Semantic Mapping in Video Retrieval

Maaike H.T. de Boer
TNO & Radboud University
maaike.deboer@tno.nl

Abstract

In the modern world, networked sensor technology makes it possible to capture the world around us in real-time. In the security domain cameras are an important source of information. Cameras in public places, bodycams, drones and recordings with smart phones are used for real time monitoring of the environment to prevent crime (*monitoring case*); and/or for investigation and retrieval of crimes, for example in evidence forensics (*forensic case*). In both cases it is required to quickly obtain the right information, without having to manually search through the data. Currently, many algorithms are available to index a video with some pre-trained concepts, such as people, objects and actions. These algorithms require a representative and large enough set of examples (training data) to recognize the concept. This training data is, however, not always present.

In this thesis, we aim to assist an analyst in their work on video stream data by providing a search capability that handles ad-hoc textual queries, i.e. queries that include concepts or events that are not pre-trained. We use the security domain as inspiration for our work, but the analyst can be working in any application domain that uses video stream data, or even indexed data. Additionally, we do only consider the technical aspects of the search capability and not on the legal, ethical or privacy issues related to video stream data. We focus on the retrieval of high-level events, such as birthday parties. We assume that these events can be composed of smaller pre-trained concepts, such as a group of people, a cake and decorations and relations between those concepts, to capture the essence of that unseen event (decompositionality assumption). Additionally, we hold the open world assumption, i.e. the system does not have complete world knowledge. Although current state of the art systems are able to detect an increasingly large number of concepts, this number still falls far behind the near infinite number of possible (textual) queries that a system needs to be able to handle.

In our aim to assist the analyst, we focus on the improvement of the visual search effectiveness (e.g. performance) by a semantic query-to-concept mapping: the mapping from the user query to the set of pre-trained concepts. We use the TRECVID Multimedia Event Detection benchmark, as it contains high-level events inspired by the security domain. In this thesis, we show that the main improvements can be achieved by using a combination of i) query-to-concept mapping based on semantic word embeddings (+12%), ii) exploiting user feedback (+26%) and iii) fusion of different modalities (data sources) (+17%).

First, we propose an incremental word2vec (**i-w2v**) method [1], which uses word2vec trained on GoogleNews items as a semantic embedding model and incrementally adds concepts to the set of selected concepts for a query in order to deal with query drift. This method improve performance in terms of MAP compared to the state of the art word2vec method and knowledge based techniques. In combination with a state of the art video event retrieval pipeline, we achieve top performance on the TRECVID MED benchmark regarding the zero-example task (MED14Test results). This improvement is, however, dependent on the availability of the concepts in the Concept Bank: without concepts related to or occurring in the event, we cannot detect the event. We, thus, need a properly composed Concept Bank to properly index videos.

Second, we propose an Adaptive Relevance Feedback interpretation method named **ARF** [2] that not only achieves high retrieval performance, but is also theoretically founded through the Rocchio algorithm from the text retrieval field. This algorithm is adjusted to the event retrieval domain in a way that the weights for the concepts are changed based on the positive and negative annotations on videos. The ARF method has higher visual search effectiveness compared to k-NN based methods on video level annotations and methods based on concept level annotations.

Third, we propose blind late fusion methods that are based on state of the art methods [3], such as average fusion or fusion based on probabilities. Especially the combination of a Joint Ratio (ratio of probabilities) and Extreme Ratio (ratio of minimum and maximum) method (**JRER**) achieves high performance in cases with reliable detectors, i.e. enough training examples. This method is not only applicable to the video retrieval field, but also in sensor fusion in general.

Although future work can be done in the direction of implicit query-to-concept mapping through deep learning methods, smartly combining the concepts and the usage of spatial and temporal information, we have shown that our proposed methods can improve the visual search effectiveness by a semantic query-to-concept mapping which brings us a step closer to a search capability that handles ad-hoc textual queries for analysts.

References

[1] Maaike de Boer, Yi-Jie Lu, Chong-Wah Ngo, Klamer Schutte, Wessel Kraaij, Zhang Hao (2017) **Semantic Reasoning in Zero Example Video Event Retrieval**. *To appear: Transactions on Multimedia Computing, Communications, and Applications*.

[2] Maaike de Boer, Geert Pingen, Douwe Knook, Klamer Schutte and Wessel Kraaij (2017) **Improving Video Event Retrieval by User Feedback**. *Multimedia Tools and Applications* pp. 1 - 21, DOI: 10.1007/s11042-017-4798-3.

[3] Maaike H.T. de Boer, Klamer Schutte, Hao Zhang, Yi-Jie Lu, Chong-Wah Ngo, Wessel Kraaij (2016) **Blind late fusion in multimedia event retrieval** . *International Journal of Multimedia Information Retrieval*, vol. 5, pp. 203 - 217.