Preface

In recent years, action recognition has received immense attention in the field of computer vision and multimedia research community due to the vast area of real-time applications such as intelligent human-robot interaction, health care, video summarization, video captioning, and surveillance monitoring. There are multiple basic semantic concepts such as human pose, facial expression, surrounding objects, background scene, and their interaction with objects which help to understand what type of action has been performed by an individual in a video. Recently deep learning has also witnessed a sequence of tremendous success in the discipline of computer vision. However, most convolutional neural networks and recurrent neural networks identify action without accomplishing a detailed and fine-grained semantic interpretation of the entities or scenes in the videos. These problems may convey the dilemmas in the interpretation of human actions which hinders the performance of recognizing the actions in videos.

In this thesis, we address the issues of related to deep learning based action recognition. In Chapter 3, we propose a deep neural network to address the issues of factorization of ambiguous action into activity performed by an actor with co-occurring objects, and underlying scenes to mitigate the influence of representation biases when they are irrelevant to the action in consideration. Our network concentrates on formulating attention mechanisms that isolate an actor from associated objects and scenes followed by capturing the long-range temporal context of a video. Chapter 4 focuses on attentive associated features of facial and gestures in the long-term context of a video for correct recognition of ambiguous actions. We propose a visual attention network to extract discriminative features of facial expressions and gestures in spatial and temporal dimensions. In Chapter 5, we concentrate on action recognition in the wild which is divided into two sub-problems: a) fall and b) unusual action recognition in videos. Chapter 5a focuses on proposing the deep neural network which incorporates semantic supervision using the per-class weight of uncertain action through class-wise weighted focal loss. We also address the class imbalance and the weak inter-class separability issues to accurately recognize fall action class from a given video. In Chapter 5b, we propose a federated contrastive learning-based unusual action recognition framework that learns the features of unusual actions to probe the behaviour of a patient dealing with mild cognitive impairment. We then resolve the issue of multi-label action recognition in Chapter 6. It resolves the issue of permutation invariant of multiple action that leads to composite action through a novel permutation invariant transformer network.

We conduct a comprehensive set of ablation studies and experiments to show the efficiency of our models. We consider accuracy as standard metrics for action recognition. We use publicly available benchmark datasets, such as UCF101, HMDB51, Breakfast, Multi-Thumos, Kinetics-400/600, ActivityNet, and OOPs to demonstrate the efficacy of proposed models. Our models outperforms the state-of-the art approaches in terms of accuracy for action recognition in wild. As per our knowledge for aforementioned issues, there is not accurate dataset for comparative study, thus create a new dataset, such as VALC, FallAction, UnusualAction, and CompositeAction dataset, which has both short and long duration videos.