Ph.D. Research Proposal

Visual Assessment of Long-term Changes of Activity levels in Elderly People

Author
Muhammad Ahmed Raza

Supervisor
Prof. Robert Bob Fisher

Institute of Perception, Action and Behaviour
School of Informatics
University of Edinburgh

Abstract

The advancements in the field of Artificial intelligence has revolutionized almost every field of life including healthcare systems. Due to more life expectancy, number of elderly people are increasing who require constant help in their daily chores and full-time attention. Research is being carried out to make the life of elders better using technology.

In this research, we propose to use only computer vision based surveillance algorithms to monitor the motor movement deterioration of elders over time. We propose to target two areas of motion assessment, i.e., action-based and joints motion analysis-based algorithms. Moreover, we aim to collect a dataset in a real household environment, where a person sits and eats his/her meals of the day. The performance of the frameworks will be judged in terms of classification and localization accuracy for action-based frameworks. Whereas, for joints motion analysis-based frameworks, the performance will be evaluated by the accuracy of classification and pose estimation.

Keywords: *Elderly healthcare, Action recognition and localization, motor deterioration*

Contents

Chapter 1

Introduction

1.1 Motivation

On a global scale, "The proportion of people aged 60 or over was just 8% in 1950 but this is projected to rise to 20% by 2050", said World Economic Forum in "Global Population Ageing: Peril or Promise?" in January 2012. the number of people growing older is increasing day by day. On the other hand, nurses and doctors are less. With this growing need, health care systems are under pressure, and the need of the hour is to automate the health care systems.

People in old age face several issues which include pathological diseases such as osteoporosis in which the bones of the person grow porous and weaker i.e. their gait is affected and vasovagal (a sudden drop in heart rate and blood pressure) which cause the person to faint and fall. In fact, these aged individuals even develop one or more long-term conditions, such as stroke, arthritis, heart disease or dementia.

Apart from these diseases, there are some other problems they face in the daily routine if timely addressed, can be very helpful. Such issues identified in [1] are: First, drastic change in pulse, which might pose heart attacks and many others that render them unconscious, resulting in falling. Second, they suffer from neurological or musculoskeletal problems which results in their distorted posture, change in gait or lack of limb movement over time which might be a potential symptom of a dangerous disease and will affect them afterwards if timely action is not taken.

1.2 Problem Definition

Motion analysis systems usually involve wearable sensors to detect problems such as if a person fell or if a person is suffering from some neurological disease such as Parkinson's. These wearable sensor-based devices have great potential to give accurate results. On the downside, elderly tend to forget things and there may also be some reluctance to use technological support, e.g. feeling that they do not want or need these devices. So, this is not a feasible solution. A alternative solution would be to develop a camera-based surveillance system which could monitor a person's daily life routine and identify if there is an anomaly or gradual change in behaviour.

This research primarily focuses on developing a camera-based system for the surveillance

of daily eating activities performed by the elderly and specifically targets the following areas,

- Estimate a view-invariant 3D upper body pose (skeleton) of a person in a real-world constrained environment
- Track the movement of joints and predict:
  - Activity-based motion quality assessment
  - Non-activity-based analysis
  - Abnormality, i.e., motor deterioration over longer periods of time

1.3 Research Questions

The proposed research targets and addresses the following four research questions:

- How can we identify motor deterioration in the elders by tracking their upper body and by using only non-intrusive and non-invasive techniques?
- Can we use RGB-D cameras to identify minute details such as small jitter or shakiness? And how well scale and view-invariant 3D pose can be estimated from an RGB-D camera?
- What are the performance metrics, their reliability and distinctive features and how to assess these reliably, given the variety of motions, day-to-day variations, and noise?
- What are the variety of eating motions and actions performed over various periods of time i.e., short and long time periods?

1.4 Claims

For our proposed algorithms,

- We claim that our one-of-a-kind unified eating action recognition framework will be an end-to-end deep learning framework that will not only localize and recognize an action but will also estimate its normalcy.
- We claim that fusion of spatio-temporal context information and attention on joint feature maps will help in understanding the scene and sub-action in a better way. Moreover, exploiting anchor based techniques will make the action localization faster and accurate.
- We claim focusing on eating activity will help us in identifying the deterioration, i.e., ease of action.
- We claim that exploring motion estimation irrespective of the underlying activity will give us the window of opportunity to analyse the person's movements on a larger scale, for example, it will give us the freedom for quantification of any action not just only eating.

1.5 Terminologies

The terminologies discussed about the research questions and their proposed solutions identified in the literature review chapter are defined in this section.

- **Trimmed Videos** contain only one of the specified action.
- **Untrimmed Videos** are that also contain multiple activities or no activity at all in some frames are referred to as untrimmed videos.
- **Fine-grained Activity Recognition** is a type that recognizes the action on a micro level. For example, 'a person is eating' is a common activity but a fine grained eating activity would be 'a person is eating a burger'.
- **Composite Action** is an action that comprises of many smaller actions.
- **Sub-Action** is a type of smaller action that occurs in composite actions. For example, 'eating' is a composite action and 'pick up a tool', 'scoop with it' and 'put in mouth' are its sub-actions.
- **Composite / Multi-Instance Multi-Level Action Recognition** is a type that recognizes a chain of multiple actions in a single untrimmed video stream.

Chapter 2

Literature Review

There are numerous problems that elders face in their daily life which hinder their ability to live independently. These include, neurological weakness, motor movement deterioration and restricted movement. Many diagnosis and prognosis techniques have been developed to tackle this problem but have limited application in reality. For example, there are wearable sensors and vision techniques that can detect the fall of a person [2]. A wearable sensor for the elder is not always a good solution because they tend to forget things or resent the intrusion. However, there are sensors embedded in mobile devices such as an inertial measurement unit (IMU) and other lightweight sensors such as the optical linear encoder (OLE) which can be effectively attached to the clothing. However, for long-term surveillance of elders movements, the above mentioned sensors suffer from problems such as drift over time which renders them infeasible.

In contrast to wearable devices, vision-based techniques solve these issues by tracking the movement of the limbs using marker-based approaches. However, these approaches are only practical or robust in constrained environments [3]. On the other hand, movement analysis using marker-less approaches might be an efficient solution to these issues. However, this is not a trivial task because extracting silhouettes largely depends on effective background subtraction [4] techniques (a priori information, i.e. model of the background is required).

Moreover, silhouette extraction loses critical information for such as whether the person was facing towards or away from the camera. Although precisely calibrated algorithms based on silhouettes using subject-specific body models have shown better results to date, their use has been limited to only laboratories due to cost constraints. The trend, however, has been to move away from the use of image silhouettes to improve robustness and reduce ambiguities. Moreover, other practical implementations for marker-less frameworks involve determining the shape/pose of the human body (skeleton) or a specific limb effectively.

In general, motion analysis can be classified into two categories: clinical and non-clinical. As the name suggests, clinical motion analysis is in a controlled environment and does not replicate a near real-world environment. The non-clinical analysis emphasizes the fact, the need to socialize in real life and and do unpredictable daily tasks. In general, camera-based motion analysis can be described as a three-step process. Pose estimation, pose tracking and motion estimation. Moreover, a brief literature review on each of the topic is given in the sections below.

2.1 Human Pose Estimation

Human pose estimation (HPE) identifies human body representations and parts from a stream of images or videos. Human body modelling is an important step to represent features and key points in a structured format. Hand-crafted techniques were used in the past until recently, with the advent of deep learning frameworks, the performance of HPE algorithms improved significantly. Hence, deep learning proved to be a breakthrough for HPE. The literature review in this section on HPE conforms to the taxonomy presented by Zheng et al. [5]. Typically there are three types of body modelling techniques, i.e., kinematic (2D / 3D HPE), planar (2D HPE) and volumetric (3D HPE usually formed from a mesh). The figure 2.1 shows three modelling types.

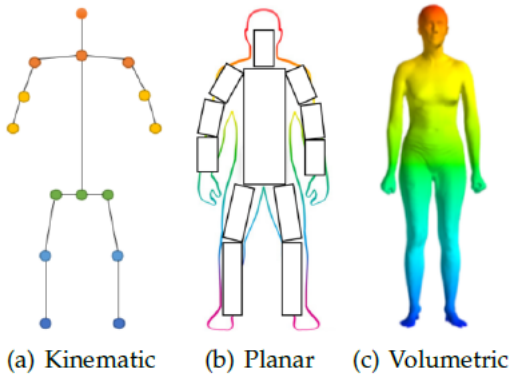

Figure 2.1: Image Credit: This image is taken from the paper of Zheng et al. [5]

2.1.1 2D Pose Estimation

2D pose estimation methods estimate the pose of single or multiple people from the discriminative features extracted from a video or image. This is the 2D pose or the spatial location of the human joint locations. 2D pose estimation can further be classified into two sub-domains, i.e., single person pose estimation and multi-person pose estimation.

Single Person Pose

2D single person pose estimation is used to extract the pose of a person when the image only contains a single person. If the image has more than one person, each of them is cropped by using full body and upper body detectors [6]. In general, there are two categories for single person pose estimation based on deep learning techniques. Regression-based and body part detection. A few major contributions in each of the areas are discussed below.

Regression-based frameworks learn a mapping from an input image to human body models or spatial joint locations. Many frameworks used regression-based learning technique to localize human joints. Toshev and Szegedy proposed a Deep Convolutional Neural Network (CNN), DeepPose [7] which cascaded deep neural network regressors to learn key points from images. Carreira et al. [8] proposed Iterative Error Feedback (IEF) which used GoogLeNet [9] as its backbone and was a model that progressively updated the initial solution by injecting the prediction error back to the input space. Li et al. [10] and Fan et al. [11] proposed multi-task

heterogeneous frameworks that comprised of two tasks at hand, i.e., finding joint locations and body part locations in an image patch. However, Luvizon et al. [12] also proposed a multi-task network that learned pose estimation and activity recognition simultaneously.

Body part detection methods predict the approximate locations of joints using heatmaps based supervised representations. The ground truth heatmap is generated by a 2D Gaussian with mean as the joint's spatial location in the image. Compared to joint coordinates, heatmaps provide better supervision as it preserves the spatial locations of the joints. Tompson et al. [13] combined a CNN based body part detector and part-based spatial-model into a unified framework. Lifshitz et al. [14] also proposed a CNN based algorithm that incorporated keypoint votes and joint probabilities for HPE. Wei et al. [15] proposed Convolutional Pose Machines (CPM), a sequential network with multi-stage processing. Newell et al. [16] proposed a stacked-hourglass (SHG) network (encoder-decoder) to capture multi-scale features. Chu et al. and Yang et al. [17], [18] respectively, built upon SHGs and improved their performance. They introduced the novel Hourglass Residual Units (HRUs) and Pyramid Residual Module (PRM) respectively to enhance the scale invariance of deep CNNs. Recently, Tang and Wu [19] showed that all the body parts are not related to each other so they proposed a part-based network that assigned different weights to each part group.

For video signals, researchers focused on Spatio-temporal modelling of the information. Jain et al. [20] designed a framework that utilized both colour and motion features. Pfister et al. [21] proposed a CNN that was able to utilize motion features to align predicted heatmaps from neighbouring frames. Luo et al. [22] utilized Long Short-Term Memory (LSTM) to capture temporal geometric consistencies from various frames.

2.1.2 3D Pose Estimation

3D HPE targets to predict locations of joints in 3D space instead of a 2D plane, like in 2D HPE. 3D pose estimation has also been of great interest as this could potentially provide structural information of the human body. 3D HPE can be classified into two categories, i.e., Pose from RGB Camera and Pose from a Depth camera.

Pose from RGB Camera

Recent progress on 2D HPE has enabled the researchers to extend the frameworks from 2D to 3D HPE. 3D pose from a monocular view is a severely ill-posed problem because many 3D poses can be projected onto the same 2D pose. This gets worse in the case of multi-person detection. Pose from an RGB camera can be divided into two categories, i.e., model-free methods and model-based methods. Model-based methods use human body models to reconstruct 3D representation., whereas model-free methods use direct estimation or 2D to 3D transformation approaches.

Direct estimation approaches infer 3D pose without explicitly estimating a 2D pose. Some recent works that particularly use this approach are [23], [24], [25] and [26]. 2D to 3D lifting approaches estimate a 2D pose as an intermediate step and is the most used technique. Tekin et al. [27] and Zhou et al. [28] used heatmaps instead of the spatial location of keypoints to project into 3D space. Some graph-based techniques for lifting 2D to 3D have also shown promising results, i.e., [29]. [30] and [31].

Model-based methods utilize a parametric (kinematic or volumetric) body model to get

a 3D pose. These methods require a model such as the kinematic model as a prior. Some researches that use kinematic model-based methods are [32], [33] and [34]. Volumetric models on the other hand provide richer information of the human body by recovering high-quality human mesh. Some researches that utilize volumetric models are [35], [36], [37] and [38].

Partial occlusion is a big challenge for single-view 3D HPE. The solution to solve this problem is to estimate a 3D view of a person from multiple poses. However, this formation causes other challenges such as resolving the corresponding location between different cameras and gets computationally expensive and memory-intensive in the case of multiple people per view. Chen et al. [39], Dong et al. [40], and others proposed multi-view matching frameworks for 3D HPE across all viewpoints. Zhang et al. [41] and Pavlakas et al. [42] accumulated the generated 3D heatmaps into a 3D structure, based on calibrated cameras. Nie at al. [43] proposed a sequential bidirectional recursive network (SeBiReNet) capable of reconstructing unseen and occluded poses. They also propose an adversarial augmentation strategy that enables us to realize view transfer on 3D poses.

Pose from Depth Camera

Most of the 3D HPE algorithms use a multi-view RGB setup. Other sensors such as Integrated Measurement Units (IMU), depth sensors and radiofrequency devices are also used for HPE. Moreover, depth sensors have gained attention due to their moderate cost and efficiency of 3D vision tasks. Depth sensors alleviate the depth ambiguity issue, which is a key problem in 3D HPE. So, if depth sensors are present along with RGB cameras HPE becomes a well-posed problem. This can be further divided into two subclasses, i.e., depth image-based and point cloud-based.

Depth image-based pose estimation has been done by using only depth data and also in combination with RGB data. Yu et al. [44] proposed DoubleFusion, which only used single view depth images to get 3D human body pose. Xiong et al. [45] proposed a network, Anchor-to-Joint regression network (A2J) which estimated anchor points from global-local spatial context information. In [46] used RGB-D cameras to capture data from multiple views and used random-forest based before incorporate information of the environment. In the end, multi-view fusion along with RGB-D optimization was used to estimate the pose. Recently, Zhi et al. [47] proposed to use high-resolution albedo texture from RGB-D video.

Point clouds can provide more information than depth images. PointNet [48] and its upgraded variants have validated good performance for various applications on segmentation and classification of the point cloud. Jiang et al. [49] combined PointNet++ [48] with SMPL to get 3D human pose. Wang et al. [50] used PointNet++ along with an attention method named spatial-temporal mesh convolution for the estimation of 3D human meshes.

2.1.3 Section Summary

This section described a brief overview of the various frameworks of the human pose estimation. It can be divided into two classes i.e., 2D and 3D pose estimation. 2D poses are further divided into regression and body-part based methods. Regression methods can learn a non-linear representation in an end-to-end fashion which is generally a fast learning paradigm. However, this gives suboptimal solutions as the problem is highly non-linear. Body-part based methods are more widely used since it uses probabilistic prediction of each pixel and heat maps which

provide rich highly preserved spatial information. However, it is highly dependent on the heat maps and its resolution. If high resolution heat maps are used, the computational complexity and the memory footprint increases significantly.

Most of the 3D pose estimation frameworks require the estimation of 2D poses prior to transforming them into 3D space. So, indirectly the performance of the 2D network affects the performance of 3D pose estimation networks. Frameworks that do not explicitly utilize 2D poses show promising results but their performance degrades when applied to unconstrained environments and occluded scenes. Anyway, to utilize and compare our results with the current state-of-the-art framework, we will focus on collecting the data in a constrained environment and find a suitable framework for our application.

As the problem defined in chapter 1, 3D pose estimated from a depth camera seems to be a better choice as it provides a better kinematic representation of the structure of a person. For this purpose, the Anchor-to-Joint (A2J) [45] framework can be used. It exploits the use of a single depth camera thus preserving the identity of the person and efficiently estimating the 3D kinematic model of the hand and human full body pose. However, A2J is sensitive to the predicted region of interest and the network is of high complexity, thus is harder to train. Moreover, it utilizes AsusXtionPRO, light depth sensor and does not provide good results with Intel Real Sense Camera, which is what we have.

