Abstract

Analyzing and understanding gestures plays a key role in our comprehension of communication. Investigating the co-occurrence of gestures and speech is currently a labor-intensive task in linguistics. Although, with advances in natural language processing methods, there having been various contributions in this field, computer vision tools and methods are not prominently used to aid the researchers in analyzing hand and body gestures.

In this thesis, we present different contributions tailored to tackle the challenges in real-world gesture retrieval which is an under-explored field in computer vision. The methods aim to systematically answer the questions of 'when' a gesture was performed and 'who' preformed it in a video. Along the way, we develop different components to address various challenges in these videos, such as the presence of multiple persons in the scene, heavily occluded hand gestures and abrupt gesture cuts due to the change of camera angle.

In contrast to the majority of the existing methods developed for gesture recognition, our proposed methods do not rely on the depth modality or sensors signals, which is available in some datasets to aid the identification of gestures. Our vision-based methods are built upon the best practices in learning the representations of complicated actions using Deep Neural Networks. We have conducted a comprehensive analysis to choose the architectures and configurations to extract discriminative spatio-temporal features. These features enable the retrieval pipeline to find the 'similar' hand gestures. We have additionally explored the notion of similarity in the context of hand gestures through field studies and experiments.

Finally, we conduct exhaustive experiments on different benchmarks and to the best of the author's knowledge, run the largest gesture retrieval evaluations using the real-world news footage, the Newscape dataset, which is a collection of more than 400 000 videos with numerous challenging scenes for a retrieval method. The assessed results by experts from the linguistics domain suggest high potential of our proposed method in inter-disciplinary research and studies.