iAutoMotion – an Autonomous Content-based Video Retrieval Engine

Luca Rossetto[1], Ivan Giangreco[1], Claudiu Tănase[1], Heiko Schuldt[1],
Stéphane Dupont[2], Omar Seddati[2], Metin Sezgin[3], and Yusuf Sahillioğlu[3]

[1] Databases and Information Systems Research Group,
Department of Mathematics and Computer Science, University of Basel, Switzerland
{luca.rossetto|ivan.giangreco|c.tanase|heiko.schuldt}@unibas.ch
[2] Research Center in Information Technologies, Université de Mons, Belgium
{stephane.dupont|omar.seddati}@umons.ac.be
[3] Intelligent User Interfaces Lab, Koç University, Turkey
{mtsezgin|ysahillioglu}@ku.edu.tr

Abstract. This paper introduces iAutoMotion, an autonomous video retrieval system that requires only minimal user input. It is based on the video retrieval engine IMOTION. iAutoMotion uses a camera to capture the input for both visual and textual queries and performs query composition, retrieval, and result submission autonomously. For the visual tasks, it uses various visual features applied to the captured query images; for the textual tasks, it applies OCR and some basic natural language processing, combined with object recognition. As the iAutoMotion system does not conform to the VBS 2016 rules, it will participate as unofficial competitor and serve as a benchmark for the manually operated systems.

1 Introduction

The stated goal of the Video Browser Showdown (VBS) [4] is to evaluate *interactive* video search systems. This is based on the implicit assumption that an interactive video search system which involves its users directly and intensely provides benefits in terms of flexibility, accuracy, and/or speed over more traditional, yet less interactive approaches. According to the general rules of the VBS 2016 competition, only interactive video search systems are entitled to participate. However, in order to evaluate the implicit assumptions according to which interactivitiy actually also implies flexibility, accuracy, and/or performance, we have developed the iAutoMotion system, an autonomous video retrieval engine.

In this paper, we present the iAutoMotion system which is designed to solve all types of tasks of the VBS with as little user interaction as possible. Intentionally, this system violates some of the rules of VBS. It will thus not participate as a regular competitor but instead, it is intended to serve as a performance benchmark so as to be able to evaluate the interactive systems against the autonomous one which is operated without a user in the loop.

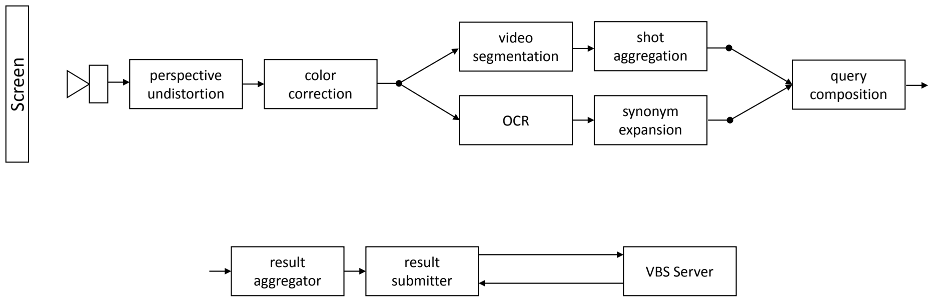

Fig. 1. Architectural Overview of iAutoMotion

The iAutoMotion system uses the same back-end as the iMotion system [3] but acquires all necessary information for query formation with a webcam. We detail the system architecture and in particular the processing pipeline of iAutoMotion in Section 2 and discuss the remaining, rather rudimentary user interaction, mainly for configuration, in Section 3. Section 4 briefly describes the implementation of iAutoMotion and Section 5 concludes.

2 System Architecture

This section describes the overall architecture and processing pipeline of iAutoMotion as depicted in Figure 1. The input video which is recorded using a webcam goes through multiple stages of processing which are explained below.

2.1 Perspective undistortion and color correction

The first part of the pipeline aims at restoring the image data originating from the webcam to its original form. It therefore has to correct for the perspective distortion created by the distance of the webcam from the center of the screen. The subsequent step in the pipeline changes color properties to compensate for factors such as illumination to optimize saturation and contrast of the acquired images as well as ensuring their accuracy in hue. This part of the pipeline requires a one-time manual calibration when iAutoMotion is set up.

2.2 Visual search and text search

The processing pipeline comes in two variants: one is tailored to the visual search task where a sequence that needs to be found is shown. The second variant is optimized for the textual search in which a textual description of the search task is given. In what follows, we describe both alternatives of the query pipeline. The switch from one mode (visual or textual) to the other is again one of the few manual tasks in iAutoMotion.

