THE UNIVERSITY *of* EDINBURGH

Visual Assessment of Changes in Activity Levels while Eating

Muhammad Ahmed Raza

Doctor of Philosophy

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

2024

Abstract

This doctoral thesis investigates non-intrusive ways to monitor the changes in activity levels while eating. Elderly individuals encounter a variety of health challenges including sarcopenia which causes a decrease in the number and size of the muscle fibre resulting from normal aging. Monitoring activity levels in the elderly is particularly crucial for healthy aging in place, as it helps detect these health-related issues early and supports independent living.

Past research provides valuable insights into the activities of daily living and various behavioral aspects of the elderly. These are mainly full-body or gait analysis-based approaches. However, to the best of our knowledge, there is a gap in research about vision-based monitoring systems to strictly analyze upper-body motion and capture valuable insights into their performance while the subjects eat in a dining room environment. We aim to address this gap in this thesis. The primary objectives of this research are to establish reliable methods to (1) monitor the eating behavior of the elderly, (2) develop a generalizable model across various subjects to minimize subjective bias, and (3) develop a complete upper-body focused pipeline that generates health statistics and gather insights into both eating behavior and musculoskeletal deterioration over time.

For this purpose, our first major contribution presents a dataset called EatSense using realsense RGB-D to monitor eating activity in a dining room environment. It comprises 135 video sequences of 27 subjects from 13 nationalities, recorded using an RGB-D camera in an uncontrolled setting, with an average duration of 11.5 minutes per video. The dataset features dense frame-wise labels for 16 atomic eating-related actions, with an average of 114.1 actions per video sequence, and provides three levels of label abstraction. EatSense uniquely focuses on upper-body posture and movements, including scenarios with and without wrist weights to simulate changes in motor function.

Two minor contributions following the EatSense dataset are: (1) Evaluation and behavioral assessment using the EatSense dataset with several action recognition and temporal action localization approaches. This study concludes that EatSense is a challenging dataset for both action recognition and temporal action localization algorithms since it has varying lengths of action instances. (2) Explore the impact of face obfuscation methods on pose-based action recognition in healthcare monitoring. The study was limited as it was only evaluated on the EatSense dataset and concluded that face

obfuscation strategies that pseudonymize facial features can preserve privacy without significantly degrading the performance of the subsequent tasks.

The second major contribution presents a vision-based approach to assess performance levels while eating, intending to monitor potential performance decline in elderly individuals. We used weights attached to the subjects' wrists to simulate mobility or motor function changes. The study compares hand-crafted feature-based regression methods (Gaussian Mixture Regression, Multilayer Perceptron, and LightGBM) against deep feature-based regression using ST-GCN. Results show that Gaussian Mixture Regression performs slightly better in predicting the degree of performance decline (i.e., weight level) across subjects.

Lastly, our third major contribution presents a comprehensive, fully autonomous vision-based pipeline for monitoring eating activities and assessing musculoskeletal health. The pipeline's key contributions include a multi-purpose video-to-report framework for long-term monitoring, improved action localization in continuous video through relaxed data augmentation and output merging techniques, and the ability to capture trends and generate insights on changes in eating behavior and upper-body movements.

Lay summary

This PhD thesis investigates methods for monitoring the activity levels of the elderly. People experience health problems as they age, which causes their muscle mass and strength to decline. It is essential to monitor their activities to identify these issues early on and assist with their independent living. In the past, studies to better understand the daily activities and behaviors of the elderly have concentrated on full-body or walking patterns. On the other hand, little research has been done on the specific use of computer vision systems to track upper-body movements, particularly during eating.

In order to fill this gap, this study presents various original contributions. One of them is a dataset named EatSense, which contains 27 individuals from 13 different countries while they eat in a dining room. Each recording lasts about 11.5 minutes on average. The dataset includes detailed labels for 16 different eating-related actions, like bringing food to the mouth, and looks at how upper-body movements change, with added wrist weights to simulate muscle function changes.

Additionally, the study assesses various methods for identifying these eating actions from the EatSense videos. It concludes that because of the different action durations (i.e., varying between 0.62 to 9 seconds), the dataset presents a challenge for the existing video understanding methods. The study also examines how to keep people's identities hidden from view in the videos while still maintaining the ability to identify what they are doing. Another section of the study examines how well various approaches—such as simulating deterioration with wrist weights—can predict changes in motor function.

Lastly, a fully automated system that monitors eating activities and evaluates muscle health using video is presented in this research. This system can generate reports over time and capture trends in eating behavior and upper-body movements, providing valuable insights into the health of elderly individuals.

Acknowledgements

All praise is due to Allah, who has given me the strength, knowledge, and perseverance to complete this thesis.

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Robert B. Fisher, for his continuous support, patience, and immense knowledge. His support, especially during the tough times, provided me with the strength and perseverance needed to continue. I could not have imagined having a better advisor and mentor for my study.

I would especially like to thank Longfei Chen whose support and encouragement were crucial in the research. I would also like to thank my peers, Nanbo Li, Christopher Lochhead, Sonali Chawla, Minseong Kim, Taichi Hosoi, Hanz Cuevas-Valesques, and Zhaole Sun for their invaluable discussions and feedback. Their friendship and encouragement kept me motivated and focused throughout this journey. I am grateful to Xiaoyan Ma (Amy) and Tiago Lé at Danu Robotics for their understanding and support, and for allowing me to balance my work commitments with my academic pursuits. Their support played a significant role in keeping me sane and completing this thesis. A special thanks go to the wonderful people I met in Japan, my friends here in Edinburgh, and back home in Pakistan, for always being there to lend an ear and offer words of comfort, which has meant the world to me. I wish them all the success and a bright future in their endeavors.

I want to thank the School of Informatics at The University of Edinburgh and the Higher Education Commission (HEC) Pakistan for providing a PhD scholarship, the Advanced Care Research Center (ACRC), and the Institute of Perception Action and Behavior (IPAB) for research support throughout my PhD. I must also acknowledge Prof. Ijaz Mansoor Qureshi for always believing in me and pushing me to pursue a PhD.

Last but certainly not least, I would like to thank my family for their unconditional love, sacrifices, and belief in me. Their support and encouragement have been the bedrock of my life, and their unwavering faith in my abilities has been a source of strength throughout my academic journey. I hereby dedicate this thesis to them, even though it is nothing compared to what they have provided.

To all of you, I extend my deepest gratitude. Cheers!

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. The work published within this thesis has been published in the following peer-reviewed articles with attribution and contribution as follows:

> **M. A. Raza**, L. Chen, L. Nanbo, and R. B. Fisher, "Eatsense: human-centric, action recognition and localization dataset for understanding eating behaviors and quality of motion assessment," Image and Vision Computing, vol. 137, p. 104762, 2023.
>
> **M. A. Raza**, and and R. B. Fisher, "Vision-based approach to assess performance levels while eating." Machine Vision and Applications 34.6 (2023): 124.
>
> **M. A. Raza**, C. Lochhead, and R. B. Fisher, "Effect of face obfuscation methods on pose-based action recognition." International Conference on AI in Healthcare, 2024.
>
> **M. A. Raza**, and R. B. Fisher, "V2R: A Fully Autonomous Vision-Based System for Analyzing Eating Behaviors and Musculoskeletal Deterioration". [Submitted]

Contribution:

- Raza: proposed and implemented ideas and theories, conducted experiments, and wrote papers.
- Chen, Nanbo: helped with dataset creation and labeling.
- Lochhead: helped with writing the introduction and background review.
- Fisher: supervised the writing, and helped with verification of the ideas, theories, and experiments.

(*Muhammad Ahmed Raza*)

Table of Contents

Chapter 1

Introduction

According to an article by World Health Organization, “Ageing and health” [1] (October 2022), one in every six people will be aged over 60 by 2030. Also, the number of people aged over 80 is expected to triple between 2020 to 2050. With an aging population that needs assistance and care, many countries are experiencing a severe shortage of doctors and trained care staff. With this escalating demand, healthcare systems face significant pressure, emphasizing the urgency to automate these systems. Additionally, the condition of many care homes is dismal, prompting elderly individuals to favor remaining in their residences.

Elderly individuals often encounter various health challenges, including pathological conditions such as osteoporosis, characterized by porous and weakened bones that can affect their gait. Additionally, conditions like vasovagal syncope, marked by a sudden drop in heart rate and blood pressure, may lead to fainting and falls. Moreover, these individuals frequently develop one or more chronic conditions over time, such as arthritis, heart disease, or dementia. In addition to these diseases, elderly individuals face several other challenges in their daily lives that, if addressed promptly, can prove to be very beneficial.

This chapter provides an overview of the thesis and its structure as follows: First, the problems related to the current healthcare monitoring systems/algorithms are presented. Second, the original contributions to solving each problem are stated. Finally, the overall structure of the thesis is presented.

1.1 Problem Statement

Elderly people need constant care and regular check-ups of their health and motor movements. Irregular check-ups and motor movement analyses do not completely monitor the health status of an individual well enough. Home-based motion analysis systems might be a good solution; however, they typically utilize wearable sensors to detect falls or carry out motion analysis. These wearable sensor-based devices hold significant potential for providing accurate results. But, there are drawbacks to this approach. Elderly individuals often experience memory lapses, and there may be a reluctance to adopt wearable technological aids [2], [3]. Consequently, this solution may not be feasible for most individuals. An alternative solution, also the motivation for this thesis, is to develop a camera-based surveillance system. Such a system could monitor an individual's daily routine and identify anomalies or gradual changes in behavior without requiring the individual to wear any additional devices.

The first step in developing such a system is to have the necessary data. For motion analysis and tracking people's daily activities, a multitude of large-scale publicly accessible datasets are available. These datasets offer performance benchmarks for many widely used algorithms and include a variety of action classes. Despite having a large total number of recording hours and a wide range of challenging scenarios, the existing datasets still lack high-quality data [4] and hence are still unable to accurately represent a particular behavior completely and help in developing sustainable and public systems [5] or have the capability to help identify a change in motion that would indicate a decline in the subject's motor movement [6].

Secondly, for understanding an individual's behavior, sub-action recognition can help analyze complex activities by breaking them down into smaller bits. Current motion analysis systems use deep-learning-based sub-action recognition networks with convoluted features that do not help healthcare workers understand the underlying root cause of the problem. For a better understanding, hand-crafted features (domain-specific engineered features) for motion analysis should be explored. These hand-crafted features are essentially the parameters of musculoskeletal movement over time that can help us comprehend the behavioral traits of individuals in a better way.

However, these sub-action recognition algorithms do not utilize full-length untrimmed videos which can provide a deeper understanding of behavior. For example, if we can identify chewing in a video sequence, we can estimate the time taken for chewing, which in turn can help us see mindless eating that can have negative ef-

fects on the health in long term. For this purpose, using a temporal action localization network that outputs temporal segments and their action class can help understand behavior within a context as they exploit inter-action temporal relationships.

Another major hurdle in developing an efficient motion analysis system is developing a generalized model that works for a large number of human subjects. Studies that involve humans are difficult because indicative features that parameterize performance changes differ greatly between individuals. To forecast performance as weights are increased, a motion model with a common set of parameters should ideally already exist. People appear to respond to the weights attached to the wrists in different ways; some appear to slouch more, while others alter their posture (dropped shoulders, etc.). This variation across subjects makes attaining a generalized model a nearly impossible task.

There has been much research done to explore non-clinical fully autonomous motion analysis systems. However, they lack several aspects: (1) they rely on deep features, (2) they explore behavioral and motion changes in a limited capacity, (3) they are hyper-focused on full/lower body (gait) analysis and utilize subjects with Parkinson's to evaluate their methodologies and (4) they have serious unaddressed ethical concerns about privacy and data security.

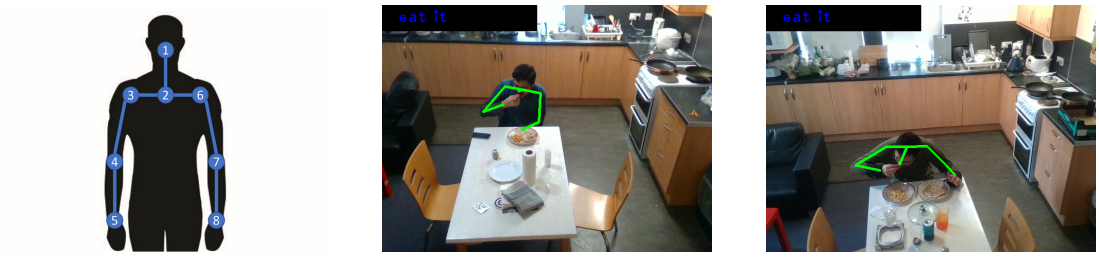
Developing a video-to-report pipeline for long-term eating behavior and musculoskeletal performance decline monitoring comprises several steps. Most importantly, it requires a Temporal Action Localization network that could accurately estimate the time segment of when an action occurred in the video. Most of the publicly available datasets contain sparse labels in the videos, which makes it easier for current temporal action localization networks to work well. However, longer videos, shorter actions, and dense labels cause the performance of these localization networks to drop [7]. Therefore, another problem is how to develop temporal action localization networks with improved performance on atomic actions that last less than a second.

1.2 Thesis Structure

This thesis is structured around 4 main chapters (chapters 3,4,5 and 6) where each extends a part of four (three published and one submitted) papers. In these four chapters we will discuss the major components required to build models for visual assessment of activity levels in the elderly:

- **Chapter 3: presents a dataset called EatSense tailored for healthcare appli-**

cations requiring long-term monitoring. It enables tracking subtle changes in movements over time for healthcare purposes and provides a comprehensive model of specific human behaviors, encompassing multiple sub-actions. In this chapter, we extensively discuss where EatSense exists among the current computer vision and health-care datasets, the semi-automatic labeling strategy used for annotations, and face obfuscation methods used to protect the privacy of the subjects in EatSense. The key features of EatSense are the introduction of challenging atomic actions for action recognition and temporal action localization frameworks, the capability to model comprehensive eating behavior in terms of a sequence of action-based behaviors, and the simulation of minor variations in motion or performance by attaching weights to the wrists of the subjects. Fig. 1.1 shows the setting of our Intel RealSense D415 system in the dining room environment. The figure also shows an image sample from our proposed dataset both without (middle figure) and with wrist weights (right figure). This work was partly adapted from the following two published papers [8], [9] with improved discussions:



(a) (b) (c)

Figure 1.1: The view from the Intel RealSene D415 in one of the dining room environments. (a) shows the eight upper-body joints 1) nose (n), 2) chest, 3) right-shoulder, 4) right-elbow, 5) right-wrist, 6) left-shoulder, 7) left-elbow, and 8) left-wrist. (b) shows the subject performing 'eat it' without weights. (c) shows the subject is performing 'eat it' with weights

Raza, Muhammad Ahmed, et al. 'EatSense: human-centric, action recognition and localization dataset for understanding eating behaviors and quality of motion assessment.' Image and Vision Computing 137 (2023): 104762.

Raza, Muhammad Ahmed, and Robert B. Fisher. 'Effect of Face Obfuscation Methods on Pose-Based Action Recognition.' International Conference on AI in Healthcare (2024).

- **Chapter 4: explores the crucial step for understanding eating behavior from videos of individuals by using two classes of action detection algorithms: (sub)-action recognition and temporal action localization (TAL).** We start by using convolution neural network (CNN)-based features but also shed light on the possibility of using domain-specific hand-crafted features for the sub-action recognition. The main motivation behind this analysis is to explore research questions such as: 1) What are some potential parameters/features (engineered using domain knowledge to emphasize explainable models) of human motion that help with sub-action recognition? 2) How do explainable features perform compared to deep-learning-based approaches? Lastly, we also explore TAL frameworks with EatSense to get useful information such as the start and end of an action, and answer the research questions: 1) How well are TAL networks able to localize actions in a densely annotated dataset? 2) How accurately do TAL networks work with hand-crafted features as compared to deep-learning-based features? This work was extended from the following published paper [8] with improved discussions and detailed TAL analysis:

  > Raza, Muhammad Ahmed, et al. 'EatSense: human-centric, action recognition and localization dataset for understanding eating behaviors and quality of motion assessment.' Image and Vision Computing 137 (2023): 104762.

