Multimedia Information Extraction Roadmap

Alexander G. Hauptmann

School of Computer Science - Carnegie Mellon University
Pittsburgh, PA 15213 USA

Critical Technical Challenges

The broad challenge in my view is to exploit multi-lingual, multimedia information from both web and TV video to allow broader understanding of different ideological, social, and cultural perspectives in different sources, for a wide variety of applications. This will involve the judicious analysis of the text and video features using a variety of machine learning and language analysis methods, as well as understanding of the video editing structure, as well as the context in which the media appears. Other challenges involve dealing with the flood of data through mechanisms for intelligent, context-driven summarization, as our brains remain limited in the amount of information they can process. Yet different challenges concern the mobile use of multimedia data, considering limited bandwidth, small displays, limited multimedia input mechanisms as well as the social network that provides the context of the media – however, the latter challenge won't be addressed in the following paragraphs.

At this point, the infrastructure for collecting massive amounts of multimedia and sensor data exists, and processing hardware is becoming relatively affordable. The open research problem is how to combine and aggregate this large amount of multimedia input and make it useful in a broad sense, for example for searching and browsing. This is one of the big challenges we face today – how to create semantically meaningful, general knowledge from multimedia data streams.

One key challenge is to understand the content of an image or video. The currently most plausible way to attack this is to develop thousands of automatically detectable semantic concepts in the video or images as a core vocabulary for highly accurate video analysis applications. Since user information needs are defined in terms of high-level, often abstract concepts, we will need thousands of mid-level 'semantic concepts' that can reliably characterize video or sensor data into classes such as outdoors, vehicles, people, running, airplanes, trees, shaking hands, and so forth, to bridge this semantic gap. The challenge is to define a good set or a taxonomy of visible things or concepts, perhaps even an ontology of visual descriptions of useful and demonstrate that we can detect these in the multimedia sensor stream with acceptable accuracy and as a result bridge the semantic gap for many useful tasks.

Existing Approaches

There is active ongoing work in semantic concept detection fro both still images and video data. These usually rely on manually created training data, that provides semantic labels for images or video at varying level of detail, sometimes at the level of a complete video clip, sometimes at the level of a shot, a single image or keyframe, and sometimes at the level of a identified region within an image or keyframe. From this data a number of 'features' are extracted, which describe video/image in terms of low-level, unambiguously computable features. These features can range from moment or histograms in different color spaces, interest points that indicate changes or boundaries, motion estimates of points, regions or overall image changes, texture and edge estimates, shape estimates, etc. This, usually high-dimensional feature data together with the 'truth' annotation is then fed into a wide variety of machine learning approaches, which is the most active area of current research, all of which try to separate data with one type of label from the other labeled data, through various projections, data reductions and generalizations. The objective is always to generalize to other things that have the same label, even if they were not in the training set, while avoiding false positives and missed detections.

Methods and Techniques

First of all, the data being exploited for information extraction needs to be scaled to much larger sizes. In comparison to speech recognition, good, recognition performance was not achieved until hundreds of hours of training data from the appropriate domains became available. Thus, more data is needed, more annotations, and more semantic concepts to provide labels for the data. What is also needed is to scale existing approaches so that can deal with datasets that are orders of magnitude larger than what has been studied so far, except perhaps inside some of the search or service provider companies.

Furthermore, the learning approaches will have to adapt to errorful labels, also known as 'weak annotations'. The most promising directions are efforts related to multiple instance learning and graphical models that discover reliable hidden labels based the existing weak annotations.

In addition, perhaps the biggest need is to explore approaches that provide good generalization across different domains, because we want to learn how to identify a concept like a chair in any environment, not just a certain type of chair when it's next to a table.

DataSets

The big 800 pound gorilla of weakly annotated multimedia data is obviously provided by the Web with wide accessibility to large amounts of all types of data. It will be the upcoming challenge It is mind boggling to consider the amount of data available on photo and video sharing websites. For example, Flickr has more than 2 billion images, with tags, groups, and descriptions, and YouTube is adding at least tens of thousand of videos daily.

However, most of this data is not quite in the right form, and each particular page doesn't contain all the right attributes to make it suitable for exploitation.

Remaining Gaps

There will likely still be semantic concepts that will not be detectable with sufficient accuracy merely by providing large amounts of labeled training data. For these, special world knowledge must be applied and exploited – along the lines of what computer vision research has focused on for many years. This requires specialized research that develops special features, exploits real-world constraints and provides high accuracy recognition. Perhaps the most successful example has been face detection, where after a focused research effort has yielded excellent results as evidenced in commercially available products.

As mentioned earlier, significant research is also needed to handle the vast amounts of available and ***relevant*** multimedia data through mechanisms for intelligent, context-driven summarization, since our human brains and ability to focus our attention will continue to be a major bottleneck. Research over the next few years, much of which has already started, must address the mobile use of multimedia data, consider its limited transmission bandwidth, miniature displays, and restricted input methods. In addition, a gap remains in the exploitation of the social network that constitutes the scaffolding and context of the media. I expect we will see significant progress in all these areas in the next five years.