

Interactive Video Retrieval

Cees Snoek

AERFAISS 2010 Summer School

Interactive Video Retrieval

In this lecture we focus on the challenges in video search, present methods how to achieve state-of-the-art performance, and indicate how to obtain improvements in the near future. Moreover, we give an overview of the latest developments and future trends in the field on the basis of the TRECVID competition – the leading competition for video search engines run by NIST.

The scientific topic of video search is dominated by five major challenges:

- a. the sensory gap between an object and its many appearances due to the accidental sensing conditions;
- b. the semantic gap between a visual concept and its lingual representation;
- c. the model gap between the amount of notions in the world and the capacity to learn them;
- d. the query-context gap between the information need and the possible retrieval solutions;
- e. the interface gap between the tiny window the screen offers to the amount of data;

The semantic gap is bridged by forming a dictionary of visual concept detectors. The largest ones to date consist of hundreds of concepts excluding concept-tailored algorithms. It would simply take too long to achieve. Instead, we come closer to the ideal of one computer vision algorithm tailored automatically to the purpose at hand by employing example data to learn from. We discuss the advantages and limitations of a machine learning approach from examples. We show for what type of concept the approach is likely to succeed or fail. In compensation for the absence of concept-specific (geometric or appearance) models, we emphasize the importance of a good feature sets. They form the basis of the observational model by all possible color, shape, texture or structure invariant features help to characterize the concept at hand. Apart from good features, the other essential component is state-of-the-art machine learning in order to get the most out of the learning data.

We integrate the features and machine learning aspects into a complete concept-based video search engine, which has successfully competed in TRECVID. The system includes computer vision, machine learning, information retrieval, and human-computer interaction. We follow the video data as they flow through the computational processes. Starting from fundamental visual features, covering local shape, texture, color, motion and the crucial need for invariance. Then, we explain how invariant features can be used in concert with kernel-based supervised learning methods to arrive at a concept detector. We discuss the important role of fusion on a feature, classifier, and semantic level to improve the robustness and general applicability of detectors. We end our component-wise decomposition of video search engines by explaining the complexities involved in delivering a limited set of uncertain concept detectors to an impatient user. For each of the components we review state-of-the-art solutions in literature, each having different characteristics and merits.

Comparative evaluation of methods and systems is imperative to appreciate progress. We discuss the data, tasks, and results of TRECVID, the leading benchmark. In addition, we discuss the many derived community initiatives in creating annotations, baselines, and software for repeatable experiments. We conclude the course with our perspective on the many challenges and opportunities ahead for the multimedia pattern recognition community.