- **Chapter 5: presents a two-step vision-based pipeline aimed at attaining a generalized model by exploring the generalization ability of existing models, using the dataset EatSense, introduced in this thesis,** for assessing the changes in the performance levels of the subjects. Several existing models do not work well on new unseen subjects that they were not trained on, since the features that indicate decay vary from person to person in EatSense. To address this, we propose a pipeline based on a robust feature selection procedure for hand-crafted features extracted using EatSense to explore the generalizability of the existing models. The pipeline involves two steps: 1) estimate the best distinctive features across all subjects (every feature is estimated from the subject's 3D positions) and 2) apply an uncertainty-aware regression model to tackle the issue. The output achieved using the proposed pipeline with Gaussain Mixture Regression (GMR) using leave-one-out training strategy is shown in Fig. 1.2. This shows that the solid line (predicted weights) is aligned to the dashed line (perfect correlation), hence demonstrating the effectiveness of the proposed

technique. We refer the readers to chapter 5 for comparison against other regression algorithms and a more detailed discussion. This work was adapted from the following published paper [10] with improved discussions:

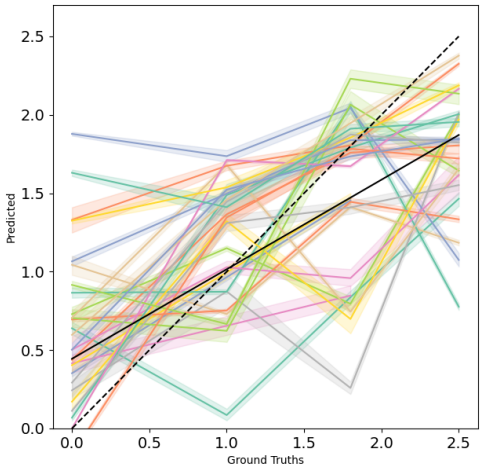


Figure 1.2: The plot shows the expected weight against the ground truth weight. The solid line represents the RANSAC-based least square fit of the data, while the dashed black line shows the perfect correlation. Every colored curve represents the outcome of a separate leave-one-out model iterated over 27 subjects. The solid-colored curves show the average of these predictions because each micromovement in the test set consists of multiple frames or clips, and the shading around each curve shows the range of one standard deviation.

> Raza, Muhammad Ahmed, and Robert B. Fisher. 'Vision-based approach to assess performance levels while eating.' Machine Vision and Applications 34.6 (2023): 124.

- **Chapter 6: presents a vision-based pipeline that aims to monitor eating activities with the objective of drawing insights into eating behaviors and musculoskeletal health.** The pipeline consists of three primary components: data recording and pre-processing, activity classification and localization, and post-processing for data analysis, including posture anomaly detection and mouthful counting. We demonstrate the effectiveness of our system with several experiments, both individual component-wise and holistic tests. This work was adapted from the following under-review paper with improved discussions:

Raza, Muhammad Ahmed, and Robert B. Fisher. 'V2R: A Fully Autonomous Vision-Based System for Analyzing Eating Behaviors and Musculoskeletal Deterioration'. [**Submitted**]

Chapter 2

Literature Review

This chapter explores the relevant background for human motion analysis for healthcare in this thesis, such as **aging, impact on health services and AI**, motion capture systems, and **current methods of assessing physical activities in the elderly**. However, the literature review on the techniques explored in this thesis to tackle the problems mentioned in 1.1 is discussed in detail in each of their respective chapters.

2.1 Aging, Impact on Health Services and AI

In Western society, nearly 30% of individuals over the age of 55 experience moderate to severe physical limitations [11]. These limitations elevate the risk of falls, institutionalization, co-morbidities, and premature death. A key factor behind these physical constraints is sarcopenia, the age-related loss of skeletal muscle mass. However, emerging evidence reveals that the decline in muscle mass is not the only factor affecting physical performance. The loss of muscle strength also plays a significant role in diminishing physical capabilities in the elderly. Extensive data suggests that motor coordination, excitation-contraction coupling, skeletal integrity, and other aspects of the nervous, muscular, and skeletal systems are crucial for maintaining physical performance in older adults [12]. These diverse factors influencing skeletal muscle performance in older adults are shown in Fig. 2.1.

Increased life expectancy often leads to living with disabilities and chronic conditions that can impede an individual's ability to carry out daily activities or sustain independent living [13]. As the aging population demands more personal attention, care and assistance, many countries face a significant shortage of care workers (e.g., home health aides). Informal caregivers worldwide, primarily women, frequently man-

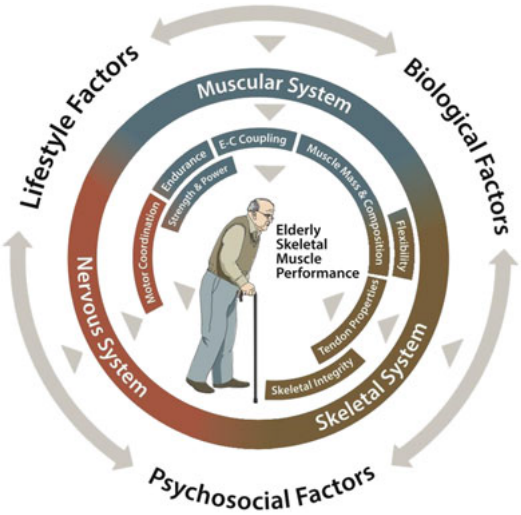


Figure 2.1: The multiple parameters of elderly skeletal muscle performance. Diagram by Tim Goheen taken from [12]

.

age other personal, family, and work responsibilities alongside providing continuous care and monitoring for their elderly relatives [14]. Due to evolving family dynamics, shrinking family sizes, increased workforce participation by women, and migration trends, the number of potential family caregivers per older adult is projected to decline sharply [15], [16].

Many older adults prefer to live independently or age in place, meaning they would rather stay in their own homes with appropriate support than move into institutional care [17]. Remote monitoring technologies, such as video cameras to observe activities at home, may help older adults live independently. However, these technologies still depend on human operators or family caregivers to watch video feeds in real time and respond based on their judgments. Consequently, they are labor-intensive and susceptible to human distractions and errors [18].

Advanced AI home-health monitoring systems, such as computer vision analytics, can classify activities like standing or walking and progressively learn what constitutes typical movements or activities for a specific older adult in their unique environment. Sensors placed in various locations within the home can track overall daily activity, time spent outside the home, walking speed, and specific locations within the home [19]. This tracking helps register the types, sequences, and duration of activities, identifying unusual movements and activities that may indicate cognitive and functional decline [20]. For example, AI monitoring programs that continuously analyze data

might detect something a human might overlook, such as an elderly individual taking increasingly longer time to balance while standing up [21].

When the automated intelligent system anticipates a decline in health from the gathered data, it can take action by sending alerts or recommendations to the elderly individual and/or their caregiver according to pre-set risk levels. By offering not only detailed real-time data but also automated notifications to ensure prompt safe care, these technologies can help avert severe health declines or injuries, potentially postponing or eliminating the necessity for expensive institutional care [22].

In summary, physical abilities decline with age due to a multitude of factors. This thesis aims to address this issue by proposing non-intrusive methods for monitoring upper-body health through the analysis of individuals' eating behaviors.

2.2 Motion Capture Systems

The term 'motion capture' (MoCap) has been defined differently by scholars based on their research areas, but it generally refers to recording the movement of objects or people. Various researchers [23, 24, 25] have identified two optical MoCap systems: marker-based and marker-less MoCap systems. Both systems have been widely used for various applications involving, rehabilitation, ergonomic risks of industrial workers, and health-care monitoring by capturing their body kinematics with smart cameras and converting the data into three-dimensional (3D) information. In this section, we present a brief overview of commonly used marker-based and marker-less MoCap systems in various applications.

2.2.1 Marker-based MoCap Systems

A marker-based motion capture (MoCap) system is a technology used to record and analyze the movement of objects or people using markers placed on the subject. Markers are reflective or LEDs attached to key points on the subject's body, such as joints and limbs. These markers serve as reference points for tracking movement. Generally, these systems consist of multiple cameras around the capture area to track the marker's positions [26]. These cameras are usually infrared to detect the reflective markers or designed to capture LED marker light.

In marker-based systems, Vicon is the most commonly used device. In [27] the researchers used a Vicon T-40 (refers to 4 mega-pixel cameras) and placed markers

on the hands of the subjects while they swam to study the feasibility of the measurement of the 3D hand kinematics. In [28] the researchers used an Opti-Track Flex3 to track the hands and head of surgeons during laparoscopic suturing to analyze the different experience levels of surgeons. To output marker information as XYZ data, the Opti-Track Flex3 employs six infrared cameras, 14 mm spherical retro-reflective markers, and a small-volume motion camera. In [29] the researchers used a Vicon MX13 (1.3-megapixel camera) and Xsens MTw to monitor the full human body to examine the feasibility and accuracy of a full-body magneto-inertial measurement units-based method for estimating the 3D body center of mass trajectory and energetics during walking.

Another type of marker-based system combines Vicon with another sensor such as an IMU. In [30] Lebleu et. al. used IMUs and a Vicon V5 (high-speed cameras) to validate the calculation of the joint angle precision and consistency using various sensor-to-segment calibrations. In [31] Lavender et. al. used an Opti-Track and EMG. The study evaluated the bio-mechanical efficacy of four devices that might be utilized by two-person teams (of firefighters) to raise patients at chair height, from a recliner chair, or the floor.

2.2.2 Marker-less MoCap Systems

Marker-less motion capture (MoCap) is a technology used to record and analyze the movement of objects or people without attaching physical markers to the subject. The marker-less system typically utilizes multiple cameras, or other types of imaging devices to capture the subject's movements. Most of these systems use cameras with depth sensors e.g., Microsoft Kinect [26]. Computer vision and machine learning techniques, such as human pose estimators and action recognition, are then applied to analyze the captured images or video streams.

2.2.2.1 Whole/Lower Body MoCap Systems

In [32], Chakraborty et. al. used a Kinect V2 to track the lower extremities to determine the stability and kinematics of joints of human gait. The researchers also used an OptoTrack to validate the results of Kinect-v2 to check for the accuracy of the marker-less motion capture against a marker-based motion capture system. In [33], the researchers used a Microsoft Kinect V2 camera and proposed a human joint tracking algorithm to assess Parkinsonian gait in older adults. Parrilla et. al. in [34] presented and used a

Move4D system to model the moving human body in 3D. Move 4D includes a collection of synchronized modules designed to scan body parts or entire bodies in motion, capturing textures, alongside a control unit and processing software. Each module is equipped with a pair of infrared (IR) cameras, an IR projector, a color (RGB) camera, and a processing unit.

2.2.2.2 Upper Body MoCap Systems

In [35, 25] the researchers compared the Kinect-V2 and Vicon marker-based systems (Vicon MX3 and Captiv L7000 respectively). Their focus was mainly on tracking and accurate estimation of joint angles of the upper body's motion. In [36] Liu et. al. estimated the upper-limb joint angles and analyzed the posture (during the task of trunk-twisting) of subjects using two Kinect-V2 cameras. More recently, in [37] Lam et. al. used two Ipad-Pros as a marker-less video recording system and validated the results with a Vicon system with a light detection and ranging scanner at two different angles to measure the active range of motion.

In summary, marker-based systems require physical markers and can only be used in a clinical or laboratory environment. On the other hand, markerless systems use cameras and other sensors for monitoring without placing any physical object on the subject. This makes it a more practical choice for monitoring activities in a home setting. Additionally, markerless systems have been validated in several studies when compared against marker-based systems and have shown similar performance [26]. For the research presented in this thesis, we focus on upper-body motion analysis in a dining room setting and hence we chose to use a marker-less system comprising of RealSense D-415 RGB-D camera to monitor activity levels in the elderly.

2.3 Pose Estimation

Markerless mocap systems rely on the algorithms followed by data recording, e.g., pose estimation or action recognition, etc. This thesis focuses on utilizing the motion of the upper body joint locations and hence also depends on accurate pose estimation. Our proposed dataset (discussed in chapter 3), was recorded in a real-world environment with the assumption that there is one person in the field of view, but recording in the dining room sometimes had instances of two or more people visible, so we dealt with this by using multi-person human pose estimators. This section briefly discusses the

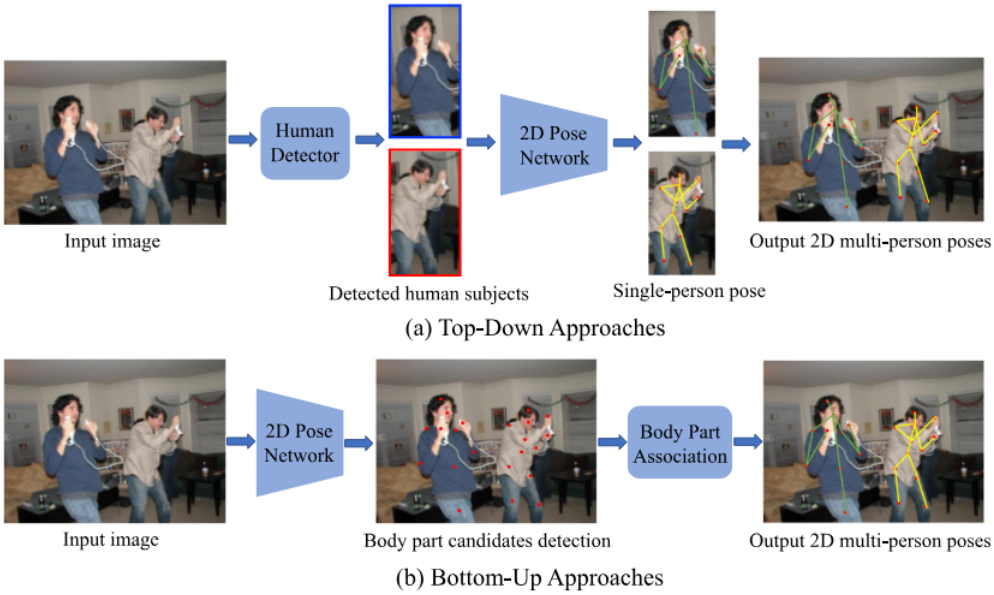


Figure 2.2: Comparison between top-down and bottom-up **2D** human pose estimators. Diagram copied from Zheng, Wu, *et al* [38].

most common and state-of-the-art multi-person 2D and 3D pose estimation algorithms.

2.3.1 2D Pose Estimation

2D human pose estimation techniques can be categorized into top-down and bottom-up approaches. Top-down approaches utilize pre-existing person detectors to extract bounding boxes (each representing an individual) from input images, followed by the application of single-person pose estimators to each bounding box to generate multi-person poses. In contrast, bottom-up approaches initially detect all body joints within an image and subsequently group these joints to form individual subjects.

As illustrated in Fig. 2.2(a), top-down approaches typically consist of two stages: a person bounding box derived from a human body detector and a single-person pose estimator used to predict keypoint locations within these bounding boxes. Cai et al. [39] introduced a multi-stage network with a Pose Refine Machine (PRM) module to find a trade-off between local and global representations in the features and a Residual Steps Network (RSN) module to learn delicate local representations by effective intra-level feature fusion strategies. Since top-down approaches depend on human detection, which is prone to failure if the limbs are occluded. To address the occlusion issue, Qiu et al. [40] created an occluded pose dataset and proposed a occluded pose estimation and correction module. Moreover, Xu et al. presented ViPNAS [41], which facilitated

pose estimation and solved the problem of occlusion by utilizing temporal information in video sequences.

The bottom-up pipeline comprises two primary steps: body joint detection, which involves extracting local features and predicting body joint candidates, and joint candidates assembly for individual bodies, which entails grouping joint candidates to construct pose representations using part association strategies, as illustrated in Fig. 2.2(b). Cao et al. [42] developed a detector named OpenPose that predicts keypoints via Part Affinity Fields (PAFs) and heatmaps. PAFs are a set of 2D vector fields that encode the position and orientation of limbs, allowing for the association of keypoints to each person. Cheng et al. [43] introduced an extension of HRNet, called the Higher Resolution Network (HigherHRNet), which resolves the issue of scale variation in bottom-up pose estimation by deconvolving the HRNet-generated high-resolution heatmaps.