2.2 Tracking

Visual Object tracking deals with tracking objects of interest in a video stream. This area has gained a lot of attention as this is a significantly important step in various applications. Due to its importance, many hand-crafted and deep learning techniques have been proposed to tackle this challenge. Recent tracking algorithms exploit the structure of the objects of interest to predict their target locations. This can broadly be divided into two categories, i.e., Correlation Filter Trackers (CFT) and Non-Correlation Filter Trackers (NCFT). The taxonomy of the tracking algorithms conforms to the one presented by Fiaz et al. [51].

2.2.1 Correlation Filter Trackers

Correlation Filters (CF) have been used in object tracking to improve efficiency and robustness. Conventionally, the purpose of using a CF is to get a map such that foreground/target (region of interest) and background have significantly different values. To keep the computational cost to a minimum, CF-based algorithms perform computations in the frequency domain. They are usually not adaptive in terms of change in scale, shape and orientation. So they often require a lot of training data. CFTs can be further sub-divided in Basic-CFTs, regularized CFTs, Fusion-based CFTs, Siamese-based, and part-based.

Basic-CFTs are trackers that use Kernel-based Correlation filters (KCF) as their baseline framework. These trackers use various feature extraction techniques from Histogram of Gradiants (HOG) to deep features such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). A line of works that follow KCF as their baseline tracker are [52], [53], [54],[55], [56], [57],[58] and [59]. Trackers such as Basic-CFTs do not perform well in the case of occlusion and most often they are not able to re-identify the object of interest. This can be solved with a larger search region but the tracker loses its discriminative power and so

there is not much difference in the foreground and background feature maps. Eventually, its performance degrades. Therefore, regularization is introduced to maintain the discriminative ability. These trackers that incorporate regularization are referred to as Regularized CFTs. This area has been explored in many researches such as, [60], [61], [62], [63] and [64].

Siamese networks join two input images and measure the similarity between the two. This is done by using shared convolutional or fully connected layers. Integration of CFTs with Siamese network is referred to as Siamese-based CFTs. Some researches that utilize Siamese-based CFTs are, [65], [66], [67] and [68]. CFT-based techniques learn the features of the target completely, i.e., as a whole. On the other hand, Part-based CFTs usually learn target features in parts. Part-based CFTs are utilized in numerous applications. Some part-based trackers developed are, [69], [70], [71] and [72]. Image fusion means fusing complementary information to improve the performance of the algorithm. This can be done as Feature-level fusion (FLF), Pixel-Level Fusion (PLF) and Decision-Level Fusion (DLF). Several types of research have proposed fusion techniques to enhance the tracking ability of the algorithms. A few of those are, [73], [74] and [75].

2.2.2 Non-Correlation Filter Trackers

Trackers that do not use correlation filters are referred to as Non-Correlation Filter Trackers (NCFTs). These are sub-divided into superpixel, Siamese-based trackers, multiple-instance-learning, sparsity, graph, and patch learning. These trackers are discriminative, except for sparsity-based trackers which are based on generative modelling.

Patch learning-based trackers exploit background and foreground patches simultaneously. Numerous samples are used to test a trained tracker and the maximum response gives the information of the position of the foreground, i.e., target. Some researches that use patch learning are, [76], [77], [78], [79] and [80]. In Multiple-Instance-Learning-based (MIL) trackers, unlike other trackers that use individual patches, training samples are bundled together and labelled as a bundle. For example, if the bundle comprises all negative samples, it is given a negative label and if it contains at least one positive, it is given a label of positive. These instances form a weak classifier and a few selected instances collectively form a strong classifier. Multiple researchers have employed MIL-based tracking, which are, [81], [82], [83] and [84].

Siamese neural network-based NCFTs work on the matching mechanism. Its learning process explores the variations in the appearance of the region of interest. Some researches that have proposed and utilized Siamese NCFTs are, [85], [86], [87] and [88]. Superpixels is a group of pixels that have the same pixel values. The background and foreground are segmented into superpixels and thus classification is done to differentiate between the two. Various superpixel based trackers have been developed in the past, which are, [89], [90] and [91]. Moreover, graph-based algorithms also use superpixels as their nodes and geometric structure as their edges. Many trackers have been developed using graphs, which are, [92], [93], and [94].

Almost all the algorithms for tracking are based on discriminative models, i.e., differentiates between background and foreground. But generative models can also be used for the task. They learn the representation of the region of interest and then search for such representation with the least reconstruction error. Sparse representation is a good example and has been used for generative models. Usually, dictionary learning is used for sparse representation. In visual object tracking, it is used to differentiate between the region of interest and background by sparsely encoding the background and foreground coefficients. A few researches in the line of

sparse trackers are, [95], [96], [97] and [98].

2.2.3 Long-Term Tracking

For long term-tracking, to learn the appearance of the region of interest completely, patches that contain all the shape variations are required. In online methods, a limited number of positive samples may also lead to over-fitting of the problem. Goodfellow et al. proposed Generative Adversarial Networks (GAN) [99] which has the inherent capability to produce realistic images. Using a similar concept, Song et al. [100] proposed Visual Tracking via Adversarial Learning (VITAL) algorithm.

Another class of long-term tracking algorithms, a sub-branch of Basic-CFTs, is called Long-term Correlation Tracking (LCT) [101]. The LCT involves the use of CF to predict the translation and scale of the region of interest. In case it fails, LCT performs re-detection, which is performed by using a random fern classifier for online re-detection. LCT selects positive samples to predict new patches as the region of interest by k-nearest neighbour (KNN) classifier. LCT has been further improved by the same authors and is named Improved Long-term Correlation Tracking (ILCT) [102]. ILCT uses a Support Vector Machine (SVM) classifier for re-detection instead of a random fern classifier.

2.2.4 Section Summary

In this section, development in the area of tracking of the keypoints / regions / objects of interest was discussed. First the tracking algorithms were divided into two categories based on their techniques, i.e., Correlation filter-based or non-Correlation filter-based. Correlation filter based methods are more robust and efficient in terms of performance. In a complete motion estimation framework, body joints will be the keypoints to track in partially occluded video streams. In this case, regularized correlation filter trackers are a better choice as they ensure that the object of interest is re-detected even after a full or partial occlusion.

2.3 Human Activity Recognition

Human Activity Recognition (HAR) is widely used for various applications such as Human-Computer Interaction (HCI) systems and augmented reality. Moreover, the field of robotics, gaming and medical systems have gained benefit from the HAR development. HAR approaches can be classified into two domains [103], i.e., HAR approaches based on feature extraction techniques, HAR approaches based on recognition stages.

2.3.1 Approach-based HAR

HAR methods can further be sub-divided based on various feature extraction processes into two classes, i.e., hand-crafted features and Feature learning.

Feature Extraction

Feature extraction is a major step for any intelligent application. Handcrafted methods for feature extraction usually depend on prior knowledge to get better discriminating features. Methods based on hand-crafted techniques are Spatio-temporal and appearance-based approaches. Hence, spatial or temporal cues are the basis of action recognition [104].

A spatial representation can be utilized in multiple ways to extract features and to form a representation worthy of action detection. There are various ways to extract and exploit spatial features. First, a silhouette [105], [106], motion history images (MHI) [107], [108] or any other holistic representation can be computed directly from the image. Second, estimate a 3D body model for example a kinematic joint model to classify the type of activity [109], [110]. Third, this representation can also help get a set of statistical local features which in turn would help us in extracting local actions [111], [112].

A temporal representation can be in various forms and utilized for action recognition. For instance, an action can be represented by a sequence of dynamics and appearance. To recognize the action these representations are divided into groups based on similarity and are called states and transition between each state is learned [113]. This could potentially be with hidden Markov model (HMM) [114] and conditional random field (CRF) [115], [116], etc. Moreover, another way of representing an action with temporal features would be to represent it as a temporal block called template [117]. Also, utilizing statistical models can be utilized to represent the temporal statistics, i.e., describe the distribution of unstructured features [118].

Appearance-based approaches use depth information, i.e., 2D stereo or 3D depth images to get features such as shape, motion or a combination of both. This estimates the spatial location of joints accurate and gives real-time performance [119]. Shape features such as contour points and geometric features help in understanding the action in applications such as hand gesture recognition [120]. Motion representations such as optical flow have been used for the activity recognition task. These are used to extract motion features and afterwards, a classifier is employed for activity recognition [121], [122]. Hybrid models that use a combination of both motion and shape features have proven that it represents the actions in an efficient way [123] and [124]. Recently, Ahad et al. [125] utilizes kinematic posture features based on 3D linear joint positions and angles between them to recognize activity.

Mainly past research focuses on either exploring spatial and temporal representations, but until recently, this trend has been shifted more towards fusing both type of features. In [126] Tomei et al., propose a graph-based action recognition framework that explores high level interactions (both spatial and temporal) between human and objects to classify the action. Zong et al. [127], proposed a motion saliency stream to capture salient motion features, i.e., temporal context features. Further they fuse it with RGB and optical flow data to extract discriminative features via a CNN. Yan et al. [128] and Peng et al. [129] explore graph-based efficient pose estimation frameworks which in turn enhances the action detection. Moreover, [128] also introduce a method to temporally upscale smaller actions so that sufficient attention is given to each of the undergoing actions.

Feature Learning

Feature Learning is the second most important step in intelligent application development. Recently, feature learning has seen a boom with the advent of deep learning techniques. Feature

learning techniques for HAR can be grouped into traditional and deep learning-based methods.

Traditional approaches for feature learning include genetic programming (GP), dictionary learning (DL) and Bayesian Networks (BN). DL gives a sparse representation of the input data by linearly combining the basis of the dictionary. This is an unsupervised, end-to-end learning procedure. Zhu et al. [130] and Xu et al. [131] propose to use dictionary learning for the HAR solution. GP is an evolutionary method, which searches a space of solutions without prior knowledge. Liu et al. [132] proposed an adaptive technique using GP to get better Spatio-temporal representations for action recognition. BNs are probabilistic models. Li et al. [133] proposed to use a BN.

Deep learning methods for HAR is a widely explored topic. However, deep learning methods require a lot of data for their training. It is usually an end to end framework that exploits multiple feature representations and abstractions to recognize some set of activities. Deep learning-based techniques can be broadly classified into Generative and Discriminative Models. Generative models are unsupervised models that learn unlabeled data distribution, i.e., used in anomaly detection. Most commonly, auto-encoders [134] and Generative Adversarial Networks (GAN) [135] are efficient learning approaches for the task at hand. Discriminative models have supervised models that exploit hierarchical learning to categorize the data into output classes. Varol et al. [136] proposed the use of a discriminative model that showed the importance of accurate optical flow for activity recognition.

2.3.2 HAR on Untrimmed Videos

Most action classifiers discussed above utilize trimmed videos for action recognition, which is unrealistic in real-world videos. In untrimmed videos, action detection is more challenging as it involves frame-level real-time object detection and optical flow-based model to mark the start and end of an activity [137].

Chen et al. [138], proposed an approach that joined space-time localization and activity categorization. First, it learned an offline binary classifier using trimmed videos for the activity of its interest. Second, untrimmed videos are introduced which are then decomposed into various 3D space-time nodes. Zhang et al. [139] proposed a framework that firstly localized the action in the video by using the same concept of knowledge transfer from trimmed videos to untrimmed videos via two deep networks. They also introduced a new self-attention module and fused the output feature maps of the module in both networks to localize the action which in turn also enhanced the recognition capability.

Song and Kim [140], proposed a two-step process, namely DeepAct. In the first stage, they extracted rich features from two 3D deep networks, i.e., I3D and C3D, and in the second stage they used LSTM based Recurrent neural network model to predict and localize the activity. On the contrary, Gleason et al. [137] proposed a frame-wise feature extraction to find Spatio-temporal proposals using hierarchical clustering and jittering techniques, while the classification of activity is done using temporal refinement I3D. In 2020, Gleason et al. [141] utilized the same idea from their previous research and defined a set of parameters to enhance the computational efficiency of their proposed approach. They also proposed another framework, namely Chunk Aggregation [142] that first breaks the video into smaller chunks, then classifies the actions and in the end re-forms the over-lapping chunks of videos by aggregation.

Single action recognition in a video stream even on untrimmed videos is still a challenging task that is usually targeted with weak video-level labels. When the number of actions in

the untrimmed videos increases it becomes more challenging to get time stamps and classify the action. Moltisanti et al. [143] targeted this problem and proposed to use single timestamps supervision by first dividing each action bound (from start to end) in an untrimmed video and represent it as a sampling distribution from timestamps. Then the classifier is used to progressively update the sampling distributions and try to localize the actions performed in the videos.

Li et al. [144], proposed to use deep reinforcement learning, i.e., Q learning algorithm for proposal generation. They proposed to use a feature extractor which is based on CNN and long-short term memory (LSTM) blocks whose output feature maps are then fed into separate branches, i.e., a classifier to predict the category of activity and a Q-network to get the proposals. Rehman et al. [145] proposed an end-to-end trainable framework that learns distinctive Spatio-temporal feature representations using 3D CNN along-with classification and localization modules with different temporal scales for activity classification and localization.

On the other hand, activity recognition researches focus on predicting the current activity, but some researchers have explored the area of predicting future activities as well. For instance, Mahmud et al. [146] worked on a dual problem, i.e., finding which activity will be performed in future and its starting time in an untrimmed video sequence. The untrimmed video goes into a Siamese like architecture with three different branches to learn activity features and predict future activity and its starting time.

2.3.3 Fine-grained Composite HAR

Fine-grained activities are a set that is visually similar i.e., with low inter-class variability. On the other hand, Composite activities are the ones that when decomposed they tend to form multiple short activities. In most scenarios, it is important to get a detailed understanding of the activity, this involves the recognizing of the individual acts in a chain of activities, along with a high-level description of the ongoing composite activity. However, composite and fine-grained activities are not entirely mutually exclusive from each other. Composite activities can also be formed from multiple fine-grained activities [147].

Awwad et al. [148] used the depth sensors in activity recognition and proposed a depth descriptor using the local depth feature (LDPT) [149] followed by fisher vector encoding to capture a distinct representation of a combination of composite and fine-grained activity recognition. Piergiovanni et al. [150] compared various recognition approaches by focusing on its temporal structure. They found that exploiting the temporal structure helps in recognizing fine-grained activities in a better way.

Weak supervision is a branch where imprecise and noisy sources are used to label data in a supervised setting. Many researchers tend to use weakly supervised labels to recognize an activity. Heidarivincheh et al. [151] proposed an approach for detecting action completion and moment of action from weak labels. They proposed to use CNN along with recurrent cells to aggregate evidence for the completion of the activity. However, Zhang et al. [152] proposed PreTrimNet, a multi-instance and multi-label (MIML) action recognition technique. First, it performs spatiotemporal pre-trimming of untrimmed videos by detecting a person in the frame. Second, it learns the three representations of features, i.e., skeletons, optical flow and RGB feature maps in a self-attention representation module. Third, they use multiple 2D, 3D convolution layers and Fully-connected layers for MIML prediction. More recently, Zhang et al. also proposed another framework, named as IONet [153] which utilize both trimmed and

untrimmed videos to learn an action recognition model by iterative optimization. However, external trimmed videos are a secondary source for improving accuracy by sharing instructive domain knowledge, but it is optional and can be removed if trimmed videos are unavailable.

Some researches have focused on the task of composite activity detection for specific applications. For example, Behera et al. [154] targeted composite activities performed in a car driver environment. They proposed Multi-stream LSTM, which takes both low-level appearance feature maps from a backbone such as VGG16 and high-level contextual information such as estimated pose to recognize the underlying activity. Aakur et al. [155] explored the surveillance application in an outdoor environment and exploited the Spatio-temporal features for localization and recognition of activities. They proposed a three-stepped approach with frame-level cascaded region proposal and detection (CRPAD), followed by tracking of the object and action recognition.

2.3.4 Section Summary

This section dealt with the literature review of the human action recognition. Action recognition has a wide range of applications and can be categorized in an approach-based fashion. Usually action recognition approaches suffer from object occlusion and illumination changes. Moreover, a motion analysis system requires composite action recognition in untrimmed videos thus making the problem multi-instance and multi-label (MIML). Moreover, it is important to utilize a better model preferably a joint model that utilizes both the kinematic model of a human and depth data.

PreTrimNet [152] explores the above mentioned avenue of multiple representations of humans for their action recognition. It also shows state-of-the-art performance on multiple benchmark datasets. Despite of many strengths such as exploitation of joints and spatio-temporal features for action localization, PreTrimNet requires skeletons and optical flow data explicitly and thus is indirectly dependent on the performance of the preprocessing step. However, we need an end-to-end network that takes a video as an input, processes it and outputs the result without depending on any external help.

On the other hand, IONet [153], from the same authors, a more robust approach which utilized shared space embeddings and modelling of action classification in an iterative fashion. They also introduce a self-attention module to eliminate background frames and exploit temporal relevance. However, its training phase requires trimmed video sequences of each action which are not always available and demand higher memory requirements.

