In Search of Video Event Semantics

M. Mazloom

Summary in English:

In search of video event semantics

In this thesis we aim to represent an event in a video using semantic features. We start from a bank of concept detectors for representing events in video. At first we considered the relevance of concepts to the event inside the video representation. Then, we concentrated on the accuracy of concept detectors. Finally, we consider the problem of searching video events with and without semantic concepts. Before reviewing the questions considered in the introduction, we will give a brief overview of the chapters.

In Chapter 2, we address the problem of video event classification using a bank of concept detector scores. Different from existing work, which simply relies on a bank containing all available detectors, we propose an algorithm that learns from examples what concepts in a bank are most informative per event, which we call the conceptlets. We model finding the conceptlets out of a large set of concept detectors as an importance sampling problem. Our proposed approximate algorithm finds the optimal conceptlets using a cross-entropy optimization. We study the behavior of video event classification based on conceptlets by performing four experiments on challenging internet video from the 2010 and 2012 TRECVID multimedia event detection tasks and Columbia's consumer video dataset. Starting from a concept bank of more than thousand pre-computed detectors, our experiments establish there are (sets of) individual concept detectors that are more discriminative and appear to be more descriptive for a particular event than others, event classification using an automatically obtained conceptlet is more robust than using all available concepts, and conceptlets obtained with our cross-entropy algorithm are better than conceptlets from state-of-the-art feature selection algorithms. What is more, the conceptlets make sense for the events of interest, without being programmed to do so.

In Chapter 3 we propose a new semantic video representation for few and zero example event detection and unsupervised video event summarization. Different from existing works, which obtain a semantic representation by training concepts over images or entire video clips, we propose an algorithm that learns a set of relevant frames as the concept prototypes from web video examples, without the need for frame-level annotations, and use them for representing an event video. We formulate the problem of learning the concept prototypes as seeking the frames closest to the densest region in the feature space of video frames from both positive and negative training videos of a target concept. We study the behavior of our video event representation based on concept prototypes by performing three experiments on challenging web videos from the TRECVID 2013 multimedia event detection task and the MED-summaries dataset. Our experiments establish that i) Event detection accuracy increases when mapping each video into concept prototype space. ii) Zero example event detection increases by analyzing each frame of a video individually in concept prototype space, rather than considering the holistic videos. iii) Unsupervised video event summarization using concept prototypes is more accurate than using video-level concept detectors.

In Chapter 4 we aim at querying web videos for complex events using only a handful of video query examples, where the standard approach learns a ranker from hundreds of examples. We consider a semantic signature representation, consisting of off-the-shelf concept detectors, to capture the variance in semantic appearance of events. Since it is unknown what similarity metric and query fusion to use in such an event retrieval setting, we perform three experiments on unconstrained web

videos from the TRECVID 2012 event detection task dataset. It reveals that retrieval with semantic signatures using normalized correlation as similarity metric is more accurate and faster than a low-level bag-of-words alternative, multiple queries are best combined using late fusion with an average operator, and event retrieval is preferred over event classification when less than eight positive video examples are available. We demonstrate a capability to yield semantic signature in a video search engine. We show how the semantic signatures provide a crude interpretation on why a certain video has been retrieved.

The aim of the Chapter 5 is event detection in video for scenarios where only few, or even zero example is available for training. For this challenging setting, the prevailing solutions in literature rely on a semantic video representation obtained from thousands of pre-trained concept detectors. Different from existing work, we propose a new semantic video representation that is based on freely available social tagged videos only, without the need for training any intermediate concept detectors. We introduce a simple algorithm that propagates tags from a video's nearest neighbors, similar in spirit to the ones used for image retrieval, but redesign it for video event detection by including video source set refinement and varying the video tag assignment. We call our approach TagBook and study its construction, descriptiveness and detection performance on the TRECVID 2013 and 2014 multimedia event detection datasets and on the Columbia Consumer Video dataset. Despite its simple nature, the proposed TagBook video representation is remarkably effective for few-example and zero-example event detection, even outperforming very recent state-of-the-art alternatives building on supervised representations.