2.3.1.1 2D to 3D Lifting

2D to 3D lifting approaches infer 3D pose from the estimated 2D off-the-shelf human pose estimators. These 2D to 3D lifting techniques typically perform better than direct estimation techniques since they take advantage of the superior performance of 2D pose estimators [38]. However, despite the state-of-the-art performance, they suffer from depth ambiguity since various 3D pose coordinates may give the same results when projected to 2D poses. A view-invariant framework was developed by Wei et al. [44] in order to lessen the impact of differing viewpoints. Within this framework, their View-Invariant Hierarchical Correction (VI-HC) network uses view-consistent constraints to refine 3D poses. The Part-centric HEatMap triplets (HEMlets) framework [45] uses three joint-heatmaps to encode the relative depth information of end-joints, linking 2D locations to 3D human poses, to address this problem. The Ray-based 3D (Ray3D) absolute estimation technique converts pixel-space input data into 3D normalized rays [46].

For our use case, we complemented 2D pose estimation technology with a depth camera and projected the 2D joints into the 3D space using the Intel Realsense API [47, 48, 49] that employs basic computer vision 2D to 3D projection techniques [50].

2.3.2 3D Pose Estimation

Similar to 2D pose estimation, 3D human pose estimation can also be divided into two categories: top-down (detection + estimation) and bottom-up (estimation + associa-

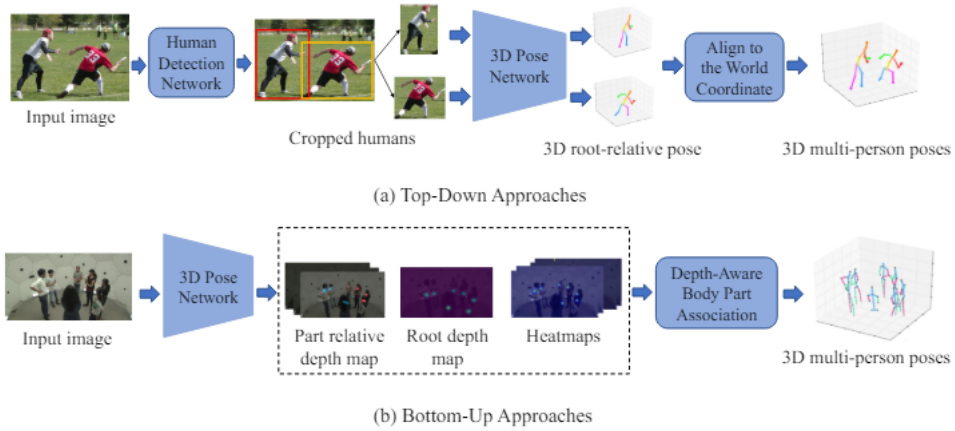


Figure 2.3: Comparison between top-down and bottom-up **3D** human pose estimators. Diagram copied from Zheng, Wu, *et al* [38].

tion). The general block diagram of these two approaches is shown in Fig. 2.3. To capitalize on their respective advantages, Cheng et al. [51] integrated top-down and bottom-up approaches for multi-person pose estimation. While the bottom-up method works well with scale differences, the top-down method works well with erroneous bounding boxes. The final output, i.e., the 3D pose, is ultimately created by combining the 3D poses produced by the two methods into a single network.

2.4 Current Methods of Assessing Physical Activities in the Elderly

Current methods of assessing physical activity in the elderly rely on accurate action recognition frameworks. These are usually followed by some metric estimation or a regression head that predicts the score or class of that particular action. The methods to assess physical activities can be divided into two categories, i.e., action quality assessment (requires underlying action information) and quality of motion assessment (does not require estimation of action category).

2.4.1 Action Quality Assessment

The field of action quality assessment (AQA) has recently expanded due to its rapidly growing real-world applications, such as healthcare and physical rehabilitation, skill training for advanced learners, and sports activity scoring. Action quality assessment

relies on activity recognition and/or segmentation, and is an effective method for measuring the action's quality that utilizes some scoring criteria. These methodologies can be divided into several categories in terms of their respective applications. We only discuss its two main applications, i.e., AQA for sports and AQA for healthcare.

2.4.1.1 AQA for Sports

In early studies on pose-based Action Quality Assessment (AQA), the process is typically divided into three main steps. First, the system tracks the locations of key body parts such as hands and feet. Next, it extracts important features like position, speed, and direction from these tracked points. Finally, a score or grade for the action's quality is calculated using either set rules or machine learning methods. For instance, Pirsiavash et al. [52] introduced a regression-based approach for evaluating action quality in Olympic sports like diving and figure skating. They first capture the athletes' body poses, which are then encoded using the Discrete Cosine Transform (DCT). These transformed pose features are fed into a Support Vector Regression (SVR) model that predicts action quality scores.

AQA primarily falls into two main methodologies based on the type of input data: pose-based methods [52], which focus on analyzing body positions, and appearance-based methods [53, 54], which rely on visual elements captured by cameras. However, obtaining accurate pose data in the sports domain is challenging [55]. The collected information is often incomplete due to athletes adopting complex positions or certain body parts being obscured. Due to these limitations of pose-based methods, researchers have shifted their focus more towards appearance-based techniques, that utilize both spatiotemporal visual features [56, 57]. These methods have achieved considerable success in recent years.

2.4.1.2 AQA for Healthcare

This area has predominantly utilized wearable sensors, as evidenced by several studies [58, 59, 59]. Seifert et al. [60] explore radar micro-Doppler signatures for gait analysis. The objective of their study is twofold: to identify changes in patterns of the gait and address the challenge of intra-motion classification for gait recognition.

In [61] the researchers utilized a Microsoft Kinect to focus on subjects who walk on stairs. They presented a learning-based method for the assessment of human movement quality in healthcare. The method consists of two key statistical models: pose

and dynamic. The pose model quantifies the likelihood of standard body positions using a probability density function, while the dynamic model accounts for temporal sequences through a continuous-state Hidden Markov Model (HMM). During the inference phase, each frame in the sequence is categorized as normal or abnormal based on the deviation of the observed data from the statistical models. This deviation is assessed using a log-likelihood metric with an empirically determined threshold.

AQA for sports (section 2.4.1.1) involve complex poses such as squatting and front somersaults that occludes some joints and lead to an unusual view of the human body which makes it tricky for normal 2D pose estimators to estimate the pose. Hence, it is beneficial to use parametric models such as [62, 63], to have both shape and pose information. Researchers are now shifting towards using such appearance-based models [56, 57]. Currently, most research on AQA for healthcare relies on pose-based models and lack the shape and orientation information of the limbs. This might be very useful in applications where the orientation of the limb such as flexion angle is required. Also, there is a possibility that AQA for healthcare might involve complex poses as well. However, the use of parametric models for pose estimation in this area is still under-explored and out of scope of this thesis.

2.4.2 Quality of Motion Assessment

The work discussed in this subsection focuses on motion-based assessment of movement in general without relying on accurate action recognition and differs from the work discussed in the previous section which was mainly based on recognized action (e.g., 'walking') quality assessment. Most of the previous research in this area involves participants diagnosed with Parkinson's disease (PD) and focuses on assessing the severity of the disease. For example, Jeon et al. [64] conduct a comparative study of various machine learning algorithms, including decision trees, discriminant analysis, support vector machines, k-nearest-neighbor, and random forests using data from wearable devices to quantify hand tremor severity.

Liu et al. [65] focus on assessing Parkinson's disease (PD) tremor severity. They address the challenge of capturing subtle and continuous tremors in different body parts, such as the jaw, hand, and leg. The authors use Eulerian video magnification for pre-processing to amplify subtle tremors and introduce a model called the global temporal-difference shift network, which focuses on the micro-temporal changes caused by the tremors. To further improve the accuracy, the model introduced

a global shift module, allowing each segment of the video to incorporate global temporal features.

Elkholy et al. (2019) created a method to evaluate neuro-musculoskeletal disorders in the elderly, such as PD, using 3D skeletal data obtained from depth cameras. Their system emphasizes three main features: velocity magnitude, asymmetry, and deformation of the center-of-mass trajectory to analyze the speed and patterns of movements. For training, they develop probabilistic models using the Gaussian Mixture Model (GMM) and Kernel Density Estimation (KDE) based on normal sequence descriptors. During the inference phase, the system calculates the likelihood that a test sequence is normal/abnormal using the trained GMM and a predetermined threshold. Additionally, a multiple linear regression model scores abnormal movements based on expert medical guidance.

Guo et al. (2022) examine video-based methods to assess Parkinson's disease severity by analyzing hand movements. They use a Graph Convolutional Network (GCN) to study the coordination of hand joints and the skeletal structure of the hand. They face challenges in extracting detailed features and ensuring model robustness. To overcome these issues, they propose a tree-structure-guided GCN enhanced with group-sparse contrastive learning. This method utilizes the natural tree structure of the human hand to create a detailed graph representation, capturing key motion features from the fingertips to the palm.

2.5 Summary

To summarize, almost all of the research in the past for action or motion quality assessment has been done either using the lower body (i.e., for gait analysis) or the full body (i.e., for individuals suffering from certain neurological issues, e.g., Parkinson's). To the best of our knowledge, no motion analysis datasets exist or any such techniques have been explored that strictly focus on analyzing the upper-body movements of individuals. This thesis fills in this gap by targeting multiple facets of the problem of motion analysis from the upper body. In the first phase, we present a video dataset of healthy aging individuals, with eating as the primary activity where we simulate changes over time using weights. Phase two of this research explores recognition models that understand the eating characteristics of individuals. The third phase explores the bias across different subjects when training machine learning models for quantifying performance decline. The last phase presents an autonomous pipeline for

analyzing both musculoskeletal performance decline and eating behavioral changes, to assist healthcare workers and provide human interpretable statistics.

Chapter 3

EatSense: Human Centric, Action Recognition and Localization Dataset

This chapter introduces a densely annotated dataset called EatSense, that captures individuals eating at a dining table in real-world, uncontrolled settings. EatSense stands out for its unobtrusive, people-centered approach, focusing on the upper body. It offers a valuable resource for analyzing eating habits and investigating changes in motion or motor skills over time.

3.1 Introduction

Numerous comprehensive datasets are accessible to the public for tasks like action recognition, temporal action localization, and tracking people's daily activities [66]. They encompass diverse action categories for recognition and temporal segments for localization, serving as benchmarks for evaluating numerous algorithms [67]. Despite their extensive coverage in terms of total recording hours and inclusion of various challenging scenarios, these datasets fall short in their capacity to model particular behaviors or identify changes in motion indicative of performance decline among subjects.

This brings us to the significance of modeling behavior and identifying minor changes. By modeling a person's behavior, such as their eating habits, we gain deeper insights into their daily practices. Moreover, the capability to detect subtle changes in motion holds immense value in situations where long-term monitoring is required, notably for elderly individuals or for evaluating changes in athletic prowess [68],[69], or assessing changes in performance after some treatment.

One way to assess performance changes in the elderly population is to record older participants over long periods. However, that would have taken weeks or months to see any noticeable changes in their performance. Also, any drastic changes would have to be reported immediately and would have raised ethical concerns resulting in halting the recording sessions So, due to limited resources and time frame, this was not possible at this stage of the study. So, to induce a simulated change in motion, weights are attached to the wrists of subjects during eating sessions[1]. Adding weights has not been previously validated to cause the same effects as geing. Weights might be one way to simulate changes in motor function, but they may not adequately capture the intricacy of natural motor performance decline, especially in older adults or people with certain medical conditions.

Introducing wrist weights alters human kinetics by potentially enhancing muscle stiffness, consequently influencing movement patterns and kinematics. The notion of employing weights to mimic upper-body degeneration has been substantiated in a prior study [70]. Similarly, in [71], a comparable approach is utilized to illustrate various gait abnormalities. Historically, various techniques have been investigated to induce temporary palm stiffness and restrict fine motor control of fingers [72], [73].

For video processing and action understanding, most publicly available datasets have shortcomings such as firstly, sparse annotations within untrimmed videos or dense labels in short untrimmed video instances, and secondly, lacking detailed sub-action level annotations, i.e., they do not have the sub-actions (atomic actions) level of annotations; instead, they typically provide only high-level action labels, such as ‘eating’ or ‘drinking’. EatSense bridges this gap by providing dense annotations for sub-actions involved in a video containing one full eating session.

EatSense contains data from 27 subjects representing 13 nationalities, thereby introducing diversity in eating methodologies, utensil preferences (forks, chopsticks, etc.), and dietary choices.

The contributions of this chapter are:

- A new untrimmed dataset named EatSense for action recognition, temporal action localization, and quality of motion assessment is presented.
  - We provide dense labels (frame-wise, $\approx$ 114.1 actions per video sequence of 11 minutes in length on average) with 3 levels of abstractions (see section 3.3.2).

[1]The weights are not intended to model aging but rather to show the detectability of motion changes and simulate performance decline.

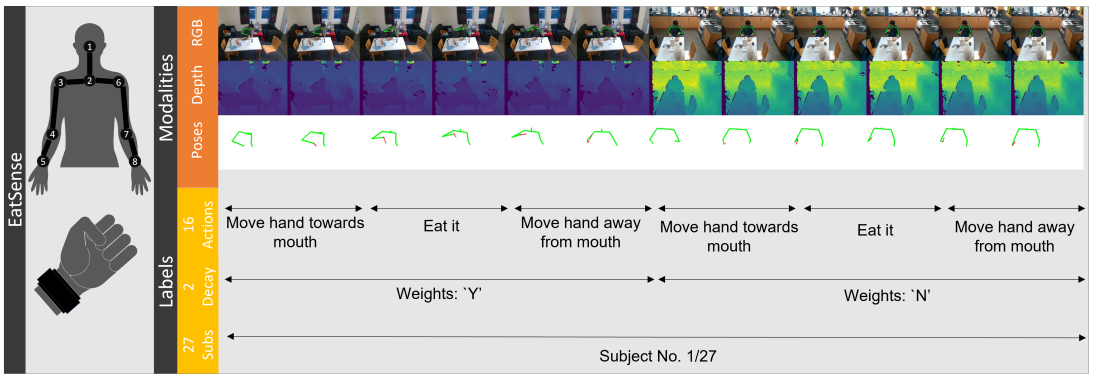


Figure 3.1: EatSense is an eating sub-action recognition/analysis dataset, that consists of multiple modalities, dense annotations (zero stride among frames), and 3 levels of abstractions of labels.

- We provide comparisons against other publicly available datasets where, unlike many datasets, EatSense contributes to both the computer-vision and health-care communities (see table 3.4).
- We use deep privacy for in-painting faces to protect the privacy of the subjects. We demonstrate the effectiveness of this technique for pose-based action recognition through experiments (see section 3.4).

3.2 Literature Review

A review of publicly available datasets related to action recognition, temporal action localization, and activities of daily living is provided.

3.2.1 Public Datasets

There are many publicly available action classification datasets. Some of the large-scale datasets can be divided into four categories according to their targeted application, i.e., datasets for action recognition (trimmed video datasets), datasets for temporal action localization (untrimmed video datasets), datasets for activities of daily living analysis, and lastly, quality of motion assessment datasets.

3.2.1.1 Trimmed Video Datasets

Datasets featuring singular actions within each video sequence are categorized as action recognition datasets, lacking the capability to spot ongoing activities temporally.

Notably, NTU-RGB-D 120 [74] stands out as one of the most extensive benchmark datasets for action recognition. Comprising 114,480 video sequences involving 106 subjects with 120 action categories. It encompasses a wide array of daily routines, group interactions, and medical scenarios. Data collection for this dataset was done with multiple sensors, including RGB, depth, and infrared technologies.

Kinetics-700 [75] represents another extensive dataset, featuring 700 distinct action categories and a vast collection of 650,317 video clips sourced from YouTube, with no fewer than 450 clips available for each action class. The action labels span a diverse range, encompassing numerous activities integral to daily life, such as, 'pouring milk', and 'drinking', among others. Moreover, Goyal et al. introduced Something-Something v2 [76], an ego-centric dataset tailored specifically to human hand gestures, including actions like placing items on surfaces or rotating objects. This dataset comprises 220,847 videos recorded at a frame rate of 12 frames per second (fps), featuring 174 distinct classes. Recorded within controlled environments, the dataset captures individuals engaging in predefined actions involving everyday objects.

Another dataset, HMDB51 [77], was assembled using a combination of digital movies and YouTube videos, with at least 101 videos per class and 7,000 manually annotated videos overall. The 51 action labels included in HMDB51 can be broadly categorized into four groups: general body movements, general facial actions, general body motions with object interaction, and general body movements for human contact. Jhuang et. al. presented J-HMDB [78] which is a subset of the HMDB51 dataset that has 21 classes and contains annotations for human joints. In the paper, these joint locations were further exploited to estimate ground-truth optical flow and segmentation.