2.4 Motion Quantification / Assessment

Automatic capture and analysis of human motion have a lot of potential applications. The quantification or estimation of human motion accuracy was quite a rarely discussed area in the past. However, recently, there has been a lot of development in the field due to the increase in the elderly population, neurological diseases and rehabilitation cases. In broad terms, motion quantification can be classified into two branches, i.e., action detection-based analysis and joints motion analysis.

2.4.1 Action Detection Based Analysis

There is a defined set of activities that are effective to monitor a person's rehabilitation process, a disorder or motor deterioration. These are most commonly referred to as activities of daily life (ADL) [156]. ADLs are the basic activities that include the fundamental activities that a person would normally perform in a routine fashion such as personal care, mobility and eating. The skills required to perform that activity are learnt in earlier stages and are more preserved even in declining cognitive functions. Another set of complex activities are referred to as Instrumental ADLs (IADL) that include other things such as managing expenses and medications. IADLs are learnt in the later stages so are not well-preserved thus can be difficult to perform.

Abnormality or neurological disorder detection by action performance

The movements of the human body can be utilized by neurologists for the prognosis or diagnosis of a potential problem. Movement disorders as classified by the Movement Disorders Society, include (most commonly studied ones are listed here) Parkinson's disease (PD), dystonia (involuntary movement), Huntington's disease (HD), ataxia (lack of coordination), tremor and essential tremor (ET), myoclonus (rapid, irregular movement), Tourette syndrome (an inherited disorder characterized by multiple vocal and motor tics) and epilepsy. As many disorders affect the movements of various limbs, many kinds of research have been carried out to diagnose the disorder before it does much damage. This can be broadly divided into three areas such as lower-body analysis, full-body analysis and accuracy of ADLs performed.

For abnormality and neurological disorder detection, many researchers have explored the area of monitoring the lower-body, i.e., the gait of the person. This involves only one activity, i.e., walking and the movement of the legs of the subject is monitored. In [157] Cho et al. proposed a three-staged framework that involved silhouette extraction, and then Principal Component Analysis on silhouettes to linearise the feature matrices. In the end, they used a minimum distance classifier to classify if a person is suffering from PD or not. Li et al. [158] proposed to use 3D skeletons to find a gait representation using the covariance-based descriptor. In the end, they classify the gait as normal, hemiplegic or parkinsonian. Nguyen et al. [159] and Khokhlova et al. [160] used an RGB-D camera to get 3D skeletons. However, the former converted the 3D skeletons into codewords and used a log-likelihood estimation, whereas the latter used an LSTM-based model to assess between normal or abnormal gait.

Some other researchers have focused only on upper or full-body to detect or classify any disorder such as epileptic seizures etc. Lu at el. [161] proposed a video analytic system that exploits multi-coloured pyjamas for segmentation and tracking. Next, they estimated limb parameters and extracted displacement and oscillation features to detect a seizure. In [162] and [163] Ahmedt et al. proposed multi-modal CNN for the assessment of epilepsy. They first detect the patient on the bed and extract spatial features such as the face, limbs and hands. Then they extract temporal features using LSTM blocks and finally a classifier to detect a seizure. Karacsony et al. [164] used infra-red (IR) videos and proposed an end-to-end deep learning framework that extracted clinically known Spatio-temporal features of seizures.

For the quality assessment of a person's movement, action recognition and the efficiency of performing that action should be quantified. To do so, a specific set of activities, i.e., ADLs and IADLs are defined and performed by the subject which in turn helps in understanding the underlying abnormality. Pirsiavash et al. [165] proposed temporal pyramids and composite object models and developed a better feature representation for action detection. In [166],

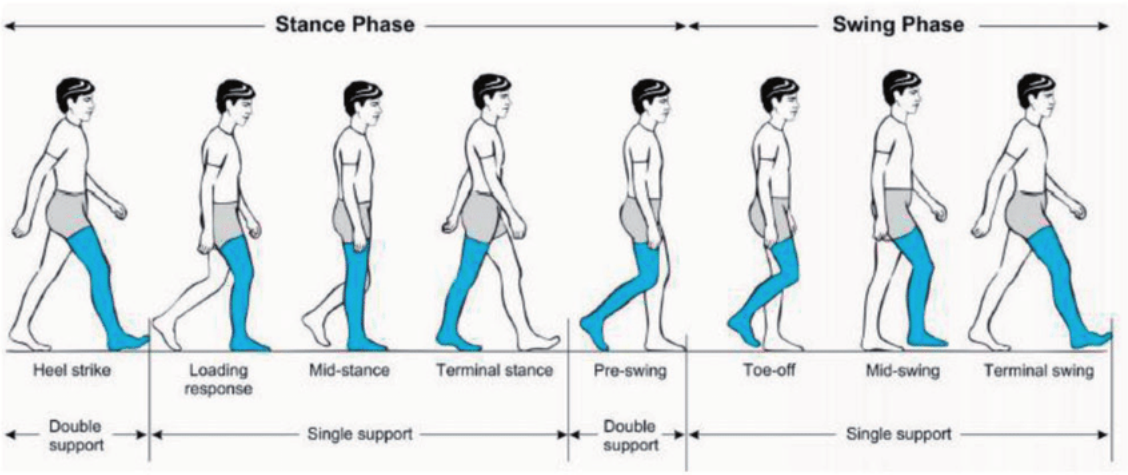

Figure 2.2: Human gait cycle. Image credit: Image taken from [170]

Elkholy et al. used a certain set of activities such as sitting, standing up, walking on a flat surface and climbing up. They proposed a framework, an end-to-end CNN, comprising of high-level and low-level feature descriptor along with two classification modules. First, for the classification of normal or abnormal motion and, second, to assess the efficiency of action performed.

Discriminative features and fundamental metrics

Better feature extraction is an important step towards the classification of normal and abnormal motions. Some researches have also explored distinct features and fundamental metrics to discriminate between pathological and normal motion. For example, using IMU or another wearable sensor, for an accurate distinction of pathology via gait analysis, researchers tend to estimate some common gait metrics such as cadence, stride length, stance phase, swing phase and walking speed etc. [167], [168] and [169]. The fig. 2.2 shows some main gait cycle characteristics.

Same metrics can also be realized for gait analysis, i.e., for vision-based solutions for gait abnormality detection. These vision-based methods for classification tend to extract high-level features from the person's silhouette or joint skeletal representation [171]. Rocha et al. [172] and Silva et al. [173] used RGB-D camera and proposed a Kinect based solution for the detection of PD. They analysed many high-level features and concluded that in PD detection via gait analysis, the velocity and acceleration of the centre shoulder play a pivotal role. Whereas, [174] emphasized that for clinical musculoskeletal modelling, better hip joint centre estimation is required as it plays an important role in the process of discriminating various pathologies.

On the other hand, other neurological diseases such as epilepsy the restricted movement of hands, the velocity and acceleration of joints, and facial expressions, are some critical factors in its detection [162], [163] and [175]. In [166], Elkholy et al. used a certain set of activities from daily life and also monitored various features of a person's motion such as Asymmetry between shoulders, the velocity of motion of joints and centre of mass.

Recently, in 2018, Luo et al. [176] proposed a framework which is, as the author's quote, the first of its kind, an AI-assisted healthcare system for the elderly which not only analyses the motion but also protects the identity of the person using it. It tracks a certain set of activities

(sitting, sleeping and using a bedside commode, etc) performed by the elders in their routine to monitor their health status over longer periods. They used multiple modalities, i.e., heat and depth maps. First, they recognize and record the activity performed by the elderly with timestamps and share that information with professional healthcare workers.

2.4.2 Joints Motion Analysis

Clinically, motion can be classified as pathological or normal by requesting the patient to perform one set of activities and monitor his condition. However, in a normal routine that is not possible. However, one can detect symptoms by monitoring the daily routine, i.e., ADLs in a controlled environment at home. Up till now, the research discussed requires at least one already planned activity to be performed by a person. However, daily, a person performs different kinds of activities that are very difficult to account for. So, surveillance methods that quantify motion rather than the activity can be more useful. However, this area hasn't been explored much.

Non-action detection (Joints motion Analysis) based motion analysis is more focused on a person's motion rather than on the accuracy of the action performed. It is usually done by extracting quality metrics from a person's motion, i.e., by tracking their silhouettes or joints. In [177], Lin et al. proposed a framework for the analysis of rehabilitation movements that segments the motion data. First, it applies segmentation as it recognizes motion candidates by velocity peaks and zero velocity crossings. Second, it uses hidden Markov models (HMM) to identify segment locations. Xue et al. [178] proposed a framework for monitoring the daily routine for the elders by tracking their skeletal joints over time and used time-series algorithms to analyse the data and the movement of the elders irrespective of the underlying activity.

2.4.3 Section Summary

This section covered literature on the topic of motion analysis by strictly using camera-based systems i.e., non-intrusive and non-invasive techniques for surveillance. There are potentially two areas that need to be explored, which are action detection based analysis (where the ease of the performance of an action is classified to judge if there is any problem) and non-action detection (joints motion analysis) based analysis (where the quality of motion is solely judged on the quality metrics extracted directly by joint locations or silhouettes from optical flow).

Recent research, such as [176] and [166] propose action based motion assessment frameworks. The former focuses on, first, action detection and second logging the action detected in a structured format for healthcare professionals. The latter, score the accuracy of the action and classify if the action is normal or abnormal. On the other hand, in the domain of monitoring daily routine irrespective of the action performed, recent research such as [178] show promising results. These focus on identifying any anomaly in their daily activities. However, to the best of my knowledge, none of the frameworks in either domain, deal with deterioration assessment and ease of action performance over long period of time.

2.5 Summary

One of the most important, essential and frequent actions that a person performs is eating. Moreover, a lot of different chains of actions (sub-actions) are performed during eating. For example, eating a butter sandwich in breakfast includes picking up a tool, scoop some butter, spread on the bread and put it in mouth. However, there is no solid research on usual sub-actions performed during eating.

To summarise, a complete motion estimation framework raises a question of how to quantify motion. This has been answered in the detailed literature review within this chapter in two ways, i.e., quantify the parameters of motion with respect to an activity and determine the the parameters of motion irrespective of the activity. On one hand, quantifying the accuracy of the motion undergoing a known activity requires a certain set of discriminative features to be extracted from a person's body preferably from his pose. On the other hand, if quantification is carried out irrespective of the activity, tracking joints and their time series analysis over time might help in detecting the problem (i.e., difficulty performing an action). However, the latter area hasn't been explored much.

Chapter 3

Methodology

We aim to develop a motion estimation system for the elders which supplies a detailed analysis of their eating behaviour. So, in this research, our focus will be to exploit both action-based and joint motion analysis (non-action based algorithms) for motion quality assessment over long periods. The overall methodology will constitute of three major parts, first video data acquisition, second, human action recognition with the accuracy of action being estimated and third, track the joints of the pose of a person and assess the motion quality.

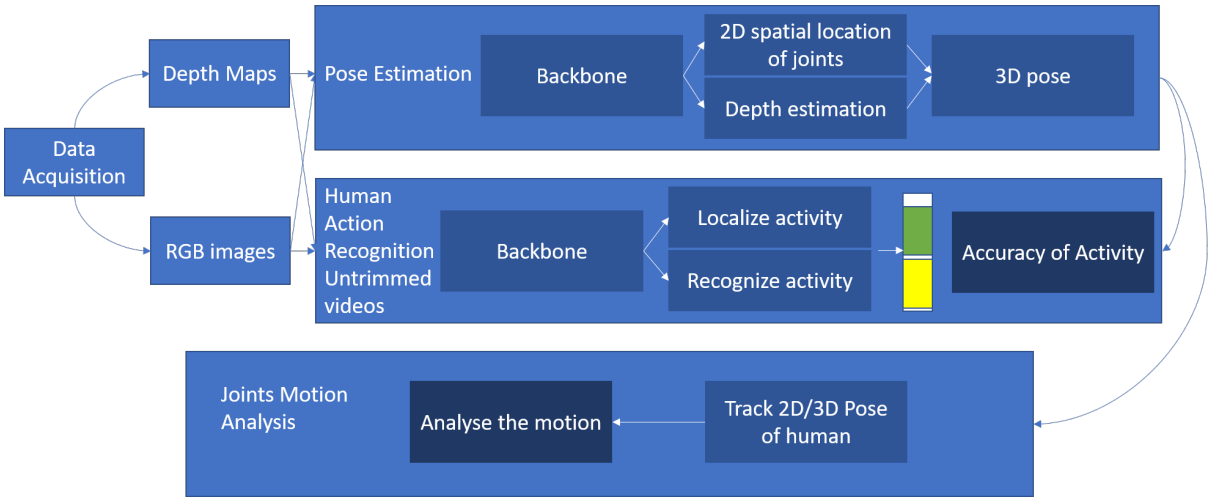

Figure 3.1: Block diagram of motion quality assessment

The main pipeline of the motion quality assessment framework is shown in fig. 3.1. First, data will be acquired from subjects under a contained and already known environment, preferably in a dining environment. Afterwards, we will formulate two end-to-end deep learning frameworks (CNN and RNN) which include a unified sub-action recognition framework that implicitly classifies deterioration of motor movement of the sub-action under consideration. A joint motion analysis framework that explores the avenue of motion assessment by tracking the joints irrespective of the ongoing activity.

3.1 Data Acquisition

Eating is one of the main, regular and most important actions of one's daily life. Monitoring it and its sub-actions for longer periods can potentially indicate any major anomaly such as the presence or start of a neurological disorder or deterioration over time for elders. So, we plan to acquire data in a known environment, focus on a dining table or a couch where a person usually sits and eats. This will be done using Intel realsense D415 and D435 RGB-D cameras, the details of the simulation environment are discussed in detail in chapter 4 section 4.2.

For that purpose, we simplify the full body pose to just the position of a few joints, which are head, neck, left and right shoulder, left and right elbow, and left and right wrists. Moreover, we have worked out four flow diagrams, each signifying common action set performed while seated on the dining table. These include sub-actions involved drinking from a cup or a glass, fig. 3.2, sub-actions involved after picking up tool in one hand, fig.3.3, atomic actions involved if a person picks tools in both hands, fig. 3.4 and if a person eats food directly with his hand, fig. 3.5.

Furthermore, we ensure to complete all the requirements set out for the approval of the university of Edinburgh's Informatics ethics committee and get their approval prior to collecting the data. Additionally, the dataset collected for this research will abide by the standards of the ethical data collection and storage by the General Data Protection Regulation (GDPR).

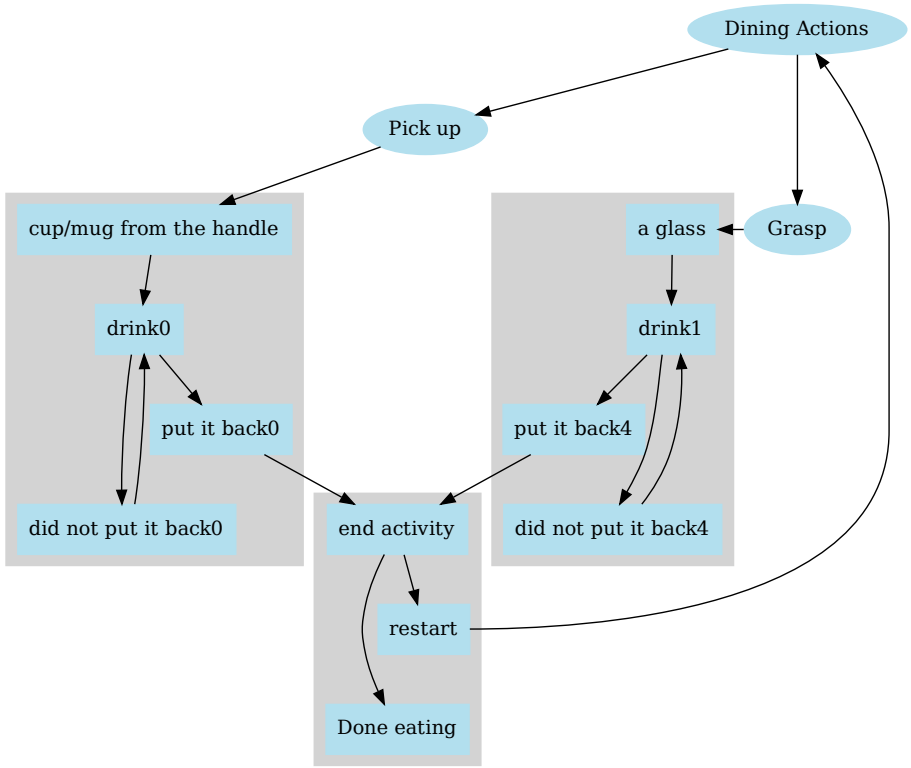

Figure 3.2: The figure shows the common actions performed by hand which are drinking something from a cup or a glass. They both involve different hand gestures, i.e., picking up or grasping respectively.

