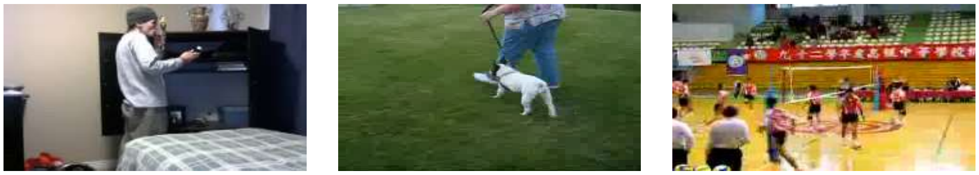
Chapter 1

Introduction

Due to the instantaneous advancement of technology such as resource-constrained devices and social media, which improve the active research field in multimedia and computer vision. The devices such as smartphones, mobile, Ipad, etc., help in recording, storing, viewing, and uploading a tremendous amount of videos around the world on social media platforms such as (Facebook, Instagram, and WhatsApp status), etc. These recorded videos contain lots of semantic information that helps in understanding the overall scenario of activities in a scene. A video is an important tool for scene understanding because it provides a temporal context for understanding the environment. It can capture changes in the environment over time, which can be used to identify objects and activities, track objects and people, and detect anomalies. In addition, a video can also help the viewer understand how the scene is being portrayed by recognizing the action. These actions are comprised of gestures of a person, the interaction with the object, or group activities, as shown in Figure 1.1. Therefore, recognizing action plays an important role in the analysis of video understanding. We have briefly summarized video-based action recognition in terms of definition, applications, challenges, issues, and our contribution in further subsections.

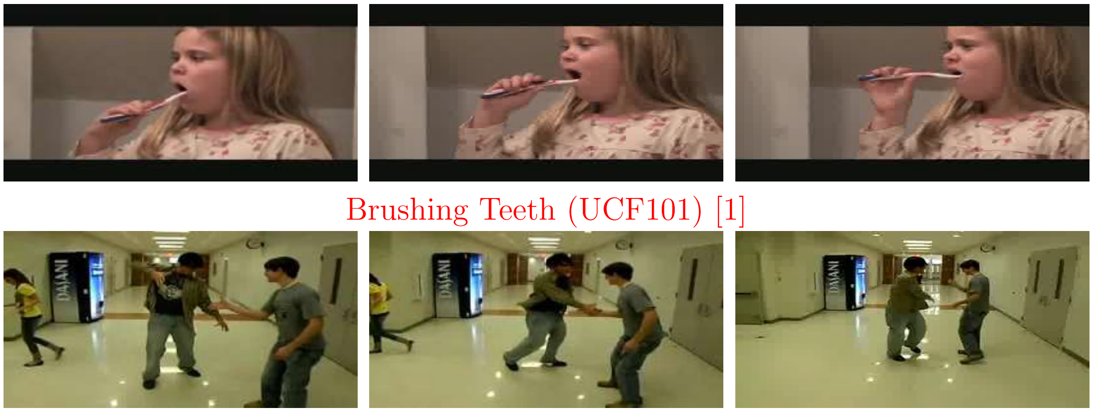


Holding WalkingWithDog VolleyballSpiking

Figure 1.1: RGB frame taken from publicly available datasets UCF101 [1] and HMDB51 [2], respectively. It depicts an action label for a sequence of movements comprising gestures, interaction, and group activities.

1.1 Definition

Action recognition (AR) is a type of computer vision technology that is used to identify and classify human actions in videos. In other words, the goal of action recognition is to identify activities according to object and scene states in a pre-segmented sequence of RGB frames, as shown in Figure 1.2. There are multiple basic semantic concepts such as human pose, surrounding objects, background scene, and interaction with objects which help to understand what type of action has been performed by an individual in a video.



Brushing Teeth (UCF101) [1]

Hand shake (HMDB51) [2]

Figure 1.2: Sequences of RGB frames from UCF101 and HMDB51 datasets that depicts a particular action.