This section shows that computer vision-based action recognition with a single clip-wide label is the targeted application for numerous large-scale datasets (including hundreds of classes). As a result, the potential applications of these datasets for healthcare or behavior modeling and understanding are severely constrained. In contrast, EatSense offers a novel dataset for motor function decline assessment and healthcare monitoring, together with full-length individual eating sessions.

3.2.1.2 Untrimmed Video Datasets

Numerous publicly accessible datasets offer avenues for exploring the temporal action localization challenge, each with different label settings. These datasets include video clips labeled with only one action, videos with sparse labels indicating prolonged periods of no activity between actions, and videos with dense labeling covering the entirety

of the video duration, either without unmarked segments or with overlapping labels occurring concurrently.

ActivityNet-1.3 [79] includes 203 routine actions, such as 'shoveling snow' and 'cleaning shoes'. These activities fall into the following general categories: cars, housekeeping, pets, interior and exterior maintenance. With 849 hours of video and an average of 137 untrimmed videos per action class, ActivityNet is a large-scale dataset. It consists of classes with 1.54 actions per 1.9-minute clip, i.e., sparse ground truth labeling. Another example of such a dataset is FineGym [80], with 530 sub-actions such as 'floor exercise' and 'vault' in uncut videos. This dataset is human-centric and only shows one subject in the field of view. This was collected from YouTube videos that show people doing different types of gymnastics.

PKU-MMD [81] is a large-scale video dataset intended for multi-modality action analysis and action recognition. It has two phases, with 51 and 49 action labels correspondingly. Ten interaction actions and forty-one daily actions make up the two groups into which the action labels can be divided. The dataset includes 2000 short videos (roughly 2 minutes) and 1076 long videos (about 4 minutes), all of which were captured from various angles.

Dense annotations can be found in multiple datasets. AVA [82] and Sphere-H130 [83] are two examples of such datasets. AVA is made up of 430 clips, each of which is a 15-minute clip that has 80 actions cropped from different movies. As a result, this dataset contains examples of numerous subjects interacting with either the surroundings or one another. The 130 sequences of 13 actions totaling roughly 70 minutes, carried out by 5 subjects in a home environment, make up the Sphere-H130 action dataset. Nevertheless, this dataset lacks real-world diversity because the subjects only carry out a limited set of tasks.

UCF-101-24 [84] is another extensive action recognition dataset, featuring RGB videos sourced from YouTube. This dataset contains 101 action categories, encompassing a total of 13,320 videos spanning 27 hours. The action categories are broadly classified into five types: 1) Human-Object Interaction, 2) Body-Motion Only, 3) Human-Human Interaction, 4) Playing Musical Instruments, and 5) Sports. In contrast, Epic-Kitchens [85] diverges from the Something-Something-v2, presenting a non-scripted dataset where subjects are instructed to perform tasks in a kitchen according to their preferences. Epic-Kitchens contains 4,053 distinct classes captured over more than 100 hours of high-definition kitchen recording sessions.

To sum up, action localization task datasets contain large numbers of untrimmed

videos and action labels; however, many of them still lack dense temporal labels, and others lack consistent sets of sub-actions within a single large action. For these reasons, they are rarely used for individual long-term behavior modeling. EatSense fills these gaps with its dense labeling for the videos and 16 sub-actions that make up the eating action.

3.2.1.3 Eating Related Datasets

A dataset for food intake gestures during meals was presented by Tang et al. [86] in a recent study. This dataset is a component of the Clemson Cafeteria Database (CCD) [87]. In a university cafeteria, 276 participants were filmed while consuming a total of 374 different foods and beverages. Three distinct gesture classes, such as bite, drink, and others, make up the dataset. In a different study, Shengjie et al. [88] used a head-mounted camera to record an ego-centric dataset in a free-living setting. They then formulated a binary classification problem to differentiate between eating and non-eating activities. Lastly, Neves et al. [89] provided a thorough analysis of methods utilized in eating gesture detection for those who are interested in this area.

OREBA (Objectively Recognizing Eating Behavior and Associated Intake) [90] is a dataset created to offer a wealth of information gathered from multiple sensors during communal meals to researchers interested in detecting intake gestures (single gestures denoted as Intake, Intake-Eat, Right, Spoon). This dataset includes video recordings made with a 360-degree camera mounted in front of the person eating and a sensor box with an accelerometer, gyroscope, and IMU fastened to each hand. Small-scale datasets including Accelerometer and audio-based Calorie Estimation (ACE) [91], Clemson [92], and Food Intake Cycle (FIC) [93] have also been presented by other research studies. These datasets primarily focus on aspects of intake gestures, like chewing and swallowing behaviors.

Men et al. introduced a dataset [94] tailored to discern high-level actions associated with specific food consumption activities, such as ‘eating a steak’ or ‘drinking from a plastic bottle’. The dataset’s principal aim was to estimate the frequency of self-feeding behaviors and gain insights into eating and drinking patterns. Utilizing Microsoft Kinect, they captured skeleton motions to facilitate analysis. On the contrary, Mobiserv-AIIA [95] was crafted for evaluating specialized meal intake strategies aimed at mitigating undernourishment or malnutrition risks. This dataset features recorded videos captured within a controlled laboratory environment, employing multiple cameras positioned at various angles for comprehensive coverage. Unlike datasets

with atomic action granularity, Mobiserv-AIIA concentrates on high-level actions such as ‘eating’, ‘drinking’, and ‘slicing’ across different meal types (breakfast, lunch, and fast food) and utilizing diverse utensils (spoon, fork, glass of water, etc.).

Previous research has presented a variety of datasets and studies focusing on recognizing actions and gestures related to activities such as eating, drinking, and swallowing. However, these studies have been limited in that they do not explicitly emphasize the most common sub-actions inherent in the eating process. Onofri et al. [96] elucidate that behavior analysis algorithms based on activity recognition necessitate two categories of knowledge: contextual knowledge and prior knowledge. Additionally, many past datasets, particularly those based on vision, lack prior knowledge as they do not include sub-action information. Consequently, they are unable to offer the capability to fully explore the complete behavior of individual subjects based on the numerous sub-actions involved during eating. EatSense bridges this gap by encompassing the 16 most common sub-actions throughout the entire eating process.

3.2.1.4 Activities of Daily Living Datasets

The MSR-Action3D [97] dataset consists of the 3D location of 20 joints in each frame. Twenty actions, including ‘tennis serve’ and ‘golf swing’, are included in the dataset. The MSR-DailyActivity dataset [98] was created to simulate a person’s daily activities while seated on a couch. It has 320 samples of 16 commonplace tasks, like ‘play guitar’ and ‘eat’.

Daily routine actions are included in some trimmed and untrimmed video-based datasets described in previous sections, such as Something-Something v2 [76], ActivityNet-1.3 [79], and Sphere-H130 [83].

The datasets mentioned above contain general activities that do not strictly conform to the definition of ADLs (i.e., based on a loose definition of ADLs). However, a subset of action categories in these datasets are ADLs. The definition and a complete list are added in the appendix B.

3.2.1.5 Quality of Motion Assessment Datasets

Numerous datasets measure the subject’s quality of motion in addition to concentrating on the ongoing activity. The purpose of Sphere-Walking [61] was to use gait analysis to assess gait motion quality. Six participants were videotaped ascending a flight of stairs for this dataset, and each participant was labeled with a score from medical

professionals. A benchmark dataset for studies on gait impairment, the Init Gait DB [99] was collected in a laboratory setting under controlled conditions. Eight distinct gait styles in which the human body's posture and limb movements changed, were modeled. It was captured with RGB cameras installed at various viewing angles.

Another dataset based on gait analysis that simulates nine different walking gait patterns is the walking gait dataset [71]. To imitate these gait patterns, a thick sole was placed in the shoe of the subject, or weights were tied at the ankle. This was recorded using Microsoft Kinect as the subject was walking on a treadmill with two flat mirrors placed behind them. Whole-body or lower-body gait analysis is also used in Sphere and other datasets. Furthermore, studies like [100], [101], and [102] provide an extensive summary of databases for gait analysis that are accessible to the general public. In-depth discussions of the difficulties and solutions surrounding gait analysis techniques are found in [100], [103], and [104].

To our understanding, none of the existing datasets are specifically tailored for evaluating human motion quality concerning action-based eating behaviors, particularly focusing on the movement of upper body joints.

3.3 EatSense Dataset

We chose eating as our performance monitoring activity because it is a set of actions performed regularly that can allow us to study changes in behavior [105] and upper-body movement that healthy individuals experience daily. Compared to other actions that may vary over time, eating is one of the most common and frequent activities in a person's daily routine [106]. Furthermore, people often continue to eat in the same ways even when they experience mild physical limitations that could impair their mobility. Finally, we think eating can be used to detect and assess motor movement changes. People's ability to move freely becomes increasingly limited as they age, which also has an impact on how well they can eat [107].

The primary goal of this dataset is to address several knowledge gaps about human eating behavior and healthcare applications by introducing a new dataset. The dataset provides the capability to localize sub-actions in videos with an average of 114.1 sub-actions in an average 11.5-minute video and provides a comprehensive labeling system with up to 16 action sub-classes, including very short actions. The dataset also places a strong emphasis on behavior understanding, especially as it relates to eating posture and hand gestures. Lastly, the dataset makes it possible to identify the decline in motor

movement, which is replicated by slightly altering upper-body movements with wrist weights. The footnote[2] contains a link to the dataset.

3.3.1 Data Collection

An RGB-D Intel RealSense camera D415 was deployed in a dining-room setting. This is a low-cost depth camera that offers accurate indoor 3D depth estimation [108]. The 2D poses were converted from a 2D to a 3D frame of reference using the depth maps. The camera was positioned to capture an oblique angle of the dining table, and it was limited to a single person per frame. Multiple locations were used for the recording, with different backgrounds and camera-to-subject distances. If another individual entered the camera's field of view or walked by the person who was eating, the frames were discarded. The recording team did not intervene or provide input during the subjects' meal times. The configuration of the camera system in the dining room is displayed in Fig. 3.1. Fig. 1.1 displays a single dataset sample with and without wrist weights. We used velcro-stitched wrist/ankle weights. It also shows the weights were attached to the subjects' wrists, denoted as joints 5 and 8.

3.3.2 Data Labelling

3.3.2.1 Poses

In the initial abstraction level, the skeleton of the upper-body pose was estimated, representing the 3D joint locations of 8 key joints: nose (1), chest (2), right shoulder (3), right elbow (4), right wrist (5), left shoulder (6), left elbow (7), and left wrist (8). HigherHRNet [109] was employed to estimate the location of 2D joints, which were subsequently projected into 3D space utilizing depth map measurements. The selection of HigherHRNet was based on empirical evaluation among various commonly utilized pose estimators to ensure accurate ground truthing of the data.

Given the importance of pose estimation in skeleton-based action recognition, an experiment was carried out to determine which algorithm would work best with EatSense. First, 100 images were sampled from the collection of videos. Second, the 2D poses of the images were carefully hand-labeled. Third, the deep learning-based pose estimation algorithms, such as OpenPose [110], darkpose [111], deeppose [112], HigherHRNet [109], and vipnas [41], were then applied to the images.

[2]https://groups.inf.ed.ac.uk/vision/DATASETS/EATSENSE/

Mean squared error (MSE) in 3D space and mean average precision (mAP) in 2D space were the two metrics used to assess the accuracy of the pose estimators. For mAP, the distance between the predicted and ground truth key points using the Gaussian kernel was used to calculate the intersection over union (IoU), also referred to as object key-point similarity (OKS) in the case of key points. The dataset includes situations in which the subject places their hand on their lap (under the table) or crosses their arms over one another, covering some joints. The tests exposed that OpenPose could not determine an individual's pose when a joint was not visible to the camera.

For both the manually labeled 2D joint locations and the 2D joint locations predicted by each classifier, the depth map was then used to project the 2D joint locations into 3D space. To measure error on common grounds, MSE (3D) was computed using only the visible joints. The mAP with the IoU [0.5,1] threshold is displayed in Table 3.1. It also shows the mean squared error (MSE) for every pose estimator. MSE emphasizes more on the distance in 3D, i.e., the higher the distance, higher MSE, whereas the mAP thresholds on distance but is calculated using only precision and recall, i.e., true/false positives. Hence, for our use case, we chose the pose estimation method with the least MSE. HigherHRNet's MSE was significantly lower (9.7×10^{-3} m) than the alternatives for roughly the same mAP, we decided to use it for estimating ground truth poses for our proposed dataset.

Table 3.1: mAP (@IoU=0.50) and MSE (3D) of the skeleton estimation as compared to hand-labelled ground-truth skeletons.

Algorithm	mAP	MSE (m)
OpenPose [110]	23.4	5.7e-2
Deeppose [112]	59.6	1.18e-2
Darkpose [111]	64.6	1.79e-2
Vipnas [41]	63.0	1.6e-2
HigherHRNet [109]	63.1	9.7e-3

3.3.2.2 Actions

For the second level of abstraction, the eating actions were broadly divided into five categories based on joint location and motion. The categories are hands-based, motion-based, head position-based, body posture-based, and others. In the third level of ab-

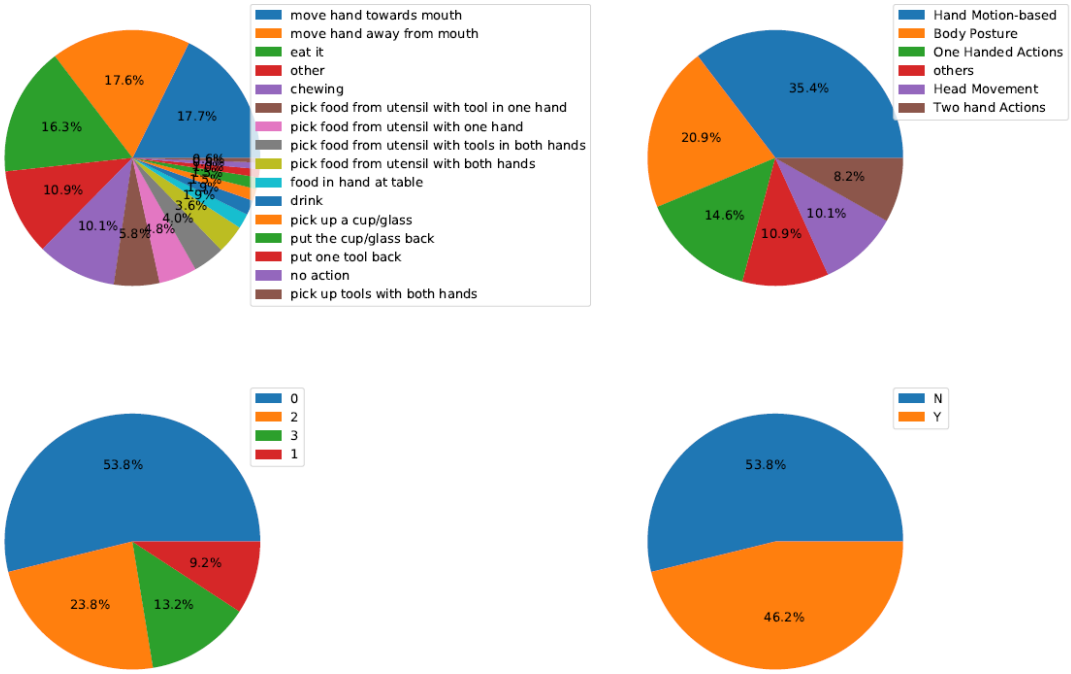


Figure 3.2: Distribution (in percentages) of various labels according to their occurrence in the dataset. Top-Left) distribution of individual 16 sub-actions. Top-Right) distribution of actions based on abstraction-level 1 for the labels. Bottom-Left) distribution of four weight classes, Bottom-Right) occurrence percentage of videos with weights 'Y' and 'N'.

straction, each category is further subdivided into sub-actions, which encompass several atomic actions some of them lasting less than or equal to a second. Our dataset approximately adheres to Zipf's law, as depicted in the pie charts in Fig. 3.2 (top-left). Given the potential unreliability of experiments on actions with few instances, we only include and present experiments involving actions with a minimum of 40 instances.

