Aligning Textual and Visual Data Towards Scalable Multimedia Retrieval

The search and retrieval of images and videos from large repositories of multimedia, is acknowledged as a hard challenge. With existing solutions, one cannot obtain detailed, semantic description for a given multimedia document. Towards addressing this challenge, we observe that several multimedia collections contain similar parallel information. 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. 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. 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, and ii. document images with lexicon. 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. 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 with commentaries, and ii. word images 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.