1.2 Applications

It is used in a variety of applications, such as video surveillance [3], human-computer interaction [4], automatic video captioning [5], video summarization [6], and Fall recognition and Elderly Monitoring [7], which are briefly described as follows:

- **Video surveillance:** The increasing need to ensure safety and privacy through surveillance has led to the rapid deployment of visual sensing devices in both public and private spaces such as offices, traffic intersections, housing apartments, shopping malls, and airports. The deployed closed-circuit television (CCTV) cameras continuously generate a large amount of visual data. This has given rise to the need for efficient video analytics solutions for the large volume of recorded or live surveillance videos.
- **Human-computer Interaction:** Human-computer interaction (HCI) is the study of how people interact with computers and other digital devices. It involves understanding user needs and designing technology that is usable, efficient, and satisfying for people to use. HCI includes research on topics such as user interface design, usability testing, human factors, and cognitive psychology, and is important in the development of software, websites, and other digital products.
- **Automatic video captioning:** Video captioning is another field of research that holds a lot of significance in real-world applications. The task is to automatically generate an English description or caption for an image or video clip. It involves analyzing its visual content and producing a description in natural language that captures the essence of the image or video. This could be very helpful in commerce, the military, and education. For instance, visually impaired people can better understand the content of images if the description is available. Advances in deep learning have led to improved performance in image captioning as well.
- **Video Summarization:** Video summarization is the process of creating a shorter version of a longer video by selecting and presenting the most important or rep-

resentative frames or segments from the original video. The goal is to condense the content of the video while preserving its key information, enabling more efficient and effective video browsing and retrieval. Video summarization techniques can be categorized as extractive, in which representative frames or segments are selected from the original video, or abstractive, in which a new summary video is created by generating new frames or segments that convey the most important information from the original video.

- **Fall Recognition:** Fall recognition is the process of detecting and identifying when a person has fallen or is at risk of falling. It often involves the use of technology such as motion sensors, cameras, and wearable devices that can detect sudden changes in movement or posture. Fall recognition systems can be used to alert caregivers or emergency services in the event of a fall, or to provide early warning signs to help prevent falls from occurring in the first place. These systems are particularly important for elderly or disabled individuals who may be more prone to falls and related injuries.
- **Elderly Monitoring:** Elderly monitoring typically refers to the use of technology or other methods to keep track of the well-being and safety of elderly individuals, especially those who live alone or require assistance with daily living activities. This can include monitoring vital signs, detecting falls, and tracking medication adherence, among other things. The aim is to provide early intervention or assistance to help seniors maintain their independence and quality of life.

1.3 Challenges

We highlight the challenges that exist in the state-of-the-art methods for application as follows: Video-based action recognition involves analyzing and classifying human actions from video data, and it is a challenging task due to several factors, including:

- **Variability in appearance:** Different people can perform the same action in different ways, making it difficult to identify a specific action based on appearance alone.
- **Background clutter:** The presence of irrelevant objects or people in the background of the video can make it difficult to accurately identify and track the subject performing the action.
- **Occlusion:** The subject's body parts may be partially or completely obscured, making it difficult to accurately track their movements.
- **Temporal dynamics:** Actions can be performed at different speeds or in different orders, making it challenging to accurately capture the temporal dynamics of the action.
- **Dataset bias:** The performance of video-based action recognition systems can be influenced by the quality and quantity of the data used to train the system. If the dataset used to train the system is biased or limited, it may not generalize well to new or diverse situations.

Addressing these challenges often requires the use of sophisticated computer vision and machine learning techniques, such as deep learning, that can extract relevant features from the video data and model the complex temporal dynamics of human actions.

1.4 Problems

We have addressed the following problems which are listed below:

- Action Recognition in Presence of Representation Bias
- Human Behaviour Traits aware Action Recognition
- Action Recognition in Wild (Fall Action and Unusual Action)
- Multi-label Action Recognition

1.5 Contributions of the Thesis

In this thesis, we investigate the deep network architecture models for action recognition in an unconstrained environment. We develop a robust deep neural network that can recognize action classes with high accuracy. We focus on the practical challenges that occur while capturing action recognition in real-life scenarios. The main contributions addressed by the thesis are summarized as follows:

- SOLUTION 1: We humans when we have to identify an action in a video, no matter how complicated the representation bias is, the recognition is subserved by both a local process aware of human-object interaction features and a global process of retrieving structural context.

  We follow a similar approach to solve the issue of representation bias. We incorporate an actor-scene factorization followed by an attention mechanism that learns both representation biases from the semantically enriched feature maps to successfully overcome the issue of misclassification.

- SOLUTION 2: In the case of an ambiguous human action class, we humans try to understand the context related to facial expressions and gestures. We pay attention to transformation invariant features to overcome the effect of ambiguous action classes in the video.