Manual ground truthing was conducted with the assistance of a video image annotator (VIA) [113]. In cases where two possible labels were correct for sub-actions, precedence was given to actions performed by the hands. For example, simultaneous actions such as 'chewing' and 'food in hand at table' was labeled as 'food in hand at table'. Similarly, instances like 'chewing' alongside 'move hand away from mouth' were annotated as 'move hand away from mouth'. Furthermore, instances where an individual was engaged in talking, using a mobile phone, or any other unlisted activity, were categorized as 'others', irrespective of concurrent eating actions. For example, 'talking', while 'chewing', or 'food in hand at table' were categorized as 'others'. Ad-

ditionally, frames depicting complete rest were labeled as 'no action'.

3.3.2.3 Weight Labels

For every participant, the researchers recorded a minimum of two sets (and as many as four sets) of videos to simulate affected motor movements. Each participant's second (and third and fourth) set had some weight fastened to their wrists to affect the properties of motion. [3]. However, in the first set, the subject was not wearing any weight and was able to move normally. Depending on whether the weight was added or not, the videos were labeled as 'Y' or 'N'. This 'Y'/'N' topology was later expanded to four distinct weight values or labels based on the weight that a subject is wearing around their wrists, which are 0kg, 1kg, 1.8kg, 2.4kg.

Several volunteers actively participated in the labeling process, which made it difficult to keep the labels consistent. We developed a two-step quality control system to attain reasonably consistent labels to address this concern. Initially, we gave the volunteers instructions on how to label eating actions using a detailed video guide that explained proper label usage for each action as well as naming conventions for actions. One of the primary researchers then carefully went over the labeled videos to make sure the quality of the labels was maintained.

3.3.3 EatSense Statistics

The dataset contains 135 video sequences of 27 subjects with different cultural backgrounds to ensure diversity in ages, ethnicity, body size, gender, and eating behaviors. These were recorded with a resolution of 640x480 at 15 frames per second (fps). Actions performed in each of the individual videos are shown at the top fig. 3.3.

Different cultures and geographical areas have different eating customs and cuisine preferences. In East Asian countries, people typically eat with chopsticks, but in South Asian countries, people typically eat with their hands, either directly from the pan or from a utensil/dish. The purpose of the subject selection was to optimize both generalizability and diversity. Subjects from thirteen different countries make up the EatSense dataset; Table 3.3 gives details on their ethnicity by region, age groups, and eating utensils used. Since different subjects choose different tools, the actions taken are inevitably subjective as well. The dataset also complies with the aforementioned

[3] Adding weights is not intended to be a model for aging or a neurological disorder, but to demonstrate that, 1) changes in motion performance can be detected and 2) the weights simulate performance decline.

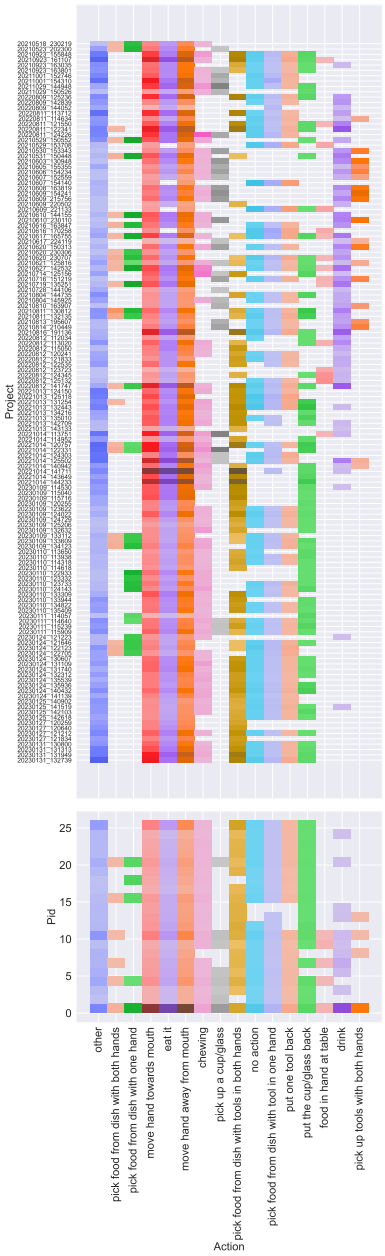


Figure 3.3: Top) Actions performed in each of the individual videos. The vertical axis shows the name of each of the individual videos, which has the format '$\{date\}\_\{unix-time\}$', collectively marked by the keyword 'Project' in the dataset. Bottom) shows the actions performed by individual subjects. The variations in the color mean the frequency of occurrence of each action, where darker shades means more instances. It has subject IDs (Pid) on the vertical axis and actions on the horizontal axis. This is a vectorized image, best viewed zoomed in.

convention, as shown in Fig. 3.3 at the bottom.

Table 3.2: Average time in seconds taken by an instance of the action and total number of instances of the action for each of the actions in the EatSense dataset

Actions	Instances	
	total no.	avg. time
chewing	795	6.165
drink	247	2.723
eat it	2630	0.717
food in hand at table	344	3.868
move hand away from mouth	2792	0.625
move hand towards mouth	2851	0.844
no action	64	9.007
other	2057	7.043
pick food from dish with both hands	282	5.342
pick food from dish with one hand	440	3.741
pick food from dish with tool in one hand	1548	3.943
pick food from dish with tools in both hands	467	6.449
pick up a cup/glass	213	1.218
pick up tools with both hands	65	1.880
put one tool back	253	1.067
put the cup/glass back	214	1.618

3.3.4 Characteristics of EatSense

EatSense has several attractive properties that distinguish it from other existing datasets.

Unlike most large-scale datasets currently available, every one of these videos has dense labels, meaning that no temporal segments are unlabeled. The dataset's current state can be readily expanded by identifying foods and tools, looking for interactions between people and objects, and determining the kinds of food that an individual eats, which could enable a thorough nutritional analysis. EatSense has consistent

Table 3.3: The table shows the diversity of subjects divided into five age groups. This shows the tools, foods, and ethnical origins of all 27 subjects involved in the dataset. Pid refers to person IDs.

Age-Groups	Pids	Genders	Tools	Foods	Ethnicity
below 30	2, 3, 4, 5, 6, 9, 10, 11, 12, 13	8M, 3F	Fork and Spoon, Fork, Spoon, No tool, Fork and Knife	Rice, Noodles, Soup, Shewarma, Apple, Toast, Only-Drinks, Roti, Egg, Steak, Sandwich, Pizza, Salad	South Asian, East Asian, British
30-39	7, 8, 15, 23, 25, 27	4M, 2F	No tool, Spoon	Shewarma, Rice	East Asian, European
40-49	19, 21, 22	1M,2F	Fork and Spoon, Fork and Knife	Rice	South-American, British
50-59	17, 25, 26	3F	Fork and Spoon, Spoon	Rice	British
Above 60	1, 16, 18, 20	3M,1F	Fork, No tool, Spoon, Fork and Knife	Rice, Roti, Soup, Wafers, Steak, Toast	American, British

backgrounds and human posture-centric action instances, in contrast to other existing datasets where a background/environment plays a significant role in differentiating between different actions.

As was previously mentioned, EatSense includes ground truth for several abstraction levels. Frame-wise action and pose labels are made available. The lower body of an individual is covered when they eat at a dining table and is not fully visible. EatSense is therefore a dataset that focuses strictly on upper-body posture. EatSense is useful for monitoring human health. For instance, it has an abstraction level of labels for which the subjects were requested to wear different weights on their wrists to replicate a person's performance decline over time. These added weights essentially

Table 3.4: Comparison of the proposed dataset EatSense against publicly available datasets used for action localization, action recognition, and healthcare research. BUC stands for human behavior understanding capability, HCC stands for healthcare capability, UV stands for the untrimmed videos, S# stands for the number of subjects which is marked multiple(M) for datasets lacking specific numbers, Lbs indicates the type of labels single (S), sparse, (Sp) and dense (D). TC stands for targetted community, i.e., computer vision (CV) and healthcare (HC). Settings refer to a controlled (C) or uncontrolled (UC) environment for recording.

Datasets	C #	BUC	HCC	UV	S#	Setting	Lbs	TC	Avg. # act.	Avg. vid. dur.
Epic-Kitchens-100 [85]	4053	✗	✗	✓	M	UC	D	CV	128.5	8.5 m
NTU-RGB-D [114]	120	✗	✗	✗	106	UC	S	CV	1	7.21 s
Kinetics [75]	700	✗	✗	✗	M	UC	S	CV	1	10 s
HMDB [77]	51	✗	✗	✓	M	UC	S	CV	1	∼2 s
J-HMDB [78]	21	✗	✗	✓	M	UC	S	CV	1	∼2 s
Something-Something-v2 [76]	174	✗	✗	✗	>1300	C	S	CV	1	4.03 s
PKU-MMD (P1) [81]	51	✗	✗	✓	66	UC	D	CV	20	3∼4 m
PKU-MMD (P2) [81]	49	✗	✗	✓	13	UC	D	CV	7	1∼2 m
Activity-Net 1.3 [79]	200	✗	✓	✓	M	UC	Sp	CV	1.54	1.9 m
THUMOS14 [115]	20	✗	✗	✓	M	UC	D	CV	15.4	1.1 m
UCF-101-24 [84]	101	✗	✗	✓	M	UC	Sp	CV	1.4	5.1 s
FineGym [80]	530	✗	✗	✓	M	UC	Sp	CV	42	2h
FineDiving [68]	52	✗	✗	✓	M	UC	Sp	CV	3.26	6.9
AVA [82]	80	✗	✗	✓	M	UC	D	CV	1380	15 m
MSR-DailyActivity [97]	16	✗	✓	✗	10	C	S	CV, HC	1	6 s
Sphere H-130 [83]	13	✗	✓	✓	5	C	D	CV, HC	13	5 m
Mobiserv-AIIA [95]	13	✗	✓	✓	12	C	Sp	CV, HC	N/A	N/A
Init Gait DB [99]	7	✗	✓	✗	10	C	S	HC	N/A	N/A
OREBA [90]	2	✓	✓	✗	100	UC	S	CV, HC	N/A	N/A
CCD [86]	5	✓	✓	✗	264	UC	S	CV, HC	N/A	N/A
EatSense (ours)	16	✓	✓	✓	27	UC	D	CV, HC	114.1	11.5m

affect the subject's movement such as slowing them down or distorting their posture, etc. It may be possible to spot a serious health issue by keeping a close eye on eating behaviors and searching for motor movement degradation.

A comparison of EatSense with other relevant datasets from the computer vision and AI healthcare communities is provided in Table 3.4. The table includes widely used healthcare-based datasets like MSRDailyActivity3D, Init Gait DB, and OREBA alongside widely used action recognition/localization datasets like Thumos14 [115], FineGym, NTU-RGB-D, and AVA. The EatSense dataset's numerous attributes are shown in the table, providing a wealth of research opportunities. These include the capacity to create models for upper-body-focused models, eating behavior analysis, action localization with dense labels, simulate performance decline in motor movement assessment, and action recognition.

3.4 Data Privacy Protection

3.4.1 Introduction

One of the special challenges associated with the vision modality is maintaining privacy. Using only second-order data extracted from the original image data is a common solution to address the issue of people's identities and homes being collected as data. Usually, this takes the form of point clouds [116], skeletal graph data [117], or silhouettes. The problems with this solution are as follows: (1) actual images need to be gathered prior to the privacy-preserving processing, and (2) making this downstream data available restricts the possibility of experimenting with alternative data representations or new pre-processing techniques.

In order to solve this, scientists have been employing identity protection algorithms driven by machine learning. These algorithms effectively maintain privacy without requiring the conversion of data into these second-order representations, allowing for the preservation and release of a large portion of the original image data for study. While machine learning-powered 'smart' methods of obfuscation are effective at minimizing this, traditional methods of obfuscating the body and face have negative effects on performance across several tasks (pose estimation, image segmentation, etc.) [118]. Researchers [119] also found that, at a low enough degree of blurring, traditional methods only result in a small loss in pose estimation performance.

Therefore, the question is, 'Do identity protection algorithms have a significant im-

pact on downstream task accuracy after pose estimation?' This section demonstrates that pose-based action recognition is essentially unaffected by privacy-preserving routines using three different face obfuscation techniques.

3.4.2 Experiments

The experiments used RGB videos from the EatSense dataset. Three strategies (face blur, deep-privacy1 [120], and deep-privacy2 [121]) were used to obfuscate the videos. Subsequently, a two-phase method was utilized for 3D pose estimation. First, ViPNAS [122] or HigherHRNet [123] were used to estimate 2D joint positions (pose). Second, depth maps were used to project the 2D joint positions into 3D space. Using ST-GCN as the action recognition framework and top-1 accuracy as the performance metric, the effectiveness of pose-based action recognition was assessed. (Note: action recognition is discussed in detail in Chapter 4). Only EatSense videos with subjects not wearing weights were used for the evaluation; thus, 27 videos, one video for each subject. Using 18 and 9 videos for training and testing, respectively, 3-fold cross-validation was carried out to produce a reliable estimate of performance.

Visual analysis of the output videos showed that, although face-blurring significantly reduced the performance of 2D pose estimation, especially in the facial region, it also considerably obscured the subject's identity. Furthermore, both the deep-privacy1 and deep-privacy2 methods in-paint a face mask over the eating utensils (*e.g.* forks and spoons) for actions like 'move hand to mouth,' 'eat it,' and 'move hand away from mouth,' where the face is partially obscured, leading to unrealistic deformed masks. A sample frame utilizing all three face-obfuscation strategies is presented in Figure 3.4. Unrealistic or deformed face obfuscation is depicted in the middle and right images.

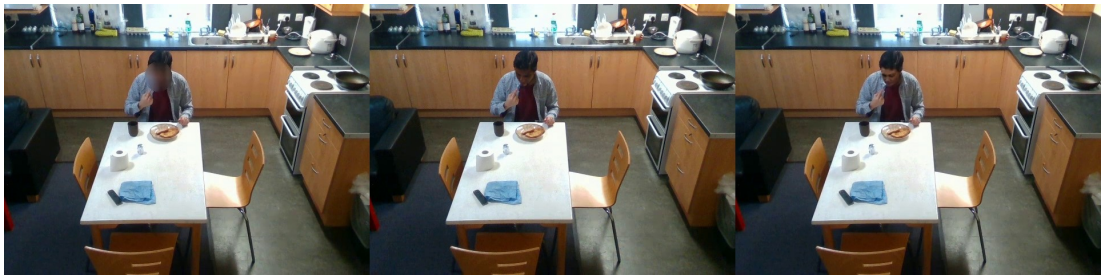


Figure 3.4: Images generated by the three face obfuscation strategies i.e., face blur (left), deep-privacy1 (middle), and deep-privacy2 (right). These show the case when the eating utensil is near the subject's mouth.

Table 3.5: The table shows the mean (μ) and standard deviation (σ) of top-1 accuracy estimated using 3-fold cross-validation on action recognition results with ST-GCN based on two 2D pose estimators and three strategies for face obfuscation.

	None		**Blur**		**deep-privacy1** [120]		**deep-privacy2** [121]	
	μ	σ	μ	σ	μ	σ	μ	σ
HigherHRNet [123]	70.33	5.57	59.11	9.95	72.08	6.41	66.38	5.40
ViPNAS [122]	67.17	8.23	57.37	10.50	66.10	7.42	64.66	5.92

The mean and standard deviations of the top-1 action recognition accuracy are shown in Table 3.5. Face-blur, mainly because of the loss of important information, clearly performs worse than the scenario where no obfuscation is applied ('None' in the table). On the other hand, even though deep-privacy1 and deep-privacy2 produce slightly damaging masks in some cases, their performance is still on par compared to the 'None' column. Consequently, while face-obfuscation may reduce the accuracy of 2D pose-estimators [118], pose-based action recognition using obfuscated data nevertheless attains a comparable degree of accuracy to that of action recognition using non-obfuscated data.

3.4.2.1 Impact of Face Obscuring

Table 3.6: The table shows the result of McNemar's test on the three cross-validation (cv1, cv2, and cv3) splits for deep-privacy1, deep-privacy2, and blurred compared to no face obfuscation i.e., 'None'. The bold p-values indicate where there is no statistically significant difference.

