM*, 52(1):77–85, 2009.
2. Hard Rock Memorabilia. URL: `http://memorabilia.hardrock.com/`, accessed on 15 July 2009.
3. Pedro Pinho, Joel Baltazar, and Fernando Pereira. Integrating Low-Level and Semantic Visual Cues for Improved Image-to-Video Experiences. In Aurélio C. Campilho and Mohamed S. Kamel, editors, *ICIAR (2)*, volume 4142 of *Lecture Notes in Computer Science*, pages 832–843. Springer, 2006.
4. Fernando Barreiro Megino, José M. Martínez Sánchez, and Víctor Valdés López. Virtual Camera Tools for an Image2Video Application. In *WIAMIS '08: Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 223–226, Washington, DC, USA, 2008. IEEE Computer Society.
5. Scalable Vector Graphics (SVG). URL: `http://www.w3.org/Graphics/SVG/`, accessed on 15 July 2009.
6. Cyril Concolato, Jean Le Feuvre, and Jean-Claude Moissinac. Design of an efficient scalable vector graphics player for constrained devices. *Consumer Electronics, IEEE Transactions on*, 54:895–903, 2008.