  We model our network based on the aforementioned process by capturing facial expressions and gestures followed by enhancing the transformation modeling capability of a network to classify action classes in a long-term scenario.

- SOLUTION 3: Humans can recognize falls action in video by relying on a combination of visual cues and contextual information. Some of the visual cues that people use to recognize fall actions include:

  - Sudden changes in body orientation and position.
  - Rapid movements and gestures indicating a loss of balance or control
  - The speed and trajectory of the fall

  - The impact of the fall on the person or surrounding objects
  - The presence or absence of a reaction to the fall, such as attempts to break the fall or get back up.
  - In addition to these visual cues, humans also use contextual information, such as the location and activity of the person before the fall, to help identify and understand what is happening in the video.

  Automated fall recognition systems attempt to replicate this process using computer vision and machine learning techniques. These systems may use a variety of visual features and motion analysis techniques to detect falls in the video, and may also incorporate contextual information, such as time of day or activity level, to improve the accuracy of their predictions.

- SOLUTION 4: Humans recognize multi-action in videos by relying on a combination of visual cues and contextual information. Automated multi-action recognition systems attempt to replicate this process using computer vision and machine learning techniques. These systems may use a variety of visual features, object detection and tracking techniques, and temporal modeling to detect and classify multiple actions in a video, and may also incorporate contextual information, such as the time of day or activity level, to improve the accuracy of their predictions.

1.6 Organization of Thesis

To organize the thesis, we have logically divided it into 7 chapters, including the current introductory chapters. We have briefly illustrated the successive chapters as follows:

- **Chapter 2: Related Work**

  In this chapter, we provide a brief summary of the literature related to action recognition techniques. The main aim of this chapter is to provide a piece of up-to-date and relevant state-of-the-art information on approaches for resolving the task of human action recognition. We have discussed and analyzed the existing leading state-of-the-art deep neural networks. These approaches are broadly classified as 1) Approaches based on single-label actions and 2) Approaches based on multi-label actions. Approaches based on single-label actions fall again into two sub-categories: (1) Habitual actions and (2) Unusual actions. Similarly, the multi-label actions fall into only Habitual action. Moreover, we also examine the essential topics where current approaches succeed and fail in identifying actions in videos.

- **Chapter 3: Action Recognition in Presence of Representation Bias**

  In this chapter, we propose a deep neural network architecture, denoted by FactorNet, for efficient action recognition in videos with long temporal duration. We design an attention mechanism that separates an actor from the associated objects and co-occurring scene followed by capturing long-range temporal context.

- **Chapter 4: Human Behaviour Traits aware Action Recognition**

  In this chapter, we propose an attention-aware deep neural network named human action attention network (HAANet) that can capture long-term temporal context to recognize actions in videos. The visual attention network extracts discriminative features of facial expressions and gestures in the spatial and temporal dimensions. We have further consolidated a class-specific attention pooling mechanism to capture transition in semantic traits over time.

- **Chapter 5a: Action Recognition in Wild: Fall Action**

  In this chapter, we propose a video-based action recognition with fall detection architecture, FallNet, which learns the features of uncertain actions related to day-to-day activities. FallNet first incorporates semantic supervision using the per-class weight of uncertain action through class-wise weighted focal loss. It addresses both the class imbalance problem and the weak inter-class separability issue. We design a joint training model to train the overall architecture efficiently in an end-to-end manner.

- **Chapter 5b: Action Recognition in Wild: Unusual Action**

  We propose a federated contrastive learning-based unusual action recognition framework, named as UnusualFedNet, that learns the features of unusual actions to probe the behaviour of a patient. Each edge device uses the proposed light-weight UBNet architecture to capture fine-grained human-object relations.

- **Chapter 6: Multi-label Complex Action Recognition**

  We propose *TIKNet*, which captures composite human actions in a long-range video. A novel Kronecker-based transformer network is proposed, which offers an effective method of building a highly extensive network while maintaining low number of parameters. Extending the scope of the proposed work, we have also addressed the issue of a different permutation of one-actions. We have discussed theoretical proof of invariance to permutation of one-actions and temporal duration of composite action. Moreover, we have discussed the efficacy of the proposed work with extensive experiments on three benchmark datasets. To the best of our knowledge, no dataset exists in the literature for ambiguous composite action pairs. We have also proposed a new dataset, named *CompositeNet*. The proposed dataset contains 31 frequently occurring composite actions with an average of 81 videos per class and an average length of 3 minutes and 30 seconds.