	Deep Privacy1 [120]			**Deep Privacy2** [121]			**blur**		
	cv1	cv2	cv3	cv1	cv2	cv3	cv1	cv2	cv3
p-value	**0.145**	**0.843**	**0.224**	2.3e-24	**0.789**	**0.064**	0.001	1e-27	5e-7
statistic	2.123	0.039	1.476	103.731	0.072	3.418	10.633	117.717	24.927

An important question is whether any of the methods for face anonymization have any impact on activity recognition. McNemar's test is a statistical test used to analyze paired nominal data. It is particularly useful when comparing the outputs of two

models on the same dataset (and test set) and helps determine if there is a significant difference between the two models. Here, McNemar's test is used to see if performance differences between each obfuscation method and the baseline 'None' is statistically significant.

We compared the predicted action classes (from each of the cross-validation splits) with three face-obfuscation methods against no obfuscation. Each of the three cross-validation sets cv1,cv2, and cv3 contained 861, 1184, and 1184 action instances, respectively. Table 3.6, summarizes the results from the McNemar test for each of the three face obfuscation models—Deep Privacy 1, Deep Privacy 2, and Blur—across three cross-validation splits. These techniques were individually compared with the baseline 'None' (i.e., no obfuscation).

For Deep Privacy 1, none of these p-values (0.145, 0.843, 0.224) are below the significance threshold of 0.05, so the null hypothesis cannot be rejected. This means that in all three cross-validation splits, there is no statistically significant difference between the performance of Deep Privacy 1 and the 'None'. The blur obfuscation technique produces very small p-values (1e-3,1e-27 and 5e-7) across all three cross-validation splits which shows a highly significant difference in performance between the blurred obfuscation and the baseline in both splits since they are smaller than the significance threshold. This supports the conclusion that the blurred technique leads to a statistically significant difference compared to the baseline.

For Deep Privacy 2, the results are more varied. It shows mixed trends in the first and second cross-validation split, and even though for the third cross-validation split the p-value is higher then the threshold of 0.05, it has a higher value of the McNemar statistic. Overall, this indicates further analysis or additional data may provide more conclusive results.

3.5 Conclusion

This chapter presents the new benchmark dataset EatSense that includes atomic actions, dense multiple abstraction levels of frame-level labels, and four weight classes (0kg,1kg,1.8kg, and 2.4kg) to simulate performance decline in motor movement. EatSense can be used as a generic training benchmark dataset for action recognition tasks specifically designed for the eating process. Furthermore, EatSense also has the capability to be used as a generic test benchmark suite for temporal action localization and action recognition.

Chapter 4

Action Recognition and Temporal Action Localization

As discussed in the previous chapter, EatSense is a dataset targeted toward the computer vision and healthcare community. In this chapter, firstly, we present an analysis of EatSense for sub-action recognition (with and without deep features). Secondly, using temporal action localization algorithms we explore how accurately can we localize the sub-actions in an untrimmed video.

4.1 Introduction

EatSense records individuals eating in an uncontrolled environment (the people face the camera and eat at a table), offering several distinctive features. Firstly, it introduces challenging atomic sub-actions for action recognition. Secondly, it consists of significantly varying durations of actions, posing huge challenges for current SOTA temporal action localization frameworks. Moreover, EatSense can be helpful in the modeling of eating behavior through a sequence of action-based behaviors. Additionally, it includes minor performance variations. We anticipate that this dataset will be valuable for future researchers, aiding in the development of robust temporal action localization networks, behavior recognition systems, and performance assessment models for eating behaviors.

We explored three research questions in this chapter:

1. For action understanding, how accurately do conventional state-of-the-art action recognition algorithms work, and how do they compare between hand-crafted and deep learning-based features?

2. How effectively can TAL networks estimate the start and end of all 16 actions using conventional temporal action localization networks?

3. Can we improve the performance of TAL networks by breaking the dataset into various subsets of actions based on similar duration?

Sub-action recognition plays a crucial role in advancing our understanding and analysis of complex human activities. Breaking down larger actions into their constituent parts enables a more precise and granular analysis of behavior. For example, in sports, recognizing the individual steps in a tennis serve (like the toss, swing, and follow-through) can give a clearer picture of the player's technique. Similarly, eating sub-actions recognition can provide a deeper understanding of the subject's long-term eating patterns, speed, and habits. For this purpose we explore (sub)action recognition techniques and conduct experiments on EatSense, employing hand-crafted feature-based approaches for explainable applications and deep learning-based approaches using RGB, RGB + optical flow, and skeleton-based techniques.

The contributions of this chapter are:

- A state model for eating micro-movements[1] that represents the most common eating behavior among subjects of all ages (see section 4.3).

- We demonstrate effective modeling of eating sub-action recognition using EatSense using both deep-learning-based networks and interpretable features (see Sections 4.4 and 4.5 respectively).

- From an explainability point of view, we demonstrate that the EatSense dataset can be used for tasks where explainability is the key, such as healthcare applications where information about individual joints is vital to understand or diagnose/track/predict a problem (see Section 4.5).

- We highlight the shortcomings of the current deep-learning-based action localization networks and provide experimental test benchmarks on EatSense (see section 4.7).

[1]Micro-movements, or sub-actions, refer to the individual and basic actions that are combined to form a single action. For instance, eating can be seen as a single action that involves several sub-actions, such as bringing the hand to the mouth.

4.2 Literature Review

This section discusses the literature review of both action classification and temporal action localization.

4.2.1 Action Classification

In general, vision-based action recognition and classification[2] frameworks can be separated into two groups according to their modalities: skeleton-based and video-based. Many researchers have investigated using hand-crafted spatial features with Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) based approaches for skeleton-based action recognition [114], [124], [125], but these disregard the human body's spatial connectivity.

To account for human joint connectivity in action recognition, researchers have proposed various approaches such as Graph Convolutional Networks (GCN) or heatmaps. Duan et al. introduced PoseConv3D (also known as PoseC3D) in [126], which utilizes a 3D heatmap volume as input for a 3D-CNN network. This design makes it less susceptible to joint estimation noise, thus enhancing robustness in action recognition. Yan et al. presented the Spatio-Temporal Graph Convolutional Network (ST-GCN) [127], which establishes both spatial and temporal graph connections. Approaches based on Adaptive Graph Convolutional Networks (AGCN) exploit the hierarchical structure of GCNs, where different layers contain multi-level semantic information to incorporate long-range dependencies of the joints for action recognition, as demonstrated in works such as [128], [129], and [130]. Recently, Chen et al. proposed a feature aggregation topology known as channel-wise topology refinement graph convolution (CTR-GC) in [131], which effectively aggregates joint features across various channels[3] and dynamically learns different topologies.

In contrast, when using RGB as a modality, Temporal Segment Networks (TSN) [132] split the video into short segments first, then samples the frames uniformly. These sampled frames are then averaged together to combine per-frame predictions. Spatial and temporal modulation blocks were introduced by Temporal Pyramid Network (TPN) [133] to align semantics and adjust the tempo among multiple levels of features extracted from the backbone. The temporal adaptive module of Temporal

[2]Note that, here, action classification and action recognition both refer to classifying trimmed videos.

[3]In GCN, channels indicate different features associated with each node, For example, in skeleton data, each joint might have multiple coordinates, and additional features such as velocity/acceleration, which can be considered as various channels.

Adaptive Network (TANet) [134] generates temporal kernels to capture global context information, which when combined with a 2D CNN, yields an effective action recognition framework.

The use of action recognition techniques in healthcare research has been the subject of numerous studies. A YOLO-based action classifier and a dataset to identify eight abnormalities, including 'backward fall' and 'chest pain', were proposed by Gul et al. in [135]. The data was gathered using a camera placed in a real-world setting. In [136], Woznowski et al. provide two granularity-based video annotation strategies—atomic labels and high-level annotations for human action recognition in healthcare—along with the full activities of a hierarchical ontology for daily living.

In order to identify eating times, Sharma et al. [137] described a convolutional neural network (CNN)-based technique for identifying hand-to-mouth gestures over prolonged periods of time, spanning from 0.5 to 15 minutes. To enhance the identification of eating events, the researchers made use of prior knowledge of other gestures. Researchers used data from the Clemson all-day dataset, which was gathered using IMU sensors that were worn around the subjects' wrists. A system for identifying eating and drinking actions was provided by Okamoto et al. [138], which also classified the food components consumed. The mouth region is detected by the system to retrieve pertinent data regarding nutritional intake.

Nevertheless, rather than identifying minor changes in a before-and-after situation, the methods associated with healthcare mainly differentiate between various anomalies. Moreover, a drawback of the majority of earlier methods is that they use deep features rather than explainable features to distinguish between two disorders. In order to better comprehend the underlying reasons for irregularities, healthcare professionals may find it more useful to employ explainable features. Furthermore, explainable features might still be quite reliable when the algorithm has difficulties distinguishing between two anomalies.

4.2.2 Temporal Action Localization

The Background Suppression Network (BSN) [139] was designed to predict the score of an action at any given time instance, along with the scores for the start and end of that specific action. It generated flexible proposals by retaining temporal positions with high scores for the action's start and end. However, these proposals were evaluated separately, overlooking the global context of the video. In contrast, the Boundary-

Matching Network (BMN) [140] aggregated the features of all proposals and evaluated them simultaneously, thereby preserving the global context of the video. Previous algorithms, including BSN and BMN, often employed an external classifier to predict action categories from video proposals and heavily relied on anchor windows.

ActionFormer [141] combined a transformer with a temporal feature pyramid network to obtain multi-scale features and recognize action categories without explicitly generating action proposals (thus eliminating the need for an external classifier) or relying on predefined anchor windows. Liu et. al. proposed the Temporal Action Detection Transfomer (TadTR) [142], where the Transformer encoder models the relationships between video snippets, capturing long-term temporal context. The decoder predicts action segments and their confidence scores. Recently, Shi et. al. proposed Temporal Action Detection with Relative Boundary Modeling TriDet [143] that targeted the problem of ambiguous boundaries. It introduced a novel Trident-head that estimates the relative probability distribution around action boundaries, for more accurate action localization.

All of the aforementioned TAL networks require a pre-trained video classification network to extract feature embeddings. This is needed due to the large footprint of untrimmed videos, so current temporal action localization networks utilize pre-computed feature embedding. Networks such as I3D [144], Temporal Segment Networks (TSN) [132] and Temporally-Sensitive Pretraining (TSP) [145] are often used for this purpose. Activity/video classification networks such as I3D, TSN, and others (that use feature embeddings) were originally designed for a classification task for trimmed videos, and do not produce suitable features for localization. However, the TSP classifier was specifically designed to assist localizers with improved temporal sensitivity by incorporating background clips and global video information.

4.3 Eating Behavioral Model

The EatSense dataset's sequences are densely labeled with 16 sub-actions of variable lengths to fully represent the eating behavior of individuals. Fig. 4.1 presents a general state diagram showing the sequential relationships between the 16 sub-actions.

Upon close inspection, it is clear that the diagram accommodates a wide range of situational variations, such as when the subject alternates between eating with one hand and one tool, or when they eat with both hands and one tool.

The actions 'eat it' and 'drink' always follow the action 'move hand towards

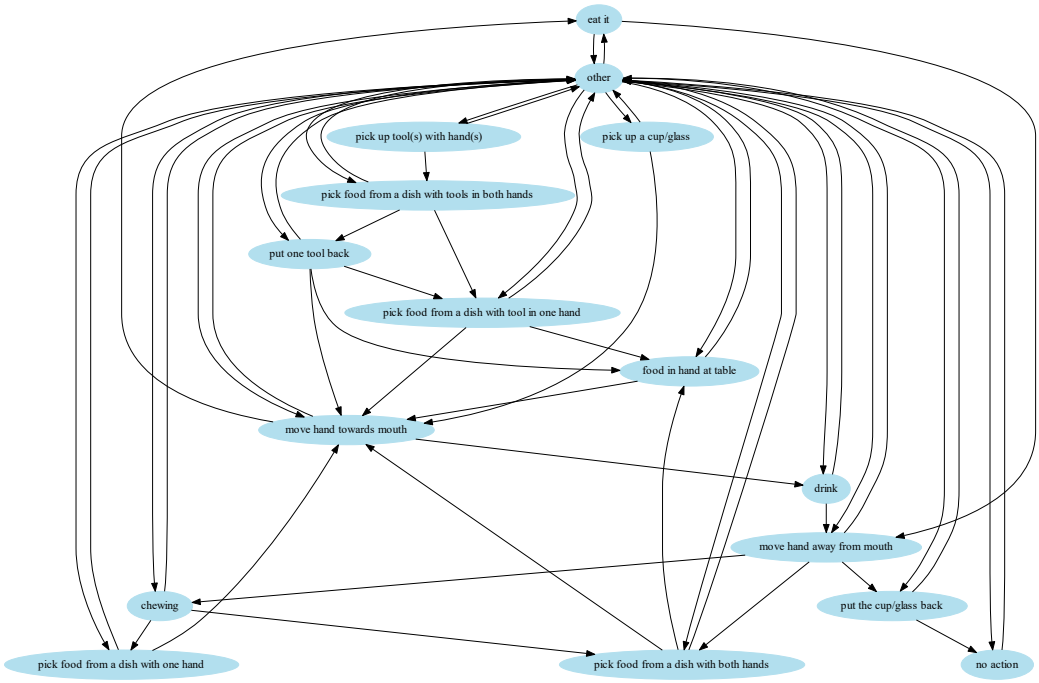


Figure 4.1: State diagram of common eating behavior with 16 action classes

mouth', and the action 'move hand away from mouth' comes next, as the eating behavior model shows. Given that the video recordings were obtained in an uncontrolled setting, the participants were allowed to converse and utilize mobile phones in the same manner as they normally would. As a result, the state diagram shows that the activity designated as 'other' can follow almost all actions.

4.4 Sub-Action Recognition: Deep Learning Based AR

For sub-action recognition, we divide the experiments based on the modality, i.e., single-modality experiments that utilize specific modalities like RGB, skeleton, and flow, and multi-modality experiments that combine RGB and flow. Also, it is assumed that the action instances are already trimmed[4]. By examining intra-action relationships without regard to past or future action occurrences, these techniques only identify the action that is currently in progress. The issue of determining when one action ends and the next one begins is not addressed in this section.

[4]Trimmed means that the full video has been divided into segments (called 'clips') that contain only a single primitive action.

4.4.1 Dataset Splits

Firstly, the data was split into clips of different activities to create a variety of actions for classification analysis. Secondly, stratified sampling was carried out on the action clips (no replacement) so that 5-fold cross-validated results could be produced. Third, five stratified splits were produced with the help of these sampled clips. For this sub-action recognition experiment, three of these five splits were used for training, and one split each for validation and testing. To perform five-fold cross-validation, the splits were permuted. For evaluation on graph-connected networks, a set of poses ($m \times 8 \times 3$ vector, where m is the number of frames) for each frame was also shuffled with the condition that the same set of actions is chosen as in the original first five splits, that use poses as the input.

4.4.2 Classification

A range of deep learning-based networks featuring distinct input modalities were assessed to classify the trimmed videos. Graph-convolutional networks (e.g., CTR-GCN [131], 2s-AGCN [130], ST-GCN [127]) and a 3D heatmap volume (e.g., PoseC-onv3D [126]) are examples of graph-based deep feature approaches that were used for skeleton-based classifiers. By using TANet [134], TPN [133], and TSN [132], we also show recognition from RGB, optical-flow (motion features), and combined RGB+Flow modalities. Additionally, a comparison was made between training the same algorithms from scratch and fine-tuning pre-trained algorithms.

The PyTorch implementation has been utilized for each of the algorithms used. Every algorithm was trained for 150 epochs. The learning rate was initially set at 1×10^{-2}, and it was multiplied by $\frac{1}{10}$ every 30^{th} epoch. These were empirically chosen. The remaining training protocols for the employed techniques were in line with those found in the original papers unless otherwise stated.

4.4.2.1 Results

In order to assess the effectiveness of modeling eating sub-actions, we calculate the macro (mean class accuracy) and Top-1 scores across all 16 classes. The top-1 (clip) and macro (class) accuracies of the networks using pre-trained models are displayed in Table 4.1, while Table 4.2 shows the performance when trained from scratch.

Table 4.1: The table shows the macro and top-1 accuracies achieved by CNNs with three modalities as input on the trimmed videos. The 'Pre-train dataset' column indicates the dataset on which that algorithm was pre-trained on. NTU-60 and Kin-400 refer to NTU-RGB-D-60 and Kinetics-400 respectively. The row 'Average' shows the mean accuracy achieved by the tested algorithms on each of the respective modalities.

