"Hey, vitrivr!" – A Multimodal UI for Video Retrieval

Prateek Goel*, Ivan Giangreco, Luca Rossetto,
Claudiu Tănase, and Heiko Schuldt

Databases and Information Systems Research Group,
Department of Mathematics and Computer Science, University of Basel, Switzerland
{ivan.giangreco|luca.rossetto|c.tanase|heiko.schuldt}@unibas.ch

Abstract. In this paper, we present a multimodal web-based user interface for the vitrivr system. vitrivr is a modern, open-source video retrieval system for searching in large collections of video using a great variety of query modes, including query-by-sketch, query-by-example and query-by-motion. With the multimodal user interface, prospective users benefit from being able to naturally interact with the vitrivr system by using spoken commands and also by applying multimodal commands which combine spoken instructions with manual pointing. While the main strength of the UI is the seamless combination of speech-based and sketch-based interaction for multimedia similarity search, the speech modality has shown to be very effective for retrieval on its own. In particular, it helps overcoming accessibility boundaries and offering retrieval functionality for users with disabilities. Finally, for a holistic natural experience with the vitrivr system, we have integrated a speech synthesis engine that returns spoken answers to the user.

1 Introduction

In recent years, the conventional approach of a user facing a mouse or keyboard for giving inputs to a system has increasingly been challenged by a –more natural– multimodal approach which considers the interaction on both the auditory and visual level. Early work in [1] presents an approach to combining speech and gesture input within a general setting. In the context of search applications, the authors of [2], for instance, show promising results when combining multiple modalities, e.g., the fusion of speech and gesture inputs in the context of a search application.

In this paper, we present a multimodal user interface for the vitrivr [3] video retrieval system which supports both manual pointing and voice commands (alone and in combination) to enhance the user experience. The vitrivr system [3] is a modern open-source video retrieval system which offers users a great variety of query paradigms, including query-by-sketch, query-by-example and query-by-motion for searching in large collections of video. It is powered by the ADAM$_{pro}$ database and the Cineast retrieval engine.

* Prateek Goel has been a Google Summer of Code '16 student with the vitrivr project.

The key advantages of the multimodal user interface include a gentle learning curve, an increased efficiency and expressiveness given by the voice interface, which seems to be helpful particularly for novice users, and a natural interaction with the system. Moreover, thanks to the accessibility of the UI, we see a strong use case for users with disabilities, as our approach does not require the explicit use of any pointing device.

2 Multimodal vitrivr UI

The primary goals of the vitrivr UI (see Figure 1) are to scaffold a query and to present the retrieved results to the user. For the first task, in the vitrivr front-end, queries are specified using one or multiple canvases which can be used to either sketch a query or use an existing input image. The visual information can be further enriched by specifying motion, e.g., to denote the motion of an object within the scene. For presenting results to a user, the vitrivr UI displays a result list of similar shots sorted by relevance.

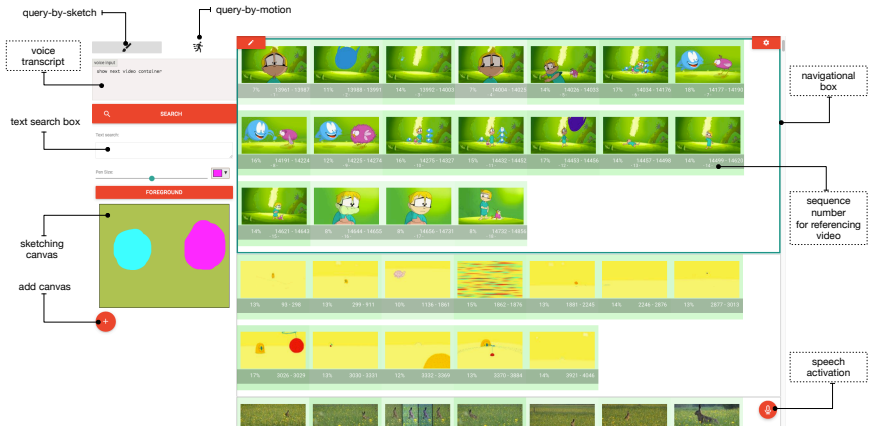

Fig. 1. Screenshot of the multimodal vitrivr UI. The explanations highlighted with a border are relevant for the speech-based UI.

Speech Interaction: We have carefully hand-crafted an ontology along the lines of the goals of the UI to support in total over 50 actions (e.g., perform a search, choose a specific color, etc.) which can be executed by more than 250 predefined, alternative spoken commands. The matching of spoken text to a command is done either based on exact rule matching (which may contain optional words, however). Furthermore, to allow for a fuzzy matching, for all sentences which could not be exactly matched, we have added a matching strategy based on n-grams; with this, we execute a command although there is no exact matching if the matching score is above a certain threshold.