Video segmentation and shot aggregation (Visual search): The part of the pipeline that is specialized on the visual tasks of the competition consists of a video segmentation and a shot aggregation module. The task of the video segmentation module is to segment the input video into visually continuous sequences which are then aggregated by the shot aggregator. This process reduces redundant queries by ensuring that only one query is performed for every shot in the input sequence while simultaneously reducing the chances of missing visual information contained in a short shot. The resulting shot aggregates are handed off to the query composition module which then queries the retrieval back-end.

OCR and semantic text expansion (Textual search): To be able to perform queries for the textual tasks of the competition, the query text has to be acquired using optical character recognition. Once the text has been obtained, all nouns are extracted from it. Because the retrieval back-end is only able to recognize a limited number of objects and concepts, synonyms and hypernyms have to be considered as well. The resulting list of detectable objects is passed to the query composition module.

2.3 Query composition

The query composition module receives data from either of the two sub-chains and builds one or multiple query objects which are then sent to the retrieval back-end. The used retrieval back-end is identical to the one used in the IMOTION system [3].

2.4 Result aggregator and submitter

The result aggregator takes all results returned by the retrieval back-end and combines them into a list of sequences which match the original input sequence. Every sequence is processed and a sub-sequence is selected which is guaranteed not to exceed the restrictions in duration imposed on all query sequences. All these results are sorted by score in decreasing order. The result submitter takes the top-n sequences and submits them one after the other to the evaluation server until it either receives feedback that the submitted sequence was indeed the correct one or the number of submitted sequences exceeds the threshold n. Because of the scoring scheme used by the evaluation server, submitting further sequences in the hope of hitting the correct one becomes quickly pointless. Therefore, for VBS n can be set to a value around or below 10.

3 User interaction

Even tough the system is designed to solve the posed challenges autonomously, it still requires a minimal amount of user interactions.

3.1 Setup and configuration

The main interaction with a human is required during the initial setup. The camera has to be mounted stably and with a clear view on the screen on which the queries of either type are going to be displayed. The bounds of this screen need to be selected to serve as input for the calibration procedure of the perspective undistortion module. Color correction, while not requiring manual input to work, can be configured to improve its results. Finally, the network needs to be set up in order to enable the system to communicate with the competition server for result submission.

3.2 Query initialization

During the competition itself, the system needs to know when a query starts and ends, and what type of query – visual or textual – should be processed. This can be signalized using a mouse or a keyboard button which needs to be held for the duration of the query. Depending on which button is pressed, the system will either use the processing sub-pipeline for the visual or the one for the textual task.

4 Implementation

The iAutoMotion system is written in Java and uses BoofCV [1] for image processing and RiTa[1] with WordNet [2] for synonym expansion. It communicates directly with the retrieval back-end bypassing the IMOTION web server. This can be done because the autonomous front-end has no need for static content such as preview images or video files.

5 Conclusions

In this paper, we have presented iAutoMotion, a video retrieval system that is fully automated at retrieval time and that just needs some manual configuration at set-up time. It will participate as unofficial competitor to VBS 2016 and serves as a benchmark for the manually operated systems.

Acknowlegements

This work was partly supported by the Chist-Era project IMOTION with contributions from the Belgian Fonds de la Recherche Scientifique (FNRS, contract no. R.50.02.14.F), the Scientific and Technological Research Council of Turkey (Tübitak, grant no. 113E325), and the Swiss National Science Foundation (SNSF, contract no. 20CH21_151571).

[1] https://rednoise.org/rita/

References

1. Peter Abeles. Boofcv. http://boofcv.org/, 2012.
2. George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
3. Luca Rossetto, Ivan Giangreco, Silvan Heller, Claudiu Tănase, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, Ozan Can Altiok, and Yusuf Sahillioğlu. IMOTION - searching for video sequences using multi-shot sketch queries. In *MultiMedia Modeling*. Springer, 2016.
4. Klaus Schoeffmann, David Ahlström, Werner Bailer, Claudiu Cobârzan, Frank Hopfgartner, Kevin McGuinness, Cathal Gurrin, Christian Frisson, Duy-Dinh Le, Manfred Del Fabro, et al. The video browser showdown: a live evaluation of interactive video search tools. *International Journal of Multimedia Information Retrieval*, 3(2):113–127, 2014.