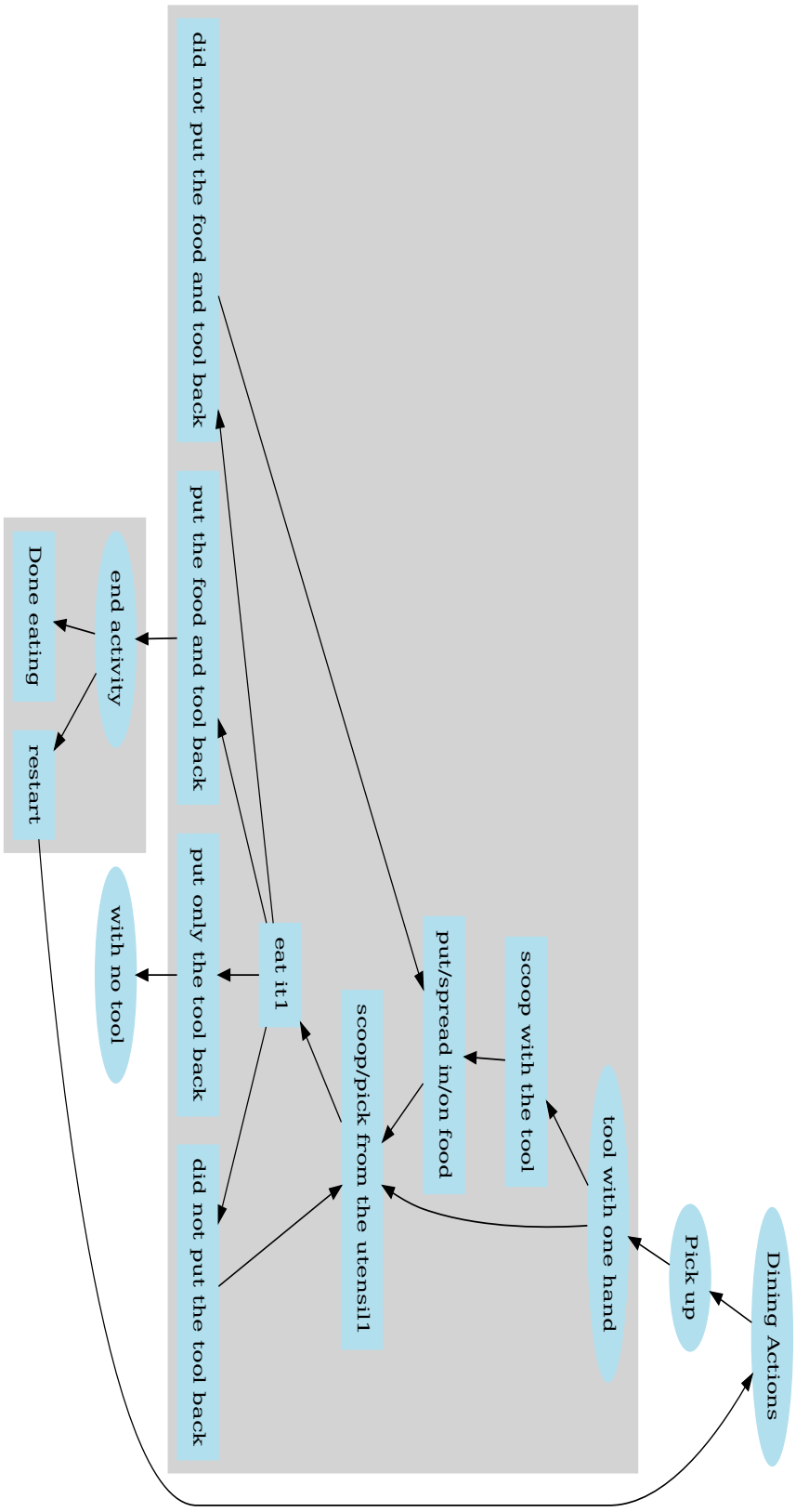

Figure 3.3: The figure shows the common actions performed by hand after picking tool in one of the hands. Mostly, the tools involved are either spoon or fork (Some cultures use different tools such as Chinese use chopsticks). These are commonly used to pick/scoop the food from the plate

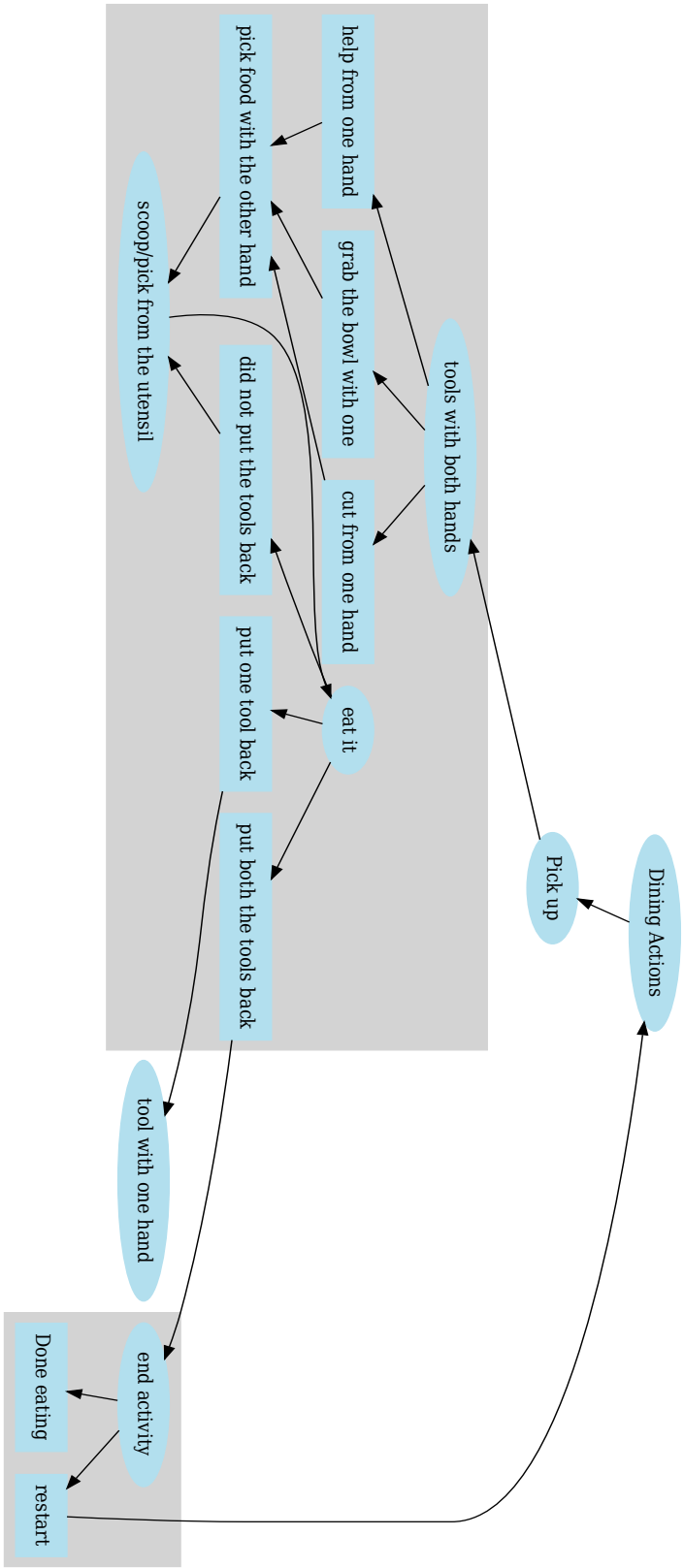

Figure 3.4: The figure shows the common actions performed by hand after picking up one tools in both the hands. Mostly, the tools involved are a combination of spoon and fork or fork and knife. These are commonly used to pick/scoop the food from the plate with the help of the second tool or cut something and then scoop the food.

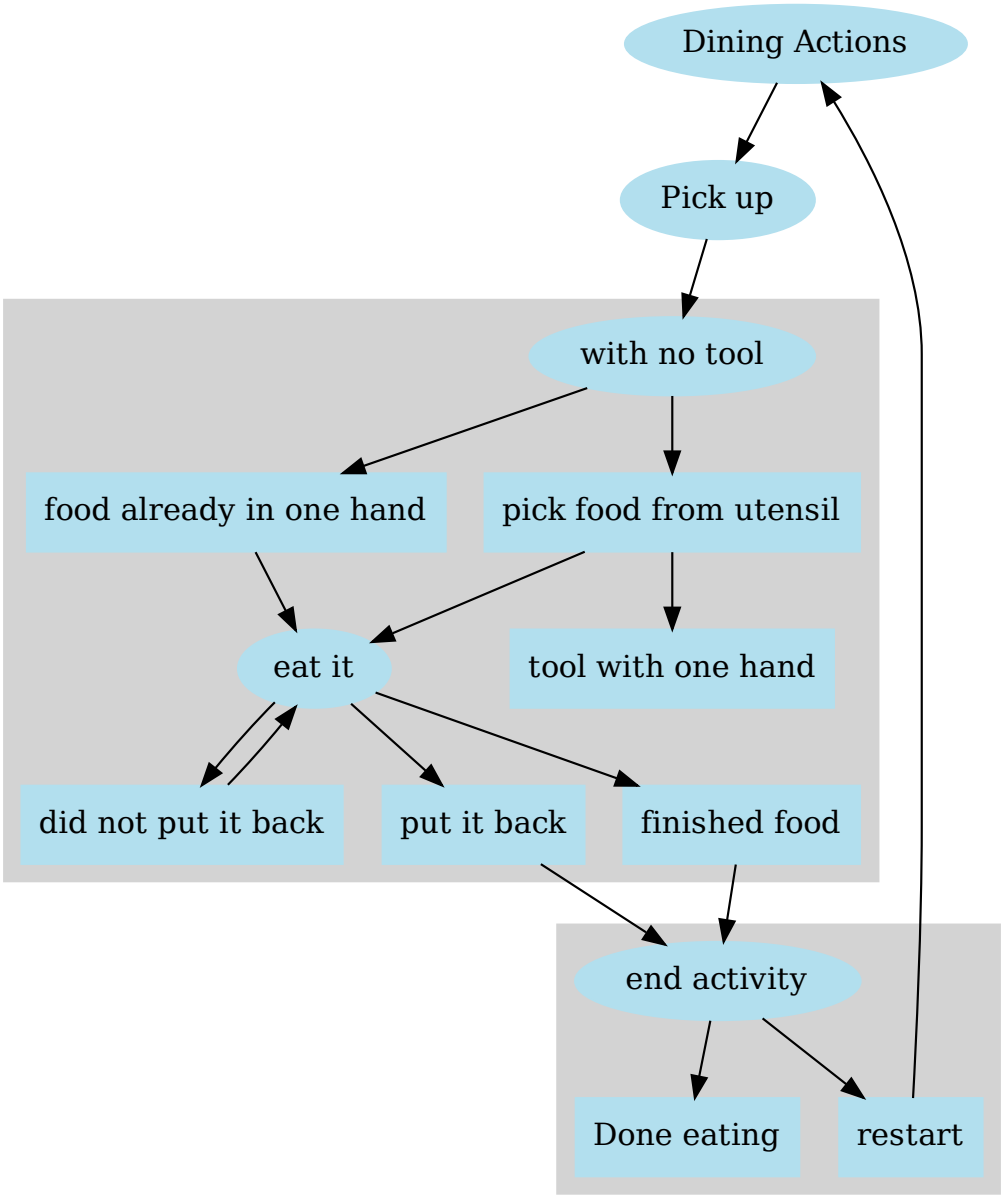

Figure 3.5: The figure shows the common actions performed by hand with no tool in hand. These involve the use of hands to directly pick up food from the utensil and eat it.

3.1.1 Data label Abstractions

Multiple levels of abstraction for the labels of data will be explored, which are shown in fig. 3.6. Altogether there are five levels of abstractions.

- First of all, the video will be divided into four portions that mark the deterioration stages over time, i.e., can eat easily, eat with difficulty (meaning: one can eat properly with-

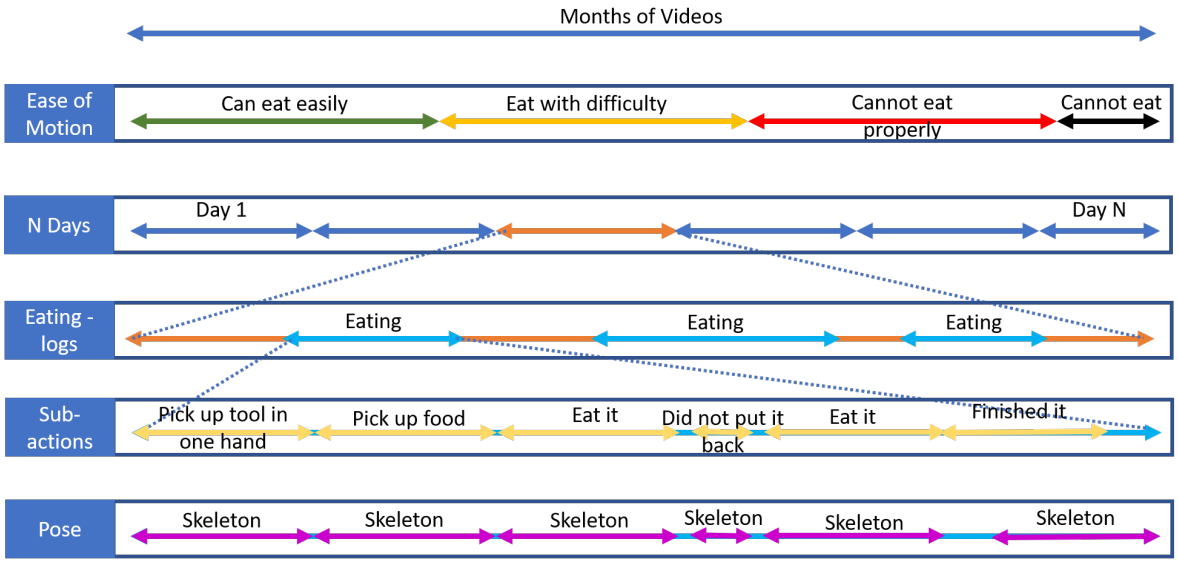

Figure 3.6: Multiple layers of abstractions for data labels

out spilling), cannot eat properly (meaning: one spills and drops food all over but can somewhat eat it) and cannot eat at all.

- Secondly, data will be divided into *N* days.
- Thirdly, these days will be further sub-divided into meals of the day (generally, breakfast, lunch and dinner).
- Fourth, each mealtime will be broken down into sub-classes of sub-actions performed while eating, such as, pick up the food, eat it and if he did not put the food back, again eat it and if the food is finished then the activity ends.
- The fifth level of abstraction will be a sub-action-wise pose, i.e., skeleton labels. This can potentially be useful for the evaluation of motor deterioration and sub-action recognition.

3.2 Action Recognition Motion Assessment

There are multiple main requirements in building a Human centred action recognition framework with the inherent capability of determining the ease of action. Firstly, as our task involves videos and sub-actions that might occur for very short time periods or may occur repeatedly, we need to be able to recognize small actions. Secondly, the network should also be able to understand the semantics and context of the scene for accurate action recognition. Thirdly, not only the network should be able to recognize the action, it should be able to temporally localize the action in an untrimmed video. Thus, overall it becomes a temporal action classification and localization problem.

3.2.1 Plan A: Simultaneous Recognition and Localization

Deep learning-based fast object detectors work on the principle of simultaneously localizing the object and recognizing it. Similarly, action recognition tasks can also be thought of simultaneously temporally localizing the action and also identifying it. This means we can think of action recognition in terms of object detection. However, localizing an object utilizes only spatial features whereas, for action, both Spatio-temporal features are needed. Fig. 3.7 shows the proposed end-to-end CNN for human sub-action recognition.