The implemented speech commands allow to specify a query, e.g., by adding keywords to the textual search, by filling the canvas with a specific color, by adjusting the pen size etc. On the other hand, the system provides navigational commands to allow the user to navigate within the result list, e.g., move down the result list, play a specific result elements and hide/highlight certain results, for instance, based on the score. For navigation, after displaying the results to the user, a navigating box is displayed, marking the current video considered, with the most similar result shots within the video being displayed and numerated. With the enumeration, a user can quickly refer to a specific result list element, e.g., by saying "play video of shot number two".

We designed the speech commands to support a natural interaction with the vitrivr system. Follow-up commands, for instance, relate to the previous command, by keeping track of the previously executed command. Consider the action of increasing the pen size: Naturally, a user would first say to the system, e.g. "increase the pen size"; for further increasing the pen size, however, repeating the command would not seem natural. Hence, the UI tracks the executed commands to support follow-up commands, such as "even more" (to adjust the pen size) or "even further" (to move within the results).

Multimodal Interaction: The spoken commands of vitrivr can be enriched by using other modalities, i.e., combining manual pointing and voice commands. We use a simple model for such a multimodal interaction: A spoken query must be followed within a short time frame (e.g., within 5 seconds) by a pointing action for correctly executing and recognising a multimodal intent. A user can for instance say "play this video" and by that start a timer which expects a click –within the predefined time frame– on a specific result item, which is then played. Similarly, she can add results to relevance feedback by saying "add these videos to positive feedback" and click on multiple videos.

Conversational Feedback: To significantly enhance the natural interaction with the system, a speech synthesis engine in the user interface responds to the commands in a conversational way. For instance, the speech engine can confirm the understanding of a command ("Ok"), give an answer to a specific question as a result of a specific command ("There are hundred results."), or ask to repeat the command in a more specific way if it was not understood ("Did you mean...?"). The latter will be used, if the recognition of the spoken command fails, but the similarity of a command within the ontology and the spoken words is greater than a certain threshold when comparing the n-grams.

3 Implementation

The vitrivr UI is browser-based. The speech recognition is implemented using annyang[1] which works on top of the W3C Web Speech API [4] offering an API for speech recognition in modern browsers. It parses the recognised words to

[1] https://www.talater.com/annyang/

commands using regular expressions. To allow for a fuzzy matching, we have built a matching strategy based on 3-grams on top, which produces a similarity score of the spoken command and a command in the ontology.

The code is available in the vitrivr open-source project.[2]

4 vitrivr in Action

vitrivr uses a large collection of free, creative commons web video. Users are able to search this collection in the context of a known-item search task with the multimodal UI, using a microphone and a mouse/pen input[3].

For this, users are able to browse through a list of present videos in the database, select a target sequence which they would like to re-find using the vitrivr system and apply the search paradigms offered: Starting, for instance, from a hand-drawn sketch by choosing via the speech-interface the brush tool and the color, and use query-by-sketching for retrieving the most similar video snippets in the database. They can, then, for example, choose a result from the result list which appears similar to the query (again using a voice commands) and use it for further searching the system. Finally, the user(s) can navigate using multimodal commands through the results and play video scenes to see if the query scene has been found.

Acknowledgments

This work was partly funded by the Swiss National Science Foundation (SNSF) in the context of the Chist-Era program IMOTION (contract no. 20CH21_151571), and the Google Summer of Code 2016 program.

References

1. Richard A. Bolt. "Put-that-there": Voice and Gesture at the Graphics Interface. In *Proc. Int. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH 1980)*, pages 262–270, Seattle, USA, 1980. ACM.
2. Larry Heck, Dilek Hakkani-Tür, Madhu Chinthakunta, Gokhan Tur, Rukmini Iyer, Partha Parthasacarthy, Lisa Stifelman, Elizabeth Shriberg, and Ashley Fidler. Multimodal Conversational Search and Browse. In *Proc. Ws. on Speech, Language and Audio in Multimedia*, Marseille, France, August 2013. IEEE.
3. Luca Rossetto, Ivan Giangreco, Claudiu Tănase, and Heiko Schuldt. vitrivr: A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections. In *Proc. Int. Conf. Multimedia (ACM MM 2016)*, pages 1183–1186, Amsterdam, Netherlands, November 2016. ACM.
4. Glen Shires and Hans Wennborg. Web Speech API Specification. W3C community group final report, W3C, October 2012. https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html, Accessed: 2017-01-11.

[2] http://vitrivr-ui.vitrivr.org
[3] A video demoing the system can be found on http://youtu.be/GCqxJ6FMlH0.