Algorithm	Pre-Train Dataset	Modality	Top-1 Acc.	Macro Acc.
CTR-GCN [131]	**NTU-60**	**Pose**	**95.1**	**91.7**
PoseConv3D [126]	NTU-60	Pose	79.1	54.9
2s-AGCN [130]	NTU-60	Pose	82.6	66.3
ST-GCN 2D [127]	NTU-60	Pose	67.4	38
ST-GCN 3D [127]	NTU-60	Pose	89.8	71.9
Average		Pose	82.8	64.5
TANet [134]	**Kin-400**	**RGB**	**87.5**	**80.5**
TPN [133]	Kin-400	RGB	87.4	79.8
TSN [132]	Kin-400	RGB	83.3	70.0
Average		RGB	86.1	77.0
TANet [134]	Kin-400	Flow	84.5	**72.6**
TPN [133]	Kin-400	Flow	82.9	66.4
TSN [132]	**Kin-400**	**Flow**	**87.2**	72.3
Average		Flow	84.8	70.5
TANet [134]	Kin-400	RGB+Flow	88.3	80.1
TPN [133]	Kin-400	RGB+Flow	88.5	81.5
TSN [132]	**Kin-400**	**RGB+Flow**	**90.2**	**82.9**
Average		RGB+Flow	89.0	81.5

4.4.2.2 Discussion

In the evaluation of deep learning-based Action Recognition techniques for eating sub-action recognition, as depicted in Table 4.1, CTR-GCN emerges as the top performer among the compared models, leveraging its channel-wise topology to dynamically learn and effectively aggregate features. Conversely, ST-GCN 2D exhibits subpar performance, likely due to the lower-quality motion features inherent in 2D poses

Table 4.2: The table shows the macro and top-1 accuracies achieved by CNNs with three modalities as input on the clipped videos. 'None' means that the algorithms were trained from scratch. The row 'Average' shows the mean accuracy achieved by the tested algorithms on each of their respective modalities.

Algorithm	Pre-Train Dataset	Modality	Top-1 Acc.	Macro Acc.
CTR-GCN [131]	**None**	**Pose**	**96.1**	**93.5**
PoseConv3D [126]	None	Pose	79.2	56.2
2s-AGCN [130]	None	Pose	83.6	65.6
ST-GCN 2D [127]	None	Pose	45.7	25.5
ST-GCN 3D [127]	None	Pose	90.1	77.9
Average		Pose	78.9	63.7
TANet [134]	**None**	**RGB**	**83.6**	**72.6**
TPN [133]	None	RGB	83.5	68.5
TSN [132]	None	RGB	82.4	61.8
Average		RGB	83.1	67.6
TANet [134]	None	Flow	82.9	**68.5**
TPN [133]	None	Flow	81.7	64.8
TSN [132]	**None**	**Flow**	**83.9**	67.5
Average		Flow	82.8	66.9
TANet [134]	None	RGB+Flow	85.7	71.7
TPN [133]	None	RGB+Flow	86.2	74.9
TSN [132]	**None**	**RGB+Flow**	**88.7**	**79**
Average		RGB+Flow	86.8	75.2

compared to their 3D counterparts. Securing the second position, ST-GCN 3D demonstrates notably improved performance in terms of both top-1 and macro accuracy. In general, the majority of deep learning-based graph convolutional networks (GCNs) achieve accuracies exceeding 70% and class-wise accuracy surpassing 50%. Furthermore, when trained from scratch, as illustrated in Table 4.2, CTR-GCN maintains its superiority over other models.

For RGB as the modality, whether trained from scratch or pre-trained on Kinetics-400, both TANet and TPN achieve nearly identical top-1 and macro accuracy, with only

slight variations in performance. However, TANet, when trained from scratch, achieves the highest class-wise accuracy (macro) of 72.6% and top-1 accuracy of 83.6%. This superiority is attributed to TANet's specialization in capturing long-term temporal dependencies, a capability that TSN and TPN lack. Notably, TANet does not perform as well when only using optical flow as input. This discrepancy may stem from the nature of how optical flow encodes motion information in videos. Optical flow offers a dense representation of motion, whereas the motion information encoded in RGB frames is more sparse, making it potentially more challenging to extract accurately.

TSN is explicitly designed to model temporal information by segmenting the video and aggregating features from each segment. This makes it well-suited for processing dense motion information like optical flow, resulting in better performance compared to TANet or TPN when using optical flow as input. Moreover, TSN employs a multi-scale temporal sampling strategy, enabling it to capture temporal information at various scales, which is particularly advantageous for processing optical flow data encoding motion information at multiple spatial and temporal scales.

In contrast, for the mixed modality (RGB+Flow), the top-1 accuracy achieved by all three algorithms is comparable. However, TSN performs better overall in terms of both top-1 accuracy and macro accuracy. This is attributed to its ability to effectively capture the temporal evolution of actions by segmenting the clip into short segments and sampling frames from each segment.

Table 4.2 presents the performance of these algorithms when trained from scratch. TANet demonstrates superiority in the RGB modality, while TSN excels with optical flow. Nevertheless, the experimental results suggest that utilizing pre-trained models enhances the accuracies achieved by these baseline algorithms for all modalities, as indicated by the averages in Tables 4.1 and 4.2.

4.5 Sub-Action Recognition: Hand-Crafted Action Recognition

This section investigates sub-action recognition using engineered features that are derived from the eight upper body joints.

4.5.1 Descriptive Features

Following the estimation of 2D human pose using HigherHRNet, 2D poses were projected into 3D space using depth maps captured by the RGBD camera. Since the projected 3D points and estimated 2D poses are absolute positions, they often shift when the surroundings or the camera's position changes. We set the camera coordinate system's origin to the subject's chest joint's 3D location to avoid this problem. As stated in eq. 4.1, we calculate the relative positions of each joint by subtracting the 3D position of the chest from the 3D joint absolute position. In this case, the absolute 3D position of the joint j, is denoted by $\vec{p}_j^{abs}$, where $j = \{1,...,7\}$ i.e., each of the seven upper-body joints except the chest, and the absolute position of the chest is indicated by $\vec{p}_c^{abs}$. The relative joint positions are then used as a feature for classification tasks.

$$\vec{p}_j^{rel} = \vec{p}_j^{abs} - \vec{p}_c^{abs} \tag{4.1}$$

4.5.1.1 Additional Spatial Features

Joint motion during different sub-actions related to eating is highly correlated. In order to take advantage of the relationships between the two arms, the Euclidean distances between the two elbows and wrists were calculated using the formula in eq. 4.2, where i stands for the elbow/wrists joint, i.e., $i \in \{w,e\}$, and l or r for the left or right joint, respectively. The relative positions of the joints $\vec{p}_j^{rel}$ are transformed into a polar coordinate system as provided by eq. 4.3 to create a more meaningful representation of a 3D point in space. Here $p_j^{rel,x}$ denotes the x coordinate of the relative position of the joint j.

The product of the polar coordinates of the joints was calculated as shown in eq. 4.4 where p_j^{polar} represents the relative polar joint position excluding the chest joint (skeleton origin), to explore the effect of the movement of joints relative to each other.

$$d_i = ||\vec{p}_{r,i}^{rel} - \vec{p}_{l,i}^{rel}||_2 \tag{4.2}$$

$$p_j^{polar} = \sqrt{(p_j^{rel,x})^2 + (p_j^{rel,y})^2 + (p_j^{rel,z})^2} \tag{4.3}$$

$$p^{prod} = \prod_{j \in \{j=1,...,7 \text{ and } j \neq c\}} p_j^{polar} \tag{4.4}$$

To get the orientation of the arms at any time instance, joint triplet angles for elbow (denoted as e) and shoulder (denoted as s) sockets were calculated (w and c represent

wrists and body center) using the equations 4.4 and 4.5 respectively, where b denotes the left or right limbs of the body.

$$p^{\theta}_{b,e} = cos^{-1}\frac{(\vec{p}^{rel}_{b,s}-\vec{p}^{rel}_{b,e})\cdot(\vec{p}^{rel}_{b,w}-\vec{p}^{rel}_{b,e})}{\|\vec{p}^{rel}_{b,s}-\vec{p}^{rel}_{b,e}\|\times\|\vec{p}^{rel}_{b,e}-\vec{p}^{rel}_{b,w}\|}$$
$$p^{\theta}_{b,s} = cos^{-1}\frac{(\vec{p}^{rel}_{b,c}-\vec{p}^{rel}_{b,s})\cdot(\vec{p}^{rel}_{b,e}-\vec{p}^{rel}_{b,s})}{\|\vec{p}^{rel}_{b,c}-\vec{p}^{rel}_{b,s}\|\times\|\vec{p}^{rel}_{b,s}-\vec{p}^{rel}_{b,e}\|} \tag{4.5}$$

Finally, we determine the joints' distance from the table's plane to take advantage of the way the human posture interacts with stationary objects in the scene. Using 3D key points on the table and the least squares method, the plane of the table was estimated. These key points were manually marked on the table in each video's opening frame, and they were then carried throughout the footage under the presumption that the table is stationary and does not move during a single eating session.

4.5.1.2 Additional Temporal Features

For reliable action recognition, temporal features are essential; however, when motion quantification is required, these features become critical. For this purpose, we estimated velocities and acceleration. The measurements of instantaneous velocity and acceleration were found to be especially noisy because of the real-world recording environment and the lack of control over the subject's performance, or sequence of actions. So we extended this feature space further with different window sizes to account for noisy measurements. The joints' velocities and accelerations were estimated as per the equations 4.6 and 4.7, respectively. Joint (j) positions at frame $t+k$ are represented by the subscripts $t+k$, where k determines how many frames are included in the temporal estimation window, i.e., the window size.

$$\vec{v}_j = \vec{p}^{rel}_{t+k+1,j} - \vec{p}^{rel}_{t-k-1,j} \tag{4.6}$$

$$\vec{a}_j = \vec{p}^{rel}_{t+k+2,j} + \vec{p}^{rel}_{t-k-2,j} - 2\vec{p}^{rel}_{t,j} \tag{4.7}$$

It's critical to utilize the values from the recent past to emphasize the causality of the current posture. Three lag relative positions (that is, $\vec{p}^{rel}_{t-1,j}, \vec{p}^{rel}_{t-2,j}, \vec{p}^{rel}_{t-3,j}$) are also used for this purpose. Additionally, a feature that included a weighted sum over the three lags was added. To emphasize recent past values more, the weights of the last three values were set. Equation 4.8 provides the weights for the past values and also depicts the equation for the moving window. The moving window is denoted by *movw*, and its

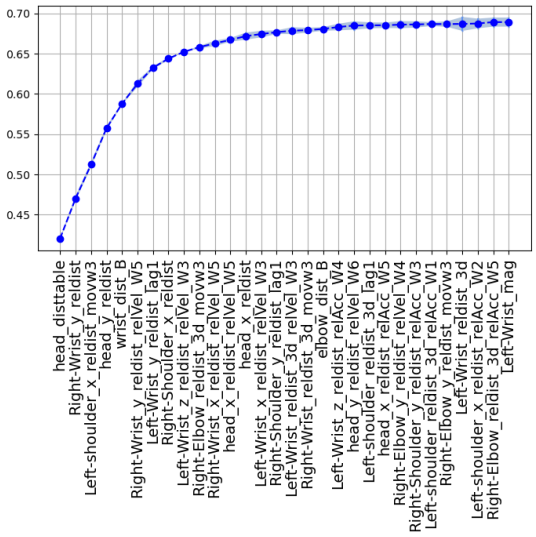


Figure 4.2: The forward sequential feature selection plots for frame-wise action recognition classification.

size is indicated by the subscript.

$$\vec{p}^{movw}_{3,j} = 0.5 \times \vec{p}^{rel}_{t-1,j} + 0.35 \times \vec{p}^{rel}_{t-2,j} + 0.15 \times \vec{p}^{rel}_{t-3,j} \tag{4.8}$$

Please note that the nomenclature and other aspects of the features are discussed in appendix A.1.

4.5.2 Classification with Hand Crafted Features

In a frame-by-frame setting, eating sub-action recognition using explainable (hand-crafted) features is investigated in this section. The skeleton's temporal and spatial domains are represented by the characteristics covered in the preceding paragraphs. There are numerous features, not all of which are helpful. To address the extensive array of features in the data, we employed a forward sequential feature selection search to identify the most contributing features. The top 30 features selected by the feature selector are depicted in Fig. 4.2. Based on these results, we determined that these 30 features are the most beneficial for classification and opted to utilize these for further experiments. The various classifiers discussed below used the vector representations of the chosen features as their input. The 16 distinct sub-actions that were labeled in the EatSense dataset were then classified using the selected features.

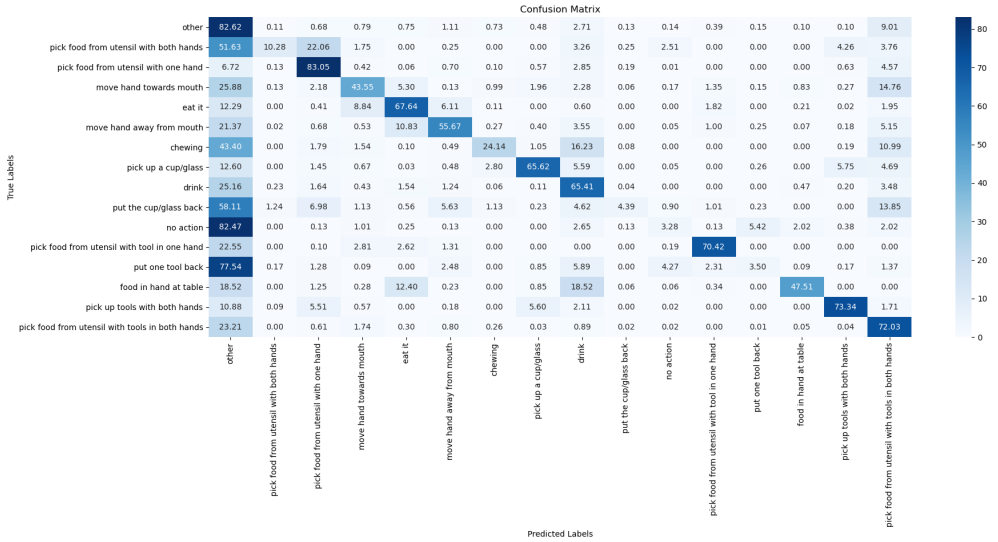


Figure 4.3: Confusion matrix for LGBM (FL) depicting percentages of (mis)classification of the classes. The rows correspond to the true data classes and the columns are the resulting classifications.]

4.5.2.1 Dataset Splits

Classification with the hand-crafted features used the same stratified sampling of the data as described in section 4.4.1. The entire dataset's clips were stratified and sampled (without replacement) into five subsets of the data splits, with each subset retaining almost the same proportion of each action's occurrence. These clips, however, were later divided into separate frames to conduct frame-by-frame analysis.

4.5.2.2 Classification

Several machine/deep learning techniques, such as Light Gradient Boosting Method (LightGBM) [146], Adaboost, Multi-layer Perceptron (MLP[5]), K-Nearest Neighbor (KNN), or Quadratic Discriminant Analysis (QDA) were investigated for classification using videos and hand-crafted features.

Additionally, LightGBM with the focal-loss objective function was tested because the number of classes in our dataset is unbalanced. Unless otherwise indicated, the hyper-parameters of the aforementioned techniques were set to their default values in sklearn (a machine learning module in Python programming language).

[5]A size 30 input layer (due to the number of the selected features, i.e., so the input matrix will be $30 \times n$ where n is the number of frames) is followed by four fully connected hidden layers with 50, 75, 45, and 25 neurons each, and then batch-normalization and RELU, in that order. Lastly, an output layer of size 16, as the output is equal to the number of classes, makes up the output layer.

Table 4.3: The table shows the frame-wise classification accuracy of testing on stratified splits and unseen full videos (marked as UV). Top-1 shows the frame-wise mean classification accuracy, Macro shows the mean classwise balanced accuracy and Std. shows the mean standard deviation (post 5-fold cross-validation). UV and FL stand for unseen videos and focal loss respectively.