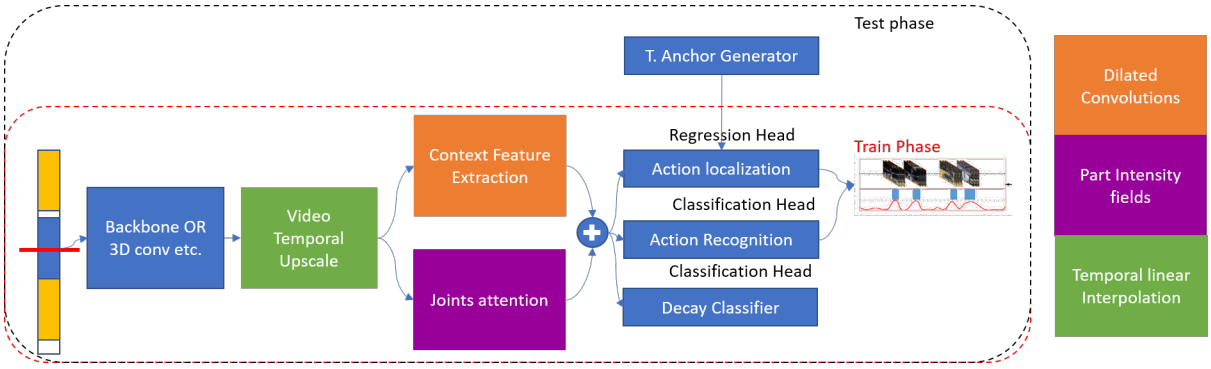

Figure 3.7: Block diagram of action recognition framework as per plan A

The proposed architecture is conceptually similar to an object detection framework with a backbone (ResNet or ResNeXt etc.), two necks fused to extract a certain set of discriminative features according to our application and four heads with various outputs that lay the foundation of a unified action detection and accuracy estimation framework. Moreover, it will utilize YOLO like training labels which incorporate time stamps for actions start and end and binary labels if there is an action and how difficult it is to perform that action.

Video Temporal Up-scale

As we want to identify sub-actions involved while eating, there is a very high chance the action clips are very short and thus does not provide much information. To solve this problem, we propose to use a temporal up-scaling technique which will try to mimic what a human does when an interesting short clip just floats away (a person scrolls back and re-play it with a slower speed). This temporal dimensional magnification can be achieved by interpolation techniques, such as linear interpolation.

Context Feature Extraction

Context information is a critical piece of information for action classification. For example, if a person picks up a burger, he is more likely to eat it. So, embedding and utilizing the context information requires a wider receptive field to cover the context regions. For this purpose, dilated convolutions can be used to extract the aforementioned information.

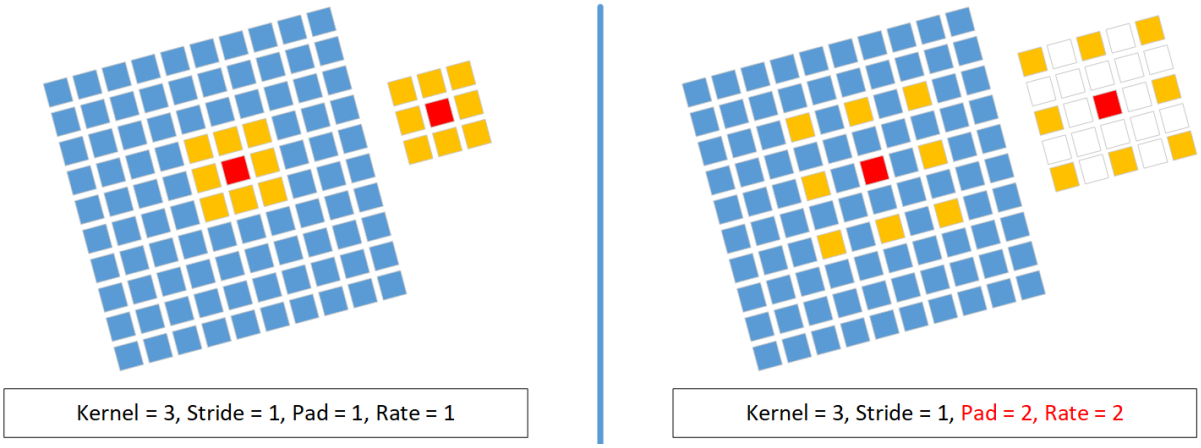

Figure 3.8: Left image shows the Standard Convolution and the right image shows Dilated or Atrous Convolution

A comparison of standard and dilated convolutions can be seen from Fig. 3.8. Consider a 1-D signal $x[k]$ and a filter $w[k]$ of size K, then by the definition of convolution, the eq. 3.2.1 represents a dilated convolution, where rate r, corresponds to stride. When $r = 1$, the above expression is reduced to a standard convolution, i.e., a standard convolution is a special case of the dilated convolution.

$$y[i] = \sum_{k=1}^{K} x[i + r.k]w[k] \tag{3.2.1}$$

After applying multiple dilated convolutions max pooling operation is applied to get dense contextual information.

Joint Attention

Attention to joints is also necessary for the ease of action classification part as it exploits the movement of each of the limbs of the person, whether s/he is stretching his arm fully or he used to do it last week but now he doesn't. To exploit attention towards the joints of a person we propose to use part intensity fields that help in identifying and localizing the body joints. In part intensity fields, first, a Gaussian with some mean and variance is placed where the key point for a body joint is, for generating ground truth for joints. Afterwards, a weighted mean squared error is calculated between the output feature map and the newly generated ground truth with Gaussian.

The loss function, a weighted mean least squared error, for part intensity fields is given in eq. 3.2.2, where S_j^* is the ground truth with gaussian confidence map and S_j^o are the output feature maps. Additionally, $\mathbf{W}(\mathbf{p})$ represents a non-negative diagonal weight matrix, $j \in 1, 2, ..., J$ which represent the joint number and pixel location $\mathbf{p} \in \mathbb{R}^2$.

$$f_{loss} = \sum_{j=1}^{J} \sum_{p} \mathbf{W}(\mathbf{p}).||\mathbf{S}_j^o(\mathbf{p}) - S_j^*(\mathbf{p})||_2^2 \tag{3.2.2}$$

Moreover, S_j^* is shown in eq. 3.2.3 where $\mathbf{x}_j$ is the ground truth position of the body joint j.

$$\mathbf{S}_j^*(\mathbf{p}) = exp(\frac{-||\mathbf{p}-\mathbf{x}_j||_2^2}{\sigma^2}) \tag{3.2.3}$$

Classification and Regression Heads

Altogether there are two classification and one regression heads as shown in fig. 3.7. Classification heads primarily predict the class of the sub-action being performed and whether it is easy for the person to perform the sub-action or the person is facing difficulties. The regression head on the other hand predicts temporal anchor localization timestamps. The accuracy of action will depend on the comparison between the action performed by a normal subject and abnormal subjects.

Train and Test Phase

In the training phase, first, the video is divided into N parts. Labels are generated accordingly. For simplicity, in this example, we assume only two action classes C_0 and C_1. Second, the first video instance goes into the backbone and neck to extract important features including context and joints. Third, a set of random anchor sizes (random start and end timestamps) are generated which try to localize the activity as best as possible. Fourth, feature maps are passed to the heads where the loss is calculated and then back-propagated to update the weights.

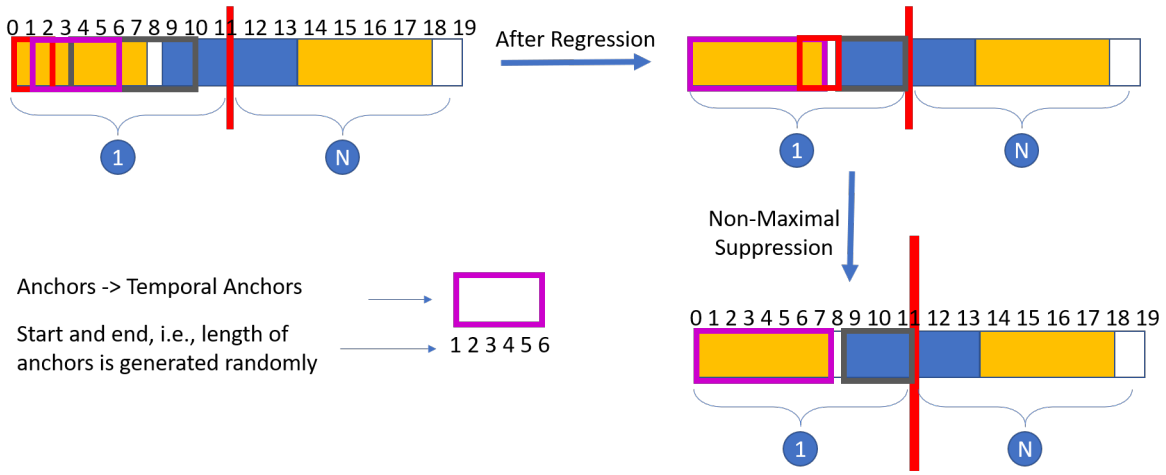

Figure 3.9: Randomly generated anchors after the regression and non-maximal suppression localizing the activity

After some epochs, the network starts to learn and its weights start to generalize the underlying multi-instance problem. In the test phase, an unknown video sequence is an input into the network which goes through the forward pass. A set of random anchors is generated which is then corrected by the regression head (action localization block) for localization.

For example, a video of 20 second length, contains two activities. In the test phase, the video is divided into N video instances (in this case $N = 2$) and the first instance is sent into the network and features are extracted. Afterwards, some anchors of random lengths are generated at first and then corrected with the predicted output of the regression head of the network. In

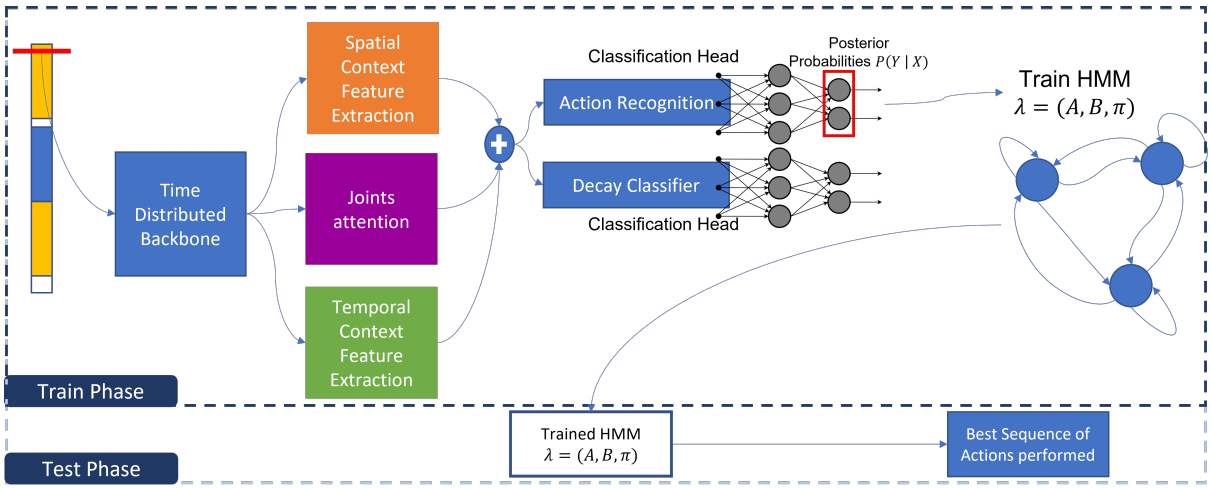

Figure 3.10: Block diagram of action recognition framework for plan B

the end, after correction by regression and non-maximal suppression (NMS), only two are left (the rest are suppressed by NMS) that localize the activity. This example is also demonstrated pictorially in fig. 3.9.

3.2.2 Plan B: Frame-by-Frame Action Recognition

We plan to use our framework in a real home environment. So it requires that the framework should be accurate and its computational time should be near-real-time. So, to address this issue, we plan to remove the temporal localization part completely as it adds a lot to the computational complexity and focus on frame-by-frame sub-action recognition. Frame-by-frame algorithms are mostly preferred in object recognition where they do not need the temporal context in any way. Also, for action recognition frameworks, in a frame-by-frame analysis some actions might have the same spatial joint locations in 2D/3D space as others, so designing a person centred frame-by-frame network might not be an effective solution.

As images in videos occur in chronological order and so to learn about movements, we need to emphasize on temporally distributed exploitation of data. For this purpose, we will use time distributed layers in the backbone. Moreover, we plan to exploit temporal context features along with spatial features such as joint locations over time and context-aware feature maps respectively. The block diagram of the proposed network is shown in fig. 3.10. Joints attention will be exploited by using part intensity fields and context-aware feature extraction by dilated convolutions, in a similar fashion mentioned in 3.2.1. Afterwards attention modules and before classification heads, long-short-term memory (LSTM) layers will be used to effectively capture temporal features temporally.

Temporal Context Feature Extraction

For temporal context feature extraction, we plan to use 3D motion history images to exploit the movements of a person in the temporal domain. The 2D motion history images (MHI) is a kind of finite-difference time-domain method. It takes images from different timestamps in a continuous image sequence and compares two or three adjacent pixels and then extract the

human body moving regions in the image by setting a threshold. However, utilizing the depth information from the depth camera we can potentially overcome the limitations of the 2D MHIs after we project them in the 3D domain.

3.2.3 Hidden Markov Model

To filter out any noise and missing data, we propose to use a hybrid CNN-HMM model to capture the essence of sub-actions or choice of sub-actions after another. Assume dataset D is the set of input images and their respective labels, i.e., $x_t \in D$ is the input image x at time t and $y \in D$ at time t is its corresponding label. Then CNN outputs a posterior probability given as $P(y_t|x_t)$ for each sample. On the other hand, a Hidden Markov Model (HMM) is initialized using the data, i.e., initial probabilities are set to be $\pi_0 = 0.5$ for each state, initial transition probabilities are set by counting the transitions from one action to another in a video sequence. Eventually, an HMM model $\lambda_0 = (A_0, B_0, \pi_0)$ is initialized.

Posterior probabilities from the CNN and initial mode λ_0 are then used to train an HMM model $\lambda_{tr} = (A_{tr}, B_{tr}, \pi_{tr})$ by using Baum-Welch algorithm.

3.2.4 Train and Test Phase

In the training phase,First HMM is initialized and then both CNN and HMM learn simultaneously. HMM is trained according to the posterior probabilities predicted by the CNN using Baum-Welch algorithm. In the test phase, first, CNN predicts a probability and then is utilized by the trained HMM to fine tune the prediction of the sequence. We plan to use Viterbi algorithm to get the best possible sequence of actions from the set of images.

3.3 Joint Motion Analysis

The main requirements for constructing a joint motion analysis, framework, first, to estimate the pose of a person in 3D space, is to identify the joint locations precisely. Second, to track the joints over time. Third, extract a discriminative motion descriptor that exploits both physical and geometric, spatial and temporal features. Last, a classifier to distinguish between classes of deterioration. The complete block diagram of the proposed framework is shown in fig. 3.11.

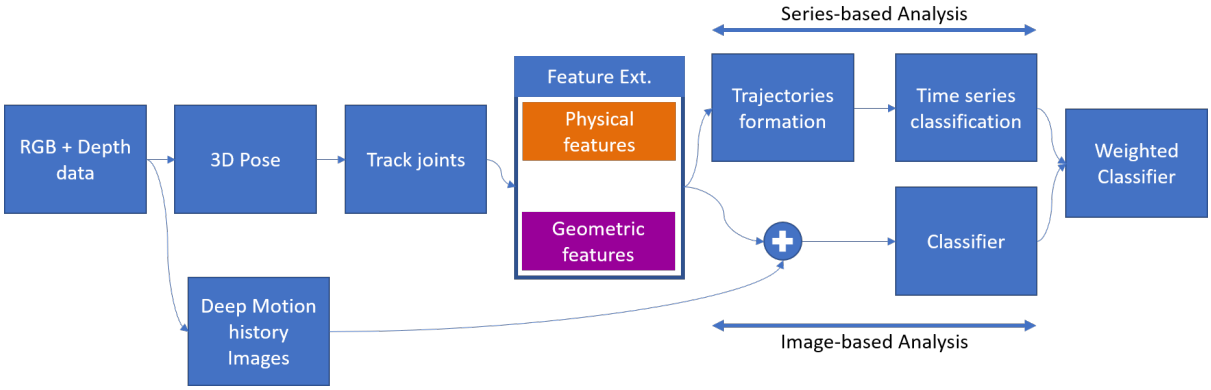

Figure 3.11: Block diagram of joint motion estimation framework

3.3.1 Pose Estimation

For the 3D pose estimation part, we simplify the full body pose to just the position of a few joints, which are head, neck, left and right shoulder, left and right elbow, and left and right wrists. We plan to use a robust framework that has been tested on various publicly datasets, an end-to-end deep learning framework known as A2J: Anchor to Joints, which uses depth maps to identify spatial key point locations and estimates depth. Its block diagram is shown in fig. 3.12.

