Aligning Textual & Visual Data: Towards Scalable Multimedia Retrieval

Pramod Sankar Kompalli
200499002

The search and retrieval of relevant images and videos from large repositories of multimedia, is acknowledged as one of the hard challenges of computer science. With existing pattern recognition solutions, one cannot obtain detailed, semantic description for a given multimedia document. Several limitations exist in feature extraction, classification schemes, along with the incompatibility of representations across domains. The situation will most likely remain so, for several years to come.

Towards addressing this challenge, we observe that several multimedia collections contain similar parallel information that are: i) semantic in nature, ii) weakly aligned with the multimedia and iii) available freely. For example, the content of a news broadcast is also available in the form of newspaper articles. If a correspondence could be obtained between the videos and such parallel information, one could access one medium using the other, which opens up immense possibilities for information extraction and retrieval. However, it is challenging to find the mapping between the two sources of data due to the unknown semantic hierarchy within each medium and the difficulty to match information across the different modalities. In this thesis, we propose novel algorithms that address these challenges.

Different $< Multimedia, Parallel\ Information >$ pairs, require different alignment techniques, depending on the granularity at which entities could be matched across them. We choose four pairs of multimedia, along with parallel information obtained in the *text* domain, such that the data is both challenging and available on a large scale. Specifically, our multimedia consists of movies, broadcast sports videos and document images, with the parallel text coming from scripts, commentaries and language resources. As we proceed from one pair to the next, we discover an increasing complexity of the problem, due to a relaxation of the temporal binding between the parallel information and the multimedia. By addressing this

challenge, we build solutions that perform increasingly fine-grained alignment between multimedia and text data.

The framework that we propose begins with an assumption that we could segment the multimedia and the text into meaningful entities that could correspond to each other. The problem then, is to identify *features* and learn to match a text-entity to a multimedia-segment (and vice versa). Such a matching scheme could be refined using additional constraints, such as temporal ordering and occurrence statistics. We build algorithms that could align across i) movies and scripts, where sentences from the script are aligned to their respective *video-shots* and ii) document images with lexicon, where the words of the dictionary are mapped to clusters of word-images extracted from the scanned books.

Further, we relax the constraint in the above assumption, such that the segmentation of the multimedia is not available *apriori*. The problem now, is to perform a joint inference of segmentation and annotation. We address this problem by building an over-complete representation of the multimedia. A large number of putative segmentations are matched against the information extracted from the parallel text, with the joint inference achieved through dynamic programming. This approach was successfully demonstrated on i) Cricket videos, which were segmented and annotated with information from online commentaries and ii) word-images, where subwords called Character N-Grams, are accurately segmented and labeled using the text-equivalent of the word.

As a consequence of the approaches proposed in this thesis, we were able to demonstrate text-based retrieval systems over large multimedia collections. The semantic level at which we can retrieve information was made possible by the annotation with parallel text information. Our work also results in a large set of labeled multimedia, which could be used by sophisticated machine learning algorithms for learning new concepts.