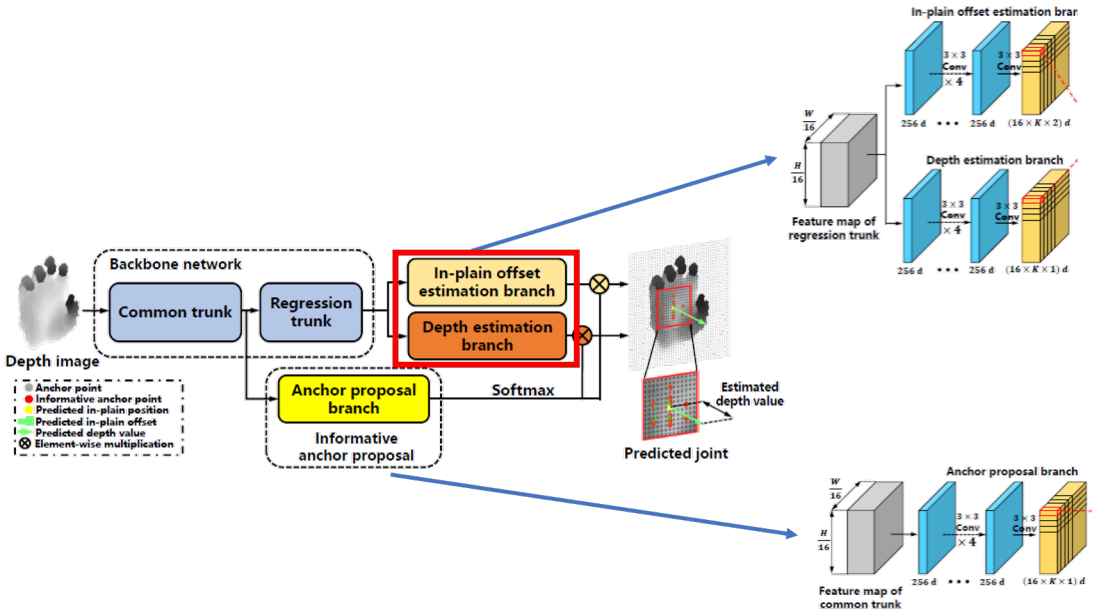

Figure 3.12: A2J: Anchor to Joints framework [45]

A2J first uses a backbone to extract features and then uses two heads to predict the 3D pose. One head finds the spatial location of the joint and the other estimates the depth value of the joints in the 3D space. The in-plain and depth value of joint j is estimated as the weighted

average shown in the eq. 3.3.1, where a is an anchor point and it belongs to anchor point set A, $S(a)$ is in-plain position of anchor point a, $P_j(a)$ is the response of anchor a towards joint j, $O_j(a)$ is the predicted in-plain offset towards joint j from anchor point a and $D_j(a)$ is the predicted depth value of joint j by anchor point a.

$$\begin{cases} \hat{S}_j = \sum_{a \in A} \tilde{P}_j(a)(S(a) + O_j(a)) \\ \hat{D}_j = \sum_{a \in A} \tilde{P}_j(a) D_j(a) \end{cases} \tag{3.3.1}$$

where,

$$\tilde{P}_j(a) = \frac{e^{P_j(a)}}{\sum_{a \in A} e^{P_j(a)}} \tag{3.3.2}$$

3.3.2 Motion Descriptor

After estimating the 3D pose of the person from depth data, the spatial location of the joints is tracked over time. Afterwards, we extract discriminative features such as various physical (acceleration and velocity) and geometric features (joint position and angles), to form meaningful motion representations/descriptor. These representations are then divided into two streams to exploit both image and sequential data analysis techniques.

3.3.3 Two Stream Classification

On one side, trajectories for each of the motion descriptor are formed along with a time-series based classifier to distinguish between deterioration classes. On the other side, the motion descriptor is fused with motion history images to explore the temporal context and then a classifier will classify between the decay of motor movement classes. Lastly, a weighted or a simple vote based classifier dependant on the result of both the streams will be designed to get a refined judgement over the deterioration of motor movement.

3.4 Summary

Generally, camera-based motion analysis systems are of two types, i.e., HAR-based and Joints motion analysis frameworks. We propose a unified deep learning end-to-end framework that estimates HAR and its ease of performance. Additionally, the joint motion analysis system explores a time series model along with an image-based feature classifier to classify motor deterioration, irrespective of the class of the activity. Overall, in this chapter, both of those topics have been explored and an approach to tackle the problem accordingly has been presented. The block diagram in 3.13 shows the big picture of the proposed algorithms.

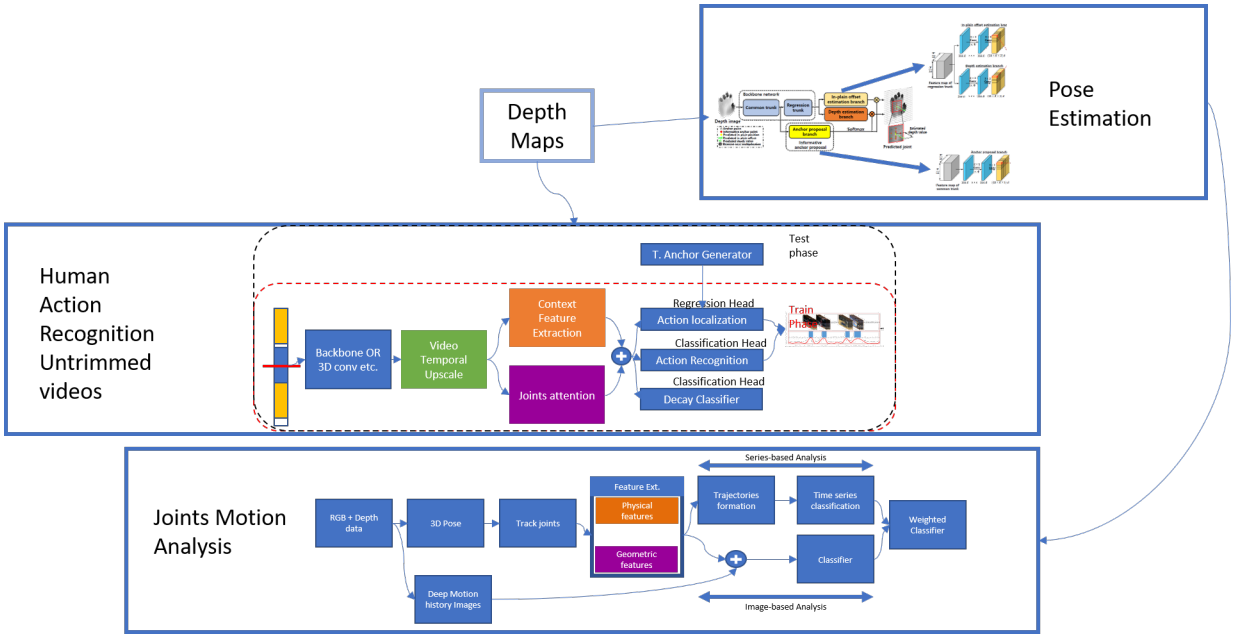

Figure 3.13: Summary of the proposed approach

Chapter 4

Evaluation

4.1 Public Datasets

There are a lot of publicly available action detection databases that include fine-grained actions. Some of these focus on everyday activities performed by an individual and some focus on minute actions in unconstrained environments but most of them only rely on RGB video feed and do not utilize depth untrimmed videos of actions. This section contains a brief description of publicly available datasets which particularly can be useful for our application.

- **NTU RGB-D** is one of the biggest benchmark datasets for action recognition. It contains daily actions, mutual actions (group actions), and medical conditions. It has been collected via multiple sensors such as RGB, depth and infra-red.
- **Sphere** is specifically designed to evaluate the quality of human movements from visual information via gait analysis. It contains walking up and down the stairs and sitting and standing actions. It has been collected via only a depth camera in a constrained and known background.
- **MSR DailyActivity3D** is designed to model daily actions performed by a person while sitting on a couch. It contains actions such as playing the guitar and watching television. It has been collected via a depth and RGB camera.

4.1.1 Proposed Dataset

We believe eating is one of the most common and frequent actions in one's daily routine and can be exploited to evaluate motor movement. So, we intend to develop a dataset for composite / sub-actions for eating activities that include actions such as 'pick a tool', 'scoop the food' and 'eat it', etc. We will use an RGB and depth camera to collect the data, whereas, for privacy reasons, only the computer will be able to see and process the collected information and other people will not be given access. Like the dataset Sphere, the dataset will be annotated with normal/abnormal sub-actions. The videos will be trimmed to primitive actions, and joint positions will be manually annotated in 2D (which allows 3D position estimation). Moreover, transition over time from 'can eat easily' to 'cannot eat' will be observed by adding weights to simulate increasing weakness.

4.2 Eating Dataset

The eating dataset will be focused on sub-actions a person performs while eating. We plan to use multiple modalities such as RGB and depth cameras for data collection. For this purpose, Intel Real Sense 435 and 415 will be used as data acquisition tools. Moreover, daily eating actions of at least two subjects of different age groups and different cultural backgrounds will be monitored. As deep learning frameworks are data-driven so we will collect daily eating data for over three months and the subject will be given an open choice to perform any action at any instance. Moreover, we plan to make this proposed dataset publicly available for deterioration assessment. So, to make it easier for the user, both trimmed and untrimmed versions of the videos of the dataset will be released and annotated.

4.2.1 Data Classes

There are various abstraction levels that we intend to explore for this dataset, which is discussed in section 3.1.1. Overall, we propose to have four classes for the assessment of gradual motor deterioration over time. Furthermore, to cover a wide spectrum of primitive actions involved in eating, we propose to use a generalized most common set of twenty-two actions. A detailed table 4.1 about the classes involved (and details on how they will be simulated) for each level of abstraction.

4.2.2 Data Labelling

Labelling videos is a laborious task and often erroneous at different levels. For labelling actions, we plan to use an open-source manual annotation tool. On the other hand, for labelling 2D pose, we plan to devise a semi-automatic neural network (network that requires user input for each prediction) spatial joint tracker. Moreover, the 2D pose will then be projected into the 3D space by using conventional computer vision techniques.

4.3 Action Recognition Evaluation

4.3.1 Metrics

For action detection frameworks based on temporal anchors which do both detection and classification simultaneously, include both true positives and false positives in the measurements. Metrics that exploit this with a threshold on IoU (if the IoU is greater than a certain threshold the detected activity is referred to as correctly localized) is required for better classification and localization activity detection. Three such metrics are mean average precision, average recall and receiver operator characteristics.

- **Mean Average Precision** (mAP)
  To evaluate the classification accuracy of the predicted temporal localization timestamps (to mark the start and end of action) and action class versus the ground truth localization timestamps and classes, mean average precision will be calculated. For mAP, the IoU threshold, which can be varied, is set to a specific value and values greater than the

Abstraction	Classes	Details
Motor Deterioration	Can eat easily	subject with no additional weight tied around his wrist
	Eat with difficulty	subject with 1 Kg of additional weight tied around his wrist
	Cannot eat properly	subject with 2.5 Kg of additional weight tied around his wrist
	Cannot eat without help	subject with 5 Kg of additional weight tied around his wrist
Eating logs	Eat	Breakfast - 6 am to 10 am
		Brunch - 10 am to 1 pm
		Lunch - 1 pm to 4 pm
		Supper - 4 pm to 7 pm
		Dinner - 7 pm to 12 Midnight
		Midnight Snack - 12 Midnight to 6 am
Eating sub actions	No activity	Recorded while eating
	Non-recognizable	
	drink	
	eat	
	Pick up a mug/cup	
	Pick up a tool in one hand	
	Pick up tools in both hands	
	Pick up food/utensil with one hand	
	Pick up food/utensil with both hands	
	Pick up food/utensil with one hand and tool from other	
	Scoop with tool	
	put in / spread on	
	scoop / pick from utensil	
	put the tool back	
	put only the tool back	
	put the food back	
	put only the food back	
	put one tool back	
	put both the tools back	
	Grasp a glass	
	put it back	
	did not put it back	

Table 4.1: Table for classes and details for different level of abstractions

threshold, either true positives, false positives etc., are used to form a confusion matrix. A confusion matrix is an $N \times N$ matrix that is used for evaluating the performance of the classification model. Two metrics that essentially summarise confusion matrix parameters are precision and recall. The equations for precision and recall are given in eq. 4.3.1 and 4.3.2, respectively.

$$p = Precision = \frac{Truepositives}{TruePositives + FalsePositives} \tag{4.3.1}$$

$$r = recall = \frac{Truepositives}{TruePositives + FalseNegatives} \tag{4.3.2}$$

If precision and recall are plotted with recall on the x-axis and precision on the y-axis of each predicted sample, we get a precision-recall curve. Generally, the area under the precision-recall curve is referred to as average precision. Average precision at a set of finely spaced recall points $M \in [0,1]$ where it interpolates the corresponding precision for a certain recall value r by taking the maximum precision whose recall value $\tilde{r} > r$ shown in eq. 4.3.3

$$AP = average\ precision = \int_{r \in M} \max_{\tilde{r} > r} p(\tilde{r}) dr \tag{4.3.3}$$

The mean over N classes is effectively the mean average precision (mAP) eq. 4.3.4.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{4.3.4}$$

- **Average Recall** (AR)
  Similarly, along with mAP, we will also compute average recall for measuring the action classification accuracy. For average recall, the IoU thresholds are varied usually from 0.5 to 1. It describes the area under the recall-IoU curve. Consequently, its mathematical form is shown in eq. 4.3.5,

$$AR = average\ recall = 2 \times \int_{0.5}^{1} recall(IoU) dIoU \tag{4.3.5}$$

- **Receiver Operator Characteristic** (ROC)
  Receiver Operator Characteristic (ROC) is also an evaluation tool for classification problems that shows the general trend of the model in terms of the separability of each of the classes. Two metrics also derived from a confusion matrix are sensitivity and specificity defined in eq. 4.3.6 and 4.3.7, respectively. The curve formed by *sensitivity* on the y-axis and $1 - specificity$ on the x-axis is referred to as the ROC curve. The area under the curve (AUC) is the measure of how effectively the classifier discriminates between the classes.

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{4.3.6}$$

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{4.3.7}$$

4.3.2 Proposed Experiments

Accuracy of action recognition frameworks can be judged based on mAP, AR and ROC. Experimentations will be carried out on some public benchmark datasets that will establish the performance of our proposed framework. Moreover, a new dataset will be developed that will label the deterioration stages of a person's movement over long periods. Usually, there are no labels or numerical quantification available for deterioration, but we will establish action and higher-level abstractions in the dataset to classify the performance of the action. The labels on various levels of abstractions for action recognition and its normalcy detection will be done manually. In our experiments, detailed ablation studies with and without attention modules, which backbone works best, for our application will be carried out.

4.4 Joints Motion Assessment

4.4.1 Metrics

Motion assessment directly depends on the effective pose estimation step. So, pose estimation is an equally important step that needs to be done accurately. For joint location and pose estimation, different metrics have been used in the past including mAP (discussed in Section 4.3.1). Three of such metrics that are used for 3D pose estimation are Mean Per Joint Position Error (MPJPE), Mean Per Vertex Error (MPVE) and 3D Percentage of Correct Keypoints (3DPCK).

- **Mean Per Joint Position Error** (MPJPE)
  The most common and widely adopted evaluation metric is MPJPE for 3D joint spatial locations. It is the Euclidean distance between estimated 3D joints and ground truth positions. This is given in eq. 4.4.1, where N is the number of joints usually 16, J_i^* and J_i are the predicted and ground truth location of the i_{th} joint in 3D space, respectively.

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} ||J_i - J_i^*||_2 \tag{4.4.1}$$

- **Mean Per Vertex Error** (MPVE)
  Along with MPJPE, we will measure the accuracy of our estimated pose in 3D using MPVE. Mean Per Vertex Error is an evaluation tool that estimates the accuracy of the pose, i.e., the 3D joint locations. MPVE measures the Euclidean distance between the ground truth and the predicted vertices. MPVE is shown in eq. 4.4.2

$$MPVE = \frac{1}{N}\sum_{i=1}^{N} ||V_i - V_i^*||_2 \tag{4.4.2}$$

- **3D Percentage of Correct Keypoints** (3DPCK)
  3D PCK is also used to measure the accuracy of localization of different key points (spatial location of joints in 3D space). The higher the PCK value the better the model performance. Detected joints are classified as correct if the distance between the ground truth joints and predicted joints are within a threshold.

As indicated in the literature, low MPJPE does not necessarily mean that the estimated pose is accurate as it depends on the predicted scale of the human skeleton. 3D PCK on the other hand is more robust to erroneous measurements but it cannot evaluate the accuracy of the predicted pose. Also, each of these metrics is designed to evaluate the accuracy of the predicted human pose in a single frame. The temporal smoothness and consistency of reconstructed human pose cannot be evaluated with these metrics. For measuring the temporal smoothness, accuracy metrics and motion estimation, metrics such as acceleration error have been proposed recently.

- **Acceleration Error** (ACC-Err)
  To measure the joint motion speed, acceleration error will be used. Acceleration error measures the distance between the predicted and ground truth 3D acceleration for each key-point in mm/s^2. It acts as a smoothness indicator for estimated motion sequences. Acceleration is calculated per frame using the finite differences between each frame.

4.4.2 Proposed Experiments

Our primary experimentation will be carried out on our proposed dataset of eating actions. Nonetheless, deterioration of movement over the long-term will be explored, utilizing a time-series modelling set up. Primarily, the evaluation will be divided into two steps. In the first step, the accuracy of pose estimation and smoothness based on the criteria defined above will be carried out. In the second step, utilizing the accelerations of the poses, time series' will be formed, and further classified for any anomaly and deterioration over time. The classification accuracy will be determined by mAP (discussed in 4.3.1).

Although the practical nature of time series analysis on data has been established in the past as it efficiently exploits patterns over longer periods. So, in our experiments, we will explore and evaluate multiple stochastic models and a detailed analysis of which works well in identifying the change in a person's movement over long periods.

Chapter 5

Summary

5.1 Conclusion

Due to the growing number of elders, the need for better and effective healthcare facilities has increased. This increase in need has put a strain on healthcare systems worldwide. At this time, scientists are exploring new and intelligent ways to help the elders live healthy life. Traditional motion estimation frameworks are usually marker-based techniques, i.e., these require intrusive sensors. However, recent researches exploit the use of marker-less techniques such as camera-based surveillance of elders to monitor their daily routine. Camera-based surveillance can potentially diagnose or prognosis important problems which can help the elders to live a more independent life without any social cut-off. This research aims to develop an effective motion assessment framework that will monitor the deterioration of motor movements of an elderly person in the long term. Moreover, as it targets the deterioration assessment potentially due to low physical activity it contributes towards the frailty assessment of the elders.

A new dataset for the deterioration assessment of motor movements will be proposed. As eating is one of the repetitive and the most common actions in a daily routine, we assume that utilizing the actions performed while eating in a home setting, can help us explore major movement events and motor deterioration. The dataset will be labelled up to primitive actions that a person performs while eating. Thus, this dataset will exploit the ease of the movements over time and classify whether the person can eat properly if he/she feels any difficulty or he/she cannot eat properly and need help.

Furthermore, the proposed research explores both avenues of camera-based surveillance, i.e., action-based and joints motion analysis. An action-based analysis is dependent on an action recognition framework, which in our case should be able to identify sub-actions. Although the action recognition and localization domain has recently seen many advancements, our proposed end-to-end, sub-action detection and deterioration assessment framework is the first of its kind. On the other hand, joints motion analysis (non-action recognition based) frameworks depend on the movements of a person, rather than the activity to assess the deterioration. There has not been much work in the domain of joints motion assessment.

5.2 Workplan

PhD Research Plan

Complete In-process planned Exceeded plan Duration

Activity	Plan Start	Plan Duration	Actual Start	Actual Duration	Percent Complete	Period (Months) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Identify Research Problems	1	1	1	1	100%	
Writing Research Proposal	2	4	2	4	60%	
Data Acquisition for Dataset	5	3			0%	
Dataset Labelling	7	3			0%	
Action Recognition framework	10	4			0%	
Deterioration Assesment	14	1			0%	
Pose Estimation Framework	15	1			0%	
Motion Assessment framework	16	4			0%	
Deterioration Assesment	20	2			0%	
Elderly Data Collection	22	4			0%	
In-home Experiments	26	6			0%	
PhD Thesis write-up	31	6			0%	

References

[1] Kaare Christensen, Gabriele Doblhammer, Roland Rau, and James W Vaupel. Ageing populations: the challenges ahead. *The lancet*, 374(9696):1196–1208, 2009.

[2] Patricia Bet, Paula C Castro, and Moacir A Ponti. Fall detection and fall risk assessment in older person using wearable sensors: A systematic review. *International journal of medical informatics*, 130:103946, 2019.

[3] Steffi L Colyer, Murray Evans, Darren P Cosker, and Aki IT Salo. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports medicine-open*, 4(1):1–15, 2018.

[4] Kaibo Fan, Ping Wang, and Shuo Zhuang. Human fall detection using slow feature analysis. *Multimedia Tools and Applications*, 78(7):9101–9128, 2019.

[5] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey, 2021.

[6] Antonio S Micilotta, Eng-Jon Ong, and Richard Bowden. Real-time upper body detection and 3d pose estimation in monoscopic images. In *European Conference on Computer Vision*, pages 139–150. Springer, 2006.

[7] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[8] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[10] Sijin Li, Zhi-Qiang Liu, and Antoni B Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 482–489, 2014.

[11] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1347–1355, 2015.

[12] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[13] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *arXiv preprint arXiv:1406.2984*, 2014.

[14] Ita Lifshitz, Ethan Fetaya, and Shimon Ullman. Human pose estimation using deep consensus voting. In *European Conference on Computer Vision*, pages 246–260. Springer, 2016.

[15] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[17] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.

[18] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *proceedings of the IEEE international conference on computer vision*, pages 1281–1290, 2017.

[19] Wei Tang and Ying Wu. Does learning specific features for related parts help human pose estimation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1107–1116, 2019.

[20] Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. Modeep: A deep learning framework using motion features for human pose estimation. In *Asian conference on computer vision*, pages 302–315. Springer, 2014.

[21] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015.

[22] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5207–5215, 2018.

[23] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014.

[24] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016.

[25] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018.

[26] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *arXiv preprint arXiv:2008.03713*, 2020.

[27] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3941–3950, 2017.

[28] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2344–2353, 2019.

[29] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2262–2271, 2019.

[30] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019.

[31] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *European Conference on Computer Vision*, pages 318–334. Springer, 2020.

[32] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *European Conference on Computer Vision*, pages 186–201. Springer, 2016.

[33] Jue Wang, Shaoli Huang, Xinchao Wang, and Dacheng Tao. Not all parts are created equal: 3d pose estimation by modeling bi-directional dependencies of body parts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7771–7780, 2019.

[34] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2020.

[35] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[36] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017.

[37] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2282–2292, 2019.

[38] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *arXiv preprint arXiv:1907.02499*, 2019.

[39] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10895–10904, 2019.

[40] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019.

[41] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision (to appear)*, December 2020.

[42] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017.

[43] Qiang Nie and Yunhui Liu. View transfer on human skeleton pose: Automatically disentangle the view-variant and view-invariant information for pose representation learning. *International Journal of Computer Vision*, 129(1):1–22, 2021.

[44] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018.

[45] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019.

[46] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. A multi-view rgb-d approach for human pose estimation in operating rooms. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 363–372. IEEE, 2017.

[47] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G Narasimhan, and Minh Vo. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In *European Conference on Computer Vision*, pages 492–509. Springer, 2020.

[48] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[49] Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. Skeleton-aware 3d human shape reconstruction from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5431–5441, 2019.

[50] Kangkan Wang, Jin Xie, Guofeng Zhang, Lei Liu, and Jian Yang. Sequential 3d human pose and shape estimation from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7275–7284, 2020.

[51] Mustansar Fiaz, Arif Mahmood, Sajid Javed, and Soon Ki Jung. Handcrafted and deep trackers: Recent visual object tracking approaches and trends. *ACM Computing Surveys (CSUR)*, 52(2):1–44, 2019.

[52] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip HS Torr. Staple: Complementary learners for real-time tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1401–1409, 2016.

[53] Zhizhen Chi, Hongyang Li, Huchuan Lu, and Ming-Hsuan Yang. Dual deep network for visual tracking. *IEEE Transactions on Image Processing*, 26(4):2005–2015, 2017.

[54] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative scale space tracking. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1561–1575, 2016.

[55] Heng Fan and Haibin Ling. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5486–5494, 2017.

[56] C. Ma, J. Huang, X. Yang, and M. Yang. Robust visual tracking via hierarchical convolutional features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (11):2709–2723, 2019. doi: 10.1109/TPAMI.2018.2865311.

[57] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson WH Lau, and Ming-Hsuan Yang. Crest: Convolutional residual learning for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 2555–2564, 2017.

[58] Mengmeng Wang, Yong Liu, and Zeyi Huang. Large margin object tracking with circulant feature maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4021–4029, 2017.

[59] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Multi-task correlation particle filter for robust object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4335–4343, 2017.

[60] Jongwon Choi, Hyung Jin Chang, Sangdoo Yun, Tobias Fischer, Yiannis Demiris, and Jin Young Choi. Attentional correlation filter network for adaptive visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4807–4816, 2017.

[61] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6638–6646, 2017.

[62] Hongwei Hu, Bo Ma, Jianbing Shen, and Ling Shao. Manifold regularized correlation object tracking. *IEEE transactions on neural networks and learning systems*, 29(5): 1786–1795, 2017.

[63] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning background-aware correlation filters for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 1135–1143, 2017.

[64] Alan Lukezic, Tomas Vojir, Luka ˇCehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6309–6318, 2017.

[65] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. Learning dynamic siamese network for visual object tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 1763–1771, 2017.

[66] Chen Huang, Simon Lucey, and Deva Ramanan. Learning policies for adaptive tracking with deep feature cascades. In *Proceedings of the IEEE international conference on computer vision*, pages 105–114, 2017.

[67] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2805–2813, 2017.

[68] Qiang Wang, Jin Gao, Junliang Xing, Mengdan Zhang, and Weiming Hu. Dcfnet: Discriminant correlation filters network for visual tracking. *arXiv preprint arXiv:1704.04057*, 2017.

[69] Kai Chen, Wenbing Tao, and Shoudong Han. Visual object tracking via enhanced structural correlation filter. *Information Sciences*, 394:232–245, 2017.

[70] Zhen Cui, Shengtao Xiao, Jiashi Feng, and Shuicheng Yan. Recurrently target-attending tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1449–1458, 2016.

[71] Si Liu, Tianzhu Zhang, Xiaochun Cao, and Changsheng Xu. Structural correlation filter for robust visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4312–4320, 2016.

[72] Alan Lukežič, Luka Čehovin Zajc, and Matej Kristan. Deformable parts correlation filters for robust visual tracking. *IEEE transactions on cybernetics*, 48(6):1849–1861, 2017.

[73] Bing Bai, Bineng Zhong, Gu Ouyang, Pengfei Wang, Xin Liu, Ziyi Chen, and Cheng Wang. Kernel correlation filters for visual tracking with adaptive fusion of heterogeneous cues. *Neurocomputing*, 286:109–120, 2018.

[74] Madan Kumar Rapuru, Sumithra Kakanuru, Pallavi M Venugopal, Deepak Mishra, and Gorthi RK Sai Subrahmanyam. Correlation-based tracker-level fusion for robust visual tracking. *IEEE Transactions on Image Processing*, 26(10):4832–4842, 2017.

[75] Guokun Wang, Jingjing Wang, Wenyi Tang, and Nenghai Yu. Robust visual tracking with deep feature fusion. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1917–1921. IEEE, 2017.

[76] Bolun Cai, Xiangmin Xu, Xiaofen Xing, Kui Jia, Jie Miao, and Dacheng Tao. Bit: Biologically inspired tracker. *IEEE transactions on image processing*, 25(3):1327–1339, 2016.

[77] Heng Fan and Haibin Ling. Sanet: Structure-aware network for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 42–49, 2017.

[78] Junyu Gao, Tianzhu Zhang, Xiaoshan Yang, and Changsheng Xu. Deep relative tracking. *IEEE Transactions on Image Processing*, 26(4):1845–1858, 2017.

[79] Bohyung Han, Jack Sim, and Hartwig Adam. Branchout: Regularization for online ensemble tracking with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3356–3365, 2017.

[80] Le Zhang, Jagannadan Varadarajan, Ponnuthurai Nagaratnam Suganthan, Narendra Ahuja, and Pierre Moulin. Robust visual tracking using oblique random forests. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5589–5598, 2017.

[81] Vijay K Sharma and Kamala K Mahapatra. Mil based visual object tracking with kernel and scale adaptation. *Signal Processing: Image Communication*, 53:51–64, 2017.

[82] Zhenjie Wang, Lijia Wang, and Hua Zhang. Patch based multiple instance learning algorithm for object tracking. *Computational intelligence and neuroscience*, 2017, 2017.

[83] Honghong Yang, Shiru Qu, and Zunxin Zheng. Visual tracking via online discriminative multiple instance metric learning. *Multimedia Tools and Applications*, 77(4):4113–4131, 2018.

[84] Chao Xu, Wenyuan Tao, Zhaopeng Meng, and Zhiyong Feng. Robust visual tracking via online multiple instance learning with fisher information. *Pattern Recognition*, 48 (12):3917–3926, 2015.

[85] Kai Chen and Wenbing Tao. Once for all: a two-flow convolutional neural network for visual tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 28 (12):3377–3386, 2017.

[86] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 fps with deep regression networks. In *European conference on computer vision*, pages 749–765. Springer, 2016.

[87] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1420–1429, 2016.

[88] Xiao Wang, Chenglong Li, Bin Luo, and Jin Tang. Sint++: Robust visual tracking via adversarial positive instance generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4864–4873, 2018.

[89] Jun Wang, Weibin Liu, Weiwei Xing, and Shunli Zhang. Two-level superpixel and feedback based visual object tracking. *Neurocomputing*, 267:581–596, 2017.

[90] Lijun Wang, Huchuan Lu, and Ming-Hsuan Yang. Constrained superpixel tracking. *IEEE transactions on cybernetics*, 48(3):1030–1041, 2017.

[91] Fan Yang, Huchuan Lu, and Ming-Hsuan Yang. Robust superpixel tracking. *IEEE Transactions on Image Processing*, 23(4):1639–1651, 2014.

[92] Dawei Du, Honggang Qi, Wenbo Li, Longyin Wen, Qingming Huang, and Siwei Lyu. Online deformable object tracking based on structure-aware hyper-graph. *IEEE Transactions on Image Processing*, 25(8):3572–3584, 2016.

[93] Tao Wang and Haibin Ling. Gracker: A graph-based planar object tracker. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1494–1501, 2017.

[94] Donghun Yeo, Jeany Son, Bohyung Han, and Joon Hee Han. Superpixel-based tracking-by-segmentation using markov chains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1812–1821, 2017.

[95] Jie Guo, Tingfa Xu, Ziyi Shen, and Guokai Shi. Visual tracking via sparse representation with reliable structure constraint. *IEEE Signal Processing Letters*, 24(2):146–150, 2016.

[96] Yang Yi, Yang Cheng, and Chuping Xu. Visual tracking based on hierarchical framework and sparse representation. *Multimedia Tools and Applications*, 77(13):16267–16289, 2018.

[97] Tianzhu Zhang, Si Liu, Changsheng Xu, Shuicheng Yan, Bernard Ghanem, Narendra Ahuja, and Ming-Hsuan Yang. Structural sparse tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 150–158, 2015.

[98] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Robust structural sparse tracking. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):473–486, 2018.

[99] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[100] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson WH Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8990–8999, 2018.

[101] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming-Hsuan Yang. Long-term correlation tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5388–5396, 2015.

[102] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Adaptive correlation filters with long-term and short-term memory for object tracking. *International Journal of Computer Vision*, 126(8):771–796, 2018.

[103] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79(41):30509–30555, 2020.

[104] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2):224–241, 2011.

[105] Federico Angelini, Zeyu Fu, Sergio A Velastin, Jonathon A Chambers, and Syed Mohsen Naqvi. 3d-hog embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4219–4223. IEEE, 2018.

[106] Luis González, Sergio A Velastin, and Gonzalo Acuna. Silhouette-based human action recognition with a multi-class support vector machine. 2018.

[107] Unaiza Ahsan, Chen Sun, and Irfan Essa. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. *arXiv preprint arXiv:1801.07230*, 2018.

[108] Earnest Paul Ijjina and Krishna Mohan Chalavadi. Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72: 504–516, 2017.

[109] Enjie Ghorbel, Rémi Boutteau, Jacques Boonaert, Xavier Savatier, and Stéphane Lecoeuche. Kinematic spline curves: A temporal invariant descriptor for fast action recognition. *Image and Vision Computing*, 77:60–71, 2018.

[110] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. Spatio-temporal attention-based lstm networks for 3d action recognition and detection. *IEEE Transactions on image processing*, 27(7):3459–3471, 2018.

[111] Shugao Ma, Jianming Zhang, Stan Sclaroff, Nazli Ikizler-Cinbis, and Leonid Sigal. Space-time tree ensemble for action recognition and localization. *International Journal of Computer Vision*, 126(2):314–332, 2018.

[112] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. Learning a deep model for human action recognition from novel viewpoints. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):667–681, 2017.

[113] Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 612–619, 2014.

[114] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.

[115] Effrosyni Mavroudi, Divya Bhaskara, Shahin Sefati, Haider Ali, and René Vidal. End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1558–1567. IEEE, 2018.

[116] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2017.

[117] Liangliang Wang, Lianzheng Ge, Ruifeng Li, and Yajun Fang. Three-stream cnns for action recognition. *Pattern Recognition Letters*, 92:33–40, 2017.

[118] A. Sharaf, M. Torki, M. E. Hussein, and M. El-Saban. Real-time multi-scale action detection from 3d skeleton data. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 998–1005, 2015. doi: 10.1109/WACV.2015.138.

[119] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[120] Wenkai Xu and Eung-Joo Lee. A novel method for hand posture recognition based on depth information descriptor. *KSII Transactions on Internet and Information Systems (TIIS)*, 9(2):763–774, 2015.

[121] Nathan Inkawhich, Matthew Inkawhich, Yiran Chen, and Hai Li. Adversarial attacks for optical flow-based action recognition classifiers. *arXiv preprint arXiv:1811.11875*, 2018.

[122] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[123] Vinay Kumar, Ankur Chaturvedi, and Anjani Kumar Rai. A framework using multiple features to detect multi-view human activity. In *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, pages 26–27, 2018.

[124] Shujah Islam, Tehreem Qasim, Muhammad Yasir, Naeem Bhatti, Hasan Mahmood, and Muhammad Zia. Single-and two-person action recognition based on silhouette shape and optical point descriptors. *Signal, Image and Video Processing*, 12(5):853–860, 2018.

[125] Md Atiqur Rahman Ahad, Masud Ahmed, Anindya Das Antar, Yasushi Makihara, and Yasushi Yagi. Action recognition using kinematics posture feature on 3D skeleton joint locations. *Pattern Recognition Letters*, 145:216–224, May 2021. ISSN 01678655. doi: 10.1016/j.patrec.2021.02.013. URL `https://linkinghub.elsevier.com/retrieve/pii/S0167865521000751`.

[126] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. Video action detection by learning graph-based spatio-temporal interactions. *Computer Vision and Image Understanding*, 206:103187, May 2021. ISSN 10773142. doi: 10.1016/j.cviu.2021.103187. URL `https://linkinghub.elsevier.com/retrieve/pii/S107731422100031X`.

[127] Ming Zong, Ruili Wang, Xiubo Chen, Zhe Chen, and Yuanhao Gong. Motion saliency based multi-stream multiplier ResNets for action recognition. *Image and Vision Computing*, 107:104108, March 2021. ISSN 02628856. doi: 10.1016/j.imavis.2021.104108. URL `https://linkinghub.elsevier.com/retrieve/pii/S0262885621000135`.

[128] Guoli Yan, Michelle Hua, and Zichun Zhong. Multi-derivative physical and geometric convolutional embedding networks for skeleton-based action recognition. *Computer Aided Geometric Design*, 86:101964, March 2021. ISSN 01678396. doi: 10.1016/j.cagd.2021.101964. URL `https://linkinghub.elsevier.com/retrieve/pii/S0167839621000108`.

[129] Wei Peng. Tripool: Graph triplet pooling for 3D skeleton-based action recognition. *Pattern Recognition*, page 12, 2021.

[130] Fan Zhu and Ling Shao. Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision*, 109(1-2):42–59, 2014.

[131] Kaiping Xu, Zheng Qin, and Guolong Wang. Recognize human activities from multi-part missing videos. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[132] Li Liu, Ling Shao, Xuelong Li, and Ke Lu. Learning spatio-temporal representations for action recognition: A genetic programming approach. *IEEE transactions on cybernetics*, 46(1):158–170, 2015.

[133] Yongqiang Li, S Mohammad Mavadati, Mohammad H Mahoor, Yongping Zhao, and Qiang Ji. Measuring the intensity of spontaneous facial action units with dynamic bayesian network. *Pattern Recognition*, 48(11):3417–3427, 2015.

[134] L Xing and XIAO Qin-kun. Human action recognition using auto-encode and pnn neural network. *Software Guide*, 1(4):1608–01529, 2018.

[135] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[136] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40 (6):1510–1517, 2017.

[137] Joshua Gleason, Rajeev Ranjan, Steven Schwarcz, Carlos Castillo, Jun-Cheng Chen, and Rama Chellappa. A Proposal-Based Solution to Spatio-Temporal Action Detection in Untrimmed Videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 141–150, Waikoloa Village, HI, USA, January 2019. IEEE. ISBN 978-1-72811-975-5. doi: 10.1109/WACV.2019.00021.

[138] Chao Yeh Chen and Kristen Grauman. Efficient Activity Detection in Untrimmed Video with Max-Subgraph Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):908–921, May 2017. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2016.2564404.

[139] Xiao-Yu Zhang, Haichao Shi, Changsheng Li, Kai Zheng, Xiaobin Zhu, and Lixin Duan. Learning Transferable Self-Attentive Representations for Action Recognition in Untrimmed Videos with Weak Supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9227–9234, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33019227.

[140] Yeongtaek Song and Incheol Kim. DeepAct: A Deep Neural Network Model for Activity Detection in Untrimmed Videos. *Journal of Information Processing Systems*, 14(1):150–161, February 2018. doi: 10.3745/JIPS.04.0059.

[141] Joshua Gleason, Carlos D. Castillo, and Rama Chellappa. Real-time Detection of Activities in Untrimmed Videos. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 117–125, Snowmass Village, CO, USA, March 2020. IEEE. ISBN 978-1-72817-162-3. doi: 10.1109/WACVW50321.2020.9096937.

[142] Joshua Gleason, Steven Schwarcz, Rajeev Ranjan, Carlos D. Castillo, Jun-Cheng Chen, and Rama Chellappa. Activity Detection in Untrimmed Videos Using Chunk-based Classifiers. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 107–116, Snowmass Village, CO, USA, March 2020. IEEE. ISBN 978-1-72817-162-3. doi: 10.1109/WACVW50321.2020.9096912.

[143] Davide Moltisanti, Sanja Fidler, and Dima Damen. Action Recognition From Single Timestamp Supervision in Untrimmed Videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9907–9916, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.01015.

[144] Nannan Li, Hui-Wen Guo, Yang Zhao, Thomas Li, and Ge Li. Active Temporal Action Detection in Untrimmed Videos Via Deep Reinforcement Learning. *IEEE Access*, 6: 59126–59140, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2872759.

[145] Md Atiqur Rahman and Robert Laganiere. Single-Stage End-to-End Temporal Activity Detection in Untrimmed Videos. In *2020 17th Conference on Computer and Robot Vision (CRV)*, pages 206–213, Ottawa, ON, Canada, May 2020. IEEE. ISBN 978-1-72819-891-0. doi: 10.1109/CRV50864.2020.00035.

[146] Tahmida Mahmud, Mahmudul Hasan, and Amit K. Roy-Chowdhury. Joint Prediction of Activity Labels and Starting Times in Untrimmed Videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5784–5793, Venice, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.616.

[147] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing Fine-Grained and Composite Activities Using Hand-Centric Features and Script Data. *International Journal of Computer Vision*, 119(3):346–373, September 2016. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-015-0851-8.

[148] Sari Awwad and Massimo Piccardi. Local depth patterns for fine-grained activity recognition in depth videos. In *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6, Palmerston North, New Zealand, November 2016. IEEE. ISBN 978-1-5090-2748-4. doi: 10.1109/IVCNZ.2016.7804453.

[149] Sari Awwad, Fairouz Hussein, and Massimo Piccardi. Local depth patterns for tracking in depth videos. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1115–1118, 2015.

[150] Aj Piergiovanni and Michael S. Ryoo. Fine-Grained Activity Recognition in Baseball Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1821–18218, Salt Lake City, UT, USA, June 2018. IEEE. ISBN 978-1-5386-6100-0. doi: 10.1109/CVPRW.2018.00226.

[151] Farnoosh Heidarivincheh, Majid Mirmehdi, and Dima Damen. Weakly-Supervised Completion Moment Detection using Temporal Attention. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1188–1196, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72815-023-9. doi: 10.1109/ICCVW.2019.00150.

[152] Xiao-Yu Zhang, Haichao Shi, Changsheng Li, and Peng Li. Multi-Instance Multi-Label Action Recognition and Localization Based on Spatio-Temporal Pre-Trimming for Untrimmed Videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12886–12893, April 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i07.6986.

[153] Xiao-Yu Zhang, Haichao Shi, Changsheng Li, Peng Li, Zekun Li, and Peng Ren. Weakly-supervised action localization via embedding-modeling iterative optimization. *Pattern Recognition*, 113:107831, May 2021. ISSN 00313203. doi: 10.1016/j.patcog.2021.107831.

[154] Ardhendu Behera, Alexander Keidel, and Bappaditya Debnath. Context-driven multi-stream lstm (m-lstm) for recognizing fine-grained activity of drivers. In *German Conference on Pattern Recognition*, pages 298–314. Springer, 2018.

[155] Sathyanarayanan Aakur, Daniel Sawyer, and Sudeep Sarkar. Fine-grained Action Detection in Untrimmed Surveillance Videos. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 38–40, Waikoloa Village, HI, USA, January 2019. IEEE. ISBN 978-1-72811-392-0. doi: 10.1109/WACVW.2019.00014.

[156] Michelle E. Mlinac and Michelle C. Feng. Assessment of Activities of Daily Living, Self-Care, and Independence. *Archives of Clinical Neuropsychology*, 31(6):506–516, 08 2016. ISSN 0887-6177. doi: 10.1093/arclin/acw049.

[157] Chien-Wen Cho, Wen-Hung Chao, Sheng-Huang Lin, and You-Yin Chen. A vision-based analysis system for gait recognition in patients with Parkinson's disease. *Expert Systems with Applications*, 36(3):7033–7039, April 2009. ISSN 09574174. doi: 10.1016/j.eswa.2008.08.076.

[158] Qiannan Li, Yafang Wang, Andrei Sharf, Ya Cao, Changhe Tu, Baoquan Chen, and Shengyuan Yu. Classification of gait anomalies from kinect. *The Visual Computer*, 34(2):229–241, February 2018. ISSN 0178-2789, 1432-2315. doi: 10.1007/s00371-016-1330-0.

[159] Trong-Nguyen Nguyen, Huu-Hung Huynh, and Jean Meunier. Skeleton-Based Abnormal Gait Detection. *Sensors*, 16(11):1792, October 2016. ISSN 1424-8220. doi: 10.3390/s16111792.

[160] Margarita Khokhlova, Cyrille Migniot, Alexey Morozov, Olga Sushkova, and Albert Dipanda. Normal and pathological gait classification LSTM model. *Artificial Intelligence in Medicine*, 94:54–66, March 2019. ISSN 09333657. doi: 10.1016/j.artmed.2018.12.007.

[161] Haiping Lu, Yaozhang Pan, Bappaditya Mandal, How-Lung Eng, Cuntai Guan, and Derrick W. S. Chan. Quantifying Limb Movements in Epileptic Seizures Through Color-Based Video Analysis. *IEEE Transactions on Biomedical Engineering*, 60(2):461–469, February 2013. ISSN 0018-9294, 1558-2531. doi: 10.1109/TBME.2012.2228649.

[162] David Ahmedt-Aristizabal, Clinton Fookes, Simon Denman, Kien Nguyen, Tharindu Fernando, Sridha Sridharan, and Sasha Dionisio. A hierarchical multimodal system for motion analysis in patients with epilepsy. *Epilepsy & Behavior*, 87:46–58, October 2018. ISSN 15255050. doi: 10.1016/j.yebeh.2018.07.028.

[163] David Ahmedt-Aristizabal, Simon Denman, Kien Nguyen, Sridha Sridharan, Sasha Dionisio, and Clinton Fookes. Understanding Patients' Behavior: Vision-Based Analysis of Seizure Disorders. *IEEE Journal of Biomedical and Health Informatics*, 23(6): 2583–2591, November 2019. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2019.2895855.

[164] Tamas Karacsony, Anna Mira Loesch-Biffar, Christian Vollmar, Soheyl Noachtar, and Joao Paulo Silva Cunha. A Deep Learning Architecture for Epileptic Seizure Classification Based on Object and Action Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4117–4121, Barcelona, Spain, May 2020. IEEE. ISBN 978-1-5090-6631-5. doi: 10.1109/ICASSP40776.2020.9054649.

[165] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854, Providence, RI, June 2012. IEEE. ISBN 978-1-4673-1228-8 978-1-4673-1226-4 978-1-4673-1227-1. doi: 10.1109/CVPR.2012.6248010.

[166] Amr Elkholy, Mohamed E. Hussein, Walid Gomaa, Dima Damen, and Emmanuel Saba. Efficient and Robust Skeleton-Based Quality Assessment and Abnormality Detection in Human Action Performance. *IEEE Journal of Biomedical and Health Informatics*, 24 (1):280–291, January 2020. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2019.2904321.

[167] Nigar Şen Köktaş, Neşe Yalabik, Güneş Yavuzer, and Robert P.W. Duin. A multi-classifier for grading knee osteoarthritis using gait analysis. *Pattern Recognition Letters*, 31(9):898–904, July 2010. ISSN 01678655. doi: 10.1016/j.patrec.2010.01.003.

[168] Can Tunca, Nezihe Pehlivan, Nağme Ak, Bert Arnrich, Gülüstü Salur, and Cem Ersoy. Inertial Sensor-Based Robust Gait Analysis in Non-Hospital Settings for Neurological Disorders. *Sensors*, 17(4):825, April 2017. ISSN 1424-8220. doi: 10.3390/s17040825.

[169] Wei-Chun Hsu, Tommy Sugiarto, Yi-Jia Lin, Fu-Chi Yang, Zheng-Yi Lin, Chi-Tien Sun, Chun-Lung Hsu, and Kuan-Nien Chou. Multiple-Wearable-Sensor-Based Gait Classification and Analysis in Patients with Neurological Disorders. *Sensors*, 18(10):3397, October 2018. ISSN 1424-8220. doi: 10.3390/s18103397.

[170] NH Nordin, Asan Muthalif, and M. Razali. Control of transtibial prosthetic limb with magnetorheological fluid damper by using a fuzzy pid controller. *Journal of Low Frequency Noise, Vibration and Active Control*, 37:146134841876617, 04 2018. doi: 10.1177/1461348418766171.

[171] Eri Ishikawa, Stephen Karungaru, and Kenji Terada. Gait features extraction method using image processing. In *2011 17th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, pages 1–6, Ulsan, Korea (South), February 2011. IEEE. ISBN 978-1-61284-677-4. doi: 10.1109/FCV.2011.5739724.

[172] Ana Patricia Rocha, Hugo Choupina, Jose Maria Fernandes, Maria Jose Rosas, Rui Vaz, and Joao Paulo Silva Cunha. Parkinson's disease assessment based on gait analysis using an innovative RGB-D camera system. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3126–3129, Chicago, IL, August 2014. IEEE. ISBN 978-1-4244-7929-0. doi: 10.1109/EMBC.2014.6944285.

[173] Joao Paulo Silva Cunha, Ana Patricia Rocha, Hugo Miguel Pereira Choupina, Jose Maria Fernandes, Maria Jose Rosas, Rui Vaz, Felix Achilles, Anna Mira Loesch, Christian Vollmar, Elisabeth Hartl, and Soheyl Noachtar. A novel portable, low-cost kinect-based system for motion analysis in neurological diseases. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2339–2342, Orlando, FL, August 2016. IEEE. ISBN 978-1-4577-0220-4. doi: 10.1109/EMBC.2016.7591199.

[174] Hans Kainz, Christopher P. Carty, Luca Modenese, Roslyn N. Boyd, and David G. Lloyd. Estimation of the hip joint centre in human motion analysis: A systematic review. *Clinical Biomechanics*, 30(4):319–329, May 2015. ISSN 02680033. doi: 10.1016/j.clinbiomech.2015.02.005.

[175] Joana Rodrigues, Paulo Maia, Hugo Miguel Pereira Choupina, and Joao Paulo Silva Cunha. On the Fly Reporting of Human Body Movement based on Kinect v2. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1546–1549, Honolulu, HI, July 2018. IEEE. ISBN 978-1-5386-3646-6. doi: 10.1109/EMBC.2018.8512454.

[176] Zelun Luo, Jun-Ting Hsieh, Niranjan Balachandar, Serena Yeung, Guido Pusiol, Jay Luxenberg, Grace Li, Li-Jia Li, N Lance Downing, Arnold Milstein, et al. Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring. *Machine Learning for Healthcare (MLHC)*, 2, 2018.

[177] Jonathan Feng-Shun Lin and Dana Kulic. Online Segmentation of Human Motion for Automated Rehabilitation Exercise Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):168–180, January 2014. ISSN 1534-4320, 1558-0210. doi: 10.1109/TNSRE.2013.2259640.

[178] David Xue, Anin Sayana, Evan Darke, Kelly Shen, Jun-Ting Hsieh, Zelun Luo, Li-Jia Li, N. Lance Downing, Arnold Milstein, and Li Fei-Fei. Vision-Based Gait Analysis for Senior Care. *arXiv:1812.00169 [cs]*, December 2018. URL `http://arxiv.org/abs/1812.00169`. arXiv: 1812.00169.