Algorithm	5-fold Cross-Valid			Unseen Accuracy	
	Top-1	Macro	Std.	UV	± 2
MLP	67.7	47.7	1.3	22.2	31.3
N. Neighbors	59.5	43.1	1.2	32.6	42.1
Decision Tree	50.8	24.4	**0.7**	20.8	24.7
Random Forest	42.3	7.6	3.3	14.3	16.9
Neural Net	42.4	7.4	2.0	11.9	12.2
AdaBoost	47.3	24.4	1.7	25.7	34.2
Naive Bayes	25.7	17.2	1.4	7.3	11.7
QDA	27.5	23.1	3.3	26.2	38.2
LGBM	67.5	47.6	1.1	36.6	44.6
LGBM (FL)	**69.5**	**48.8**	0.9	**38.9**	**44.9**

4.5.2.3 Experiments

For frame-by-frame analysis, we conducted two experiments: (i) Training and testing on stratified splits using 5-fold cross-validation. (ii) Training on stratified splits and testing on unseen full videos. These experiments allowed us to thoroughly evaluate the performance of our hand-crafted features-based classification model under different conditions and assess its generalization capabilities on previously unseen data (data unused for training or testing).

4.5.2.4 Results

Table 4.3 presents the results of both experiments. The first three columns display the mean top-1 accuracy and standard deviation achieved by each algorithm with 5-fold cross-validation. The last two columns show the accuracy of the trained model when applied to unseen videos. Notably, LightGBM outperforms the other algorithms in

both experiments. Since the action ground truth is manually labeled, the start and end frame times of the actions may be inaccurate. To mitigate human labeling errors and address temporal classifier offsets in the labels, a windowed search was conducted. This search involved examining frames within a range of sizes ranging from ± 0 to ± 21 frames (equivalent to $\pm$ 0 seconds to $\pm$ 1.5 seconds) to identify the correct label within that frame range, between the ground truth and the predictions (discussed in detail in Appendix A.2). Table 4.3 presents the accuracy achieved with a window size of ± 2 (corresponding to $\pm$ 0.3 seconds).

Figure 4.3 shows the confusion matrix of the best performing classifier as compared to others, i.e., LGBM (FL). In the confusion matrix, the true label is represented by each row in the matrix, and the predicted label is represented by each column. The off-diagonal entries display the misclassifications, or instances where the model predicted the wrong class, whereas the diagonal entries indicate the number of accurate classifications.

For instance, with the first class 'Other', 82.62% (44,123 samples out of 53,404) samples were correctly classified, but the model misclassified a portion of them into other actions, such as 9.01% samples predicted as 'Pick food from utensil with tools in both hands'. Some classes show high confusion such as 'Drink', where 65.41% of samples were correctly classified, but significant confusion exists with other actions like 'Other' (25.16% samples misclassified). Hence, it is observed most of the actions are confused with the class 'Other'. This is because 'Other' is a very noisy action since it can contain all kinds of actions a human does on a table including, using a mobile phone and chatting (e.g., chatting is similar to 'No action' since it doesn't involve a lot of hand gestures).

Overall, the confusion matrix reveals areas where the model struggles, especially with actions that might have similar hand movements or objects involved, such as confusion between 'Move hand towards mouth' and 'Eat it'.

4.6 Comparison Summary of Hand-Crafted versus Deep-Learning based Features

In Tables 4.1 and 4.2, the focus is on the top-1 and macro accuracies achieved by CNN-based models using different input modalities (Pose, RGB, Flow, RGB+Flow) on trimmed video snippets for video-clip classification. Table 4.3 on the other hand,

presents frame-wise classification top-1 accuracy of different algorithms on stratified splits and unseen full videos. Although both hand-crafted and deep-learning based feature analysis utilize the same train and test set, they do not have identical input information, i.e., the hand-crafted analysis uses framewise-spatiotemporal features whereas deep-learning based analysis utilizes deep features extracted from entire action clips (either RGB, pose, or flow) as an input.

For a very superficial comparison, if we ignore the fact of different inputs and focus only on top-1 accuracies achieved by the models, the CNN-based models in Tables 4.1 and 4.2 consistently achieve higher accuracies compared to the models in Table 4.3.

4.7 Temporal Action Localization

Temporal action localization (TAL) allows for a detailed analysis of the sequence, duration of actions, and start and end of action instances, which is vital for understanding complex behaviors and their context. These temporal action localization algorithms leverage images to analyze inter-action temporal relationships. One of the primary contributions of EatSense lies in its dense labeling approach, ensuring that there are no unlabeled segments within the approximately 11.5-minute-long videos, on average. These characteristics open up ways to explore temporal action localization techniques for understanding the eating behaviors of individuals using the EatSense dataset.

4.7.1 TAL using EatSense

In this experiment, we perform TAL using the EatSense dataset to understand if we can localize densely annotated sub-actions that have hugely varying lengths (from less than a second to 9 seconds on average) in a video sequence. We use BMN [140], ActionFormer [141] and TadTR [142] to perform localization with all 16 sub-actions within the EatSense dataset. Most of the TAL approaches require a feature embedding estimated prior to training the models since training the model directly with the videos has a significant computational overhead. For this purpose, we extract the feature embedding using specialized deep networks such as TSN [132] and TSP [145] that emphasize more on making the boundaries of the action instances more prominent in the untrimmed videos. We used the author's original code base to perform TAL in EatSense.

4.7.1.1 Dataset Splits

The 135 untrimmed EatSense videos were randomly divided into three splits, training on 96, validation on 24, and testing on 15 videos.

4.7.1.2 Action Localizers

Videos (typically in non-overlapping snippets) are fed into an action localization framework's (pre-trained) visual encoder, which represents the video as a high-level feature set before being further processed for action recognition and localization. Utilizing the TSN [132] and TSP [145] pipelines, we extracted visually encoded features for tasks related to temporal localization (determining when the video stream switches from one activity to another). In order to obtain the dense high-level feature set needed to detect atomic actions for EatSense, we employed overlapping snippets.

Feature estimation for this TAL usually requires some hyper-parameters to be set according to the dataset. Two hyper-parameters that were particularly important were feature stride (skip frames in the video sequence, i.e., stride of 1 means no skipped frames, and stride of 2 means use every other frame) and clip length (number of frames, i.e., length of the video snippet). We empirically chose these hyper-parameters and generated TSP features using a stride of 2 and clip-length of 5. Then, using BMN [140], ActionFormer [141], TadTR [142], and TriDet [143] these high-level features were utilized to produce segment proposals for evaluation using the EatSense dataset. We also tested TadTR with hand-crafted (HC) features. Ultimately, TSN (trained and evaluated as in section 4.4.2) was used to classify the proposals produced by BMN, while ActionFormer implicitly categorized the generated proposals into action classes.

4.7.1.3 Results

Each of the TAL networks outputs temporal segments, i.e., t_s and t_e (start and end of the segment) with the corresponding action class and its confidence score. Afterward, Non-Maximal Suppression (NMS) helps eliminate redundant action proposals by retaining only the most confident detections and removing those with significant temporal overlap (based on temporal intersection over union) and lower confidence scores. This process enhances the precision of the model by ensuring that only the most relevant action segments are kept hence reducing false positives in the detection set.

Mean average precision (mAP) is used as a metric for the performance evaluation

on EatSense for action localization tasks. It measures how accurately a network can identify both the occurrence and temporal boundaries of specific actions within a video sequence. The metric combines precision (the fraction of correctly identified action instances among all predicted instances) and recall (the fraction of correctly identified action instances among all actual instances) across various confidence thresholds. The area under the precision-recall curve is calculated to get the AP for that specific action class and mAP is obtained by averaging the AP values across all action classes. The results are shown in table 4.4, where @0.1, @0.3, and @0.5 denote the temporal IoU (tIoU) threshold levels.

Table 4.4: The table depicts the mean average precision (mAP) for the action localization task using CNNs on untrimmed videos.

		Mean Average Precision			
Algorithm	Feature	@0.1	@0.3	@0.5	Avg.
TadTR [142]	TSP	**15.41**	7.84	**4.80**	**9.35**
TadTR [142]	HC	2.25	0.93	0.27	1.15
TriDet [143]	TSP	6.76	3.6	0.41	3.59
ActionFormer [141]	TSP	14.04	**7.91**	3.43	8.46
BMN [140] + TSN	TSN	2.27	1.12	0.60	1.33

4.7.1.4 Discussion

Table 4.4 compares the performance of four algorithms, TadTR with TSP and handcrafted features, TriDet, and ActionFormer with TSP features and BMN + TSN with TSN features, in terms of Mean Average Precision (mAP) at IoU thresholds of 0.1, 0.3, and 0.5, as well as their average mAP scores. TadTR (with TSP features) has the highest average mAP 9.35 and excels at the 0.1 and 0.5 thresholds with scores of 15.41 and 4.80, respectively. ActionFormer shows the second-best performance with an average mAP of 8.46, performing best at the 0.3 threshold of 7.91. TriDet, with TSP features, has a moderate average mAP of 3.59, while BMN + TSN, using TSN features, achieves an average mAP of 1.33 across all thresholds. TadTR with handcrafted feature space performs the worst with 1.15 mAP.

As shown in table 4.4, the untrimmed videos in EatSense pose special difficulties for current action localization networks. As tIoUs increase, localization algorithm

performance decreases. This is because higher tIoU thresholds demand more precise localization of actions within a video. Small deviations in the predicted start and end times of an action can significantly reduce the overlap with the ground truth. Another possibility is that the ground truth annotations may have some level of inconsistency, and higher tIoU thresholds leave less room for these variations. This means that even well-performing models might suffer in evaluation due to slight discrepancies between the model's predictions and the ground truth. Thus, human labeling error can affect performance metrics such as mAP even more significantly for smaller actions (the smallest action in EatSense lasts only 0.62 seconds). Hence, EatSense is a very challenging dataset for temporal action localization frameworks.

Also, to the best of our knowledge, there are no specialized off-the-shelf skeleton-based temporal action localization frameworks available, so we only performed temporal action localization using RGB data in EatSense.

4.7.2 Analysis of TAL using TadTR and EatSense

In the previous section, we explored temporal action location networks and observed that the performance of the TAL algorithms is below par when used with the EatSense dataset. Three potential causes for their under-performance are: (1) the smaller actions are too small to detect and localize, (2) large actions (such as 'other') are noisy and hence indistinguishable from each other, and (3) a combination of these hugely varying lengths and noisy actions makes the performance go down. The research question that we aim to explore with this experiment is: how accurately can we localize actions if we divide the action classes based on how long an action lasts, i.e., into three categories of 'Short', 'Medium', and 'Long' action instances?

In this section, we analyze TadTR (one of the TAL networks) in detail using Hand-crafted features extracted from EatSense. We chose TadTR for this analysis because it performed slightly better as compared to others with TSP features. However, we chose hand-crafted features since it gives us more flexibility and help us understand the root cause of the bad performance of the TAL networks.

4.7.2.1 Dataset Splits

In this experiment, firstly, we extracted the top 30 most contributing features via feature selection using LightGBM as a frame-wise classification with macro-accuracy as the metric. Secondly, similar to the previous experiment we randomly divided the dataset

into three splits, i.e., training on 96, validation on 24, and testing on 15 videos.

4.7.2.2 Experimental Setup

We divided the actions into three categories based on their average length. The average time of the actions was given in chapter 3 Table 3.2. We classified the actions that last less than 2 seconds on average as 'Short' (7 action classes), actions that last between 2 and 4 as 'Medium' (4 action classes), and anything beyond this as 'Long' (5 action classes). The whole video contains all the action instances, but for this experiment, we restricted our labels strictly to only the actions under observation, and the rest of them were clipped out. We used each of the categories of actions to train a separate TadTR model. After careful observations, we found that the action class 'other' is the noisiest, and 'no action' occurs a few times in the whole dataset. So, to see the performance of TadTR without these actions, we clipped out 'other' and 'no action' classes and performed two tests named, 'Long*' where all the actions in the 'Long' category are included besides these two, and 'Mixed' where these two actions were clipped out and the rest of the 14 action classes were included.

4.7.2.3 Results

The top 3 rows of the table 4.5 show the results of TadTR after independently training three different models and testing them with the same videos as originally selected. However, we also clipped out the irrelevant action classes according to their group in the test set, e.g., a testing model trained for 'Short' instances was tested on re-compiled videos that only contained short action instances, and the rest of them were clipped out. The second-last row shows the results of the performance of TadTR over long action instances after excluding the 'other' and 'no action' classes. The last row (marked as 'Mixed' in the category column) shows the results achieved by TadTR if we localize 14 actions and leave the two actions that are the noisiest (i.e., 'other') and less frequent (i.e., 'no action').

4.7.2.4 Discussion

Table 4.5 compares the performance of TadTR when the action classes in EatSense are categorized into 'Short', 'Medium', and 'Long' instances based on the average length of action. In the table, we see that TadTR performs the best with 'Medium'. It achieves almost 5% more accuracy than the 'Short' action category. This difference in

Table 4.5: The table gives the mean average precision (mAP) of TadTR on 'Short', 'Medium', and 'Long' action instances. 'Long*' means long actions excluding 'other' and 'no action'. The 'Mixed' category shows the mAP achieved when temporal localization is performed on 14 actions (excluding 'other' and 'no action') instead of 16.

		Mean Average Precision			
Algorithm	Category	@0.1	@0.3	@0.5	Avg.
TadTR [142]	Short	31.67	23.57	13.93	23.05
TadTR [142]	Medium	36.67	29.76	17.74	28.06
TadTR [142]	Long	17.74	11.14	4.23	11.03
TadTR [142]	Long*	30.00	21.28	7.95	19.73
TadTR [142]	Mixed	27.35	21.03	13	20.46

the performance of the 'Short' and the 'Medium' categories is believed to be because of the high sensitivity of the metric (mean average precision) being used.

On the other hand, TadTR achieves the lowest mAP with 'Long' action instances. These 'Long' instances include actions that are believed to be the most sporadic, (i.e., 'other') and less frequent (i.e., 'no action'). We can observe the performance difference caused by these two actions in the row marked as 'Long*'. The accuracy achieved when these two actions were excluded is almost 9% higher than when these two action classes were included. This can also be validated, shown by the category 'Mixed' in the table when we tested TadTR's performance with all 14 actions and excluded these two actions and it achieved 20.46 mAP (as compared to 1.12 for all 16 actions as shown in the Table 4.4).

4.8 Conclusion

In the evaluation of deep learning-based action recognition (AR) techniques for eating sub-action recognition, CTR-GCN emerges as the top performer among the compared models, leveraging its channel-wise topology to dynamically learn and effectively aggregate features. Conversely, ST-GCN 2D exhibits subpar performance, likely due to the lower-quality motion features inherent in 2D poses compared to their 3D counterparts. Overall, the experimental results suggest that utilizing pre-trained models enhances the accuracies achieved by these baseline algorithms for all modalities.

Furthermore, we also demonstrated the use of hand-crafted features for AR using EatSense. This was to emphasize the use of hand-crafted features for explainable applications. The experimental results indicate that LightGBM outperforms other algorithms in both mean top-1 accuracy and the accuracy of models when applied to unseen videos, as measured with 5-fold cross-validation. Since this is a frame-by-frame analysis, to minimize the human labeling errors and address temporal classifier offsets, a windowed search was conducted, examining frames within a range of $\pm$ 1 to $\pm$ 21 frames (equivalent to $\pm$ 0.2 seconds to $\pm$ 3 seconds) to better align predictions with the ground truth. This windowed search was carried out on a video sequence, rather than action snippets and overall it improved the performance at least by 6% for AdaBoost, QDA, and LightGBM.

We also explored four temporal action localization frameworks (BMN, ActionFormer, TriDet, and TadTR) with EatSense's untrimmed videos utilizing TSN/TSP/HC feature embeddings. Overall, their performance was below par for localizing all 16 actions in EatSense, especially at higher tIoU thresholds. This was potentially due to two reasons: 1) the huge variability of the length of the temporal segments (the lengths vary from 0.62 to 9 seconds on average) for each of the action classes present in EatSense, 2) mAP's high sensitivity to human ground-truth labeling errors. This indicates that EatSense is more challenging for localization than other publicly available datasets. This underscores the need for developing new approaches for untrimmed video understanding that are more robust in terms of handling actions of varying lengths.

Lastly, to understand if the performance can be improved if we restrict our train and test sets on a subset of actions with similar duration, we see an increase in the accuracy. These results indicate that TadTR performs best on medium-duration actions and struggles with longer actions, especially at higher IoU thresholds, but performance improves when noisy actions are excluded from consideration.

Chapter 5

Quality of Motion Assessment

In this chapter, firstly we demonstrate the effectiveness of simulating performance decline by attaching different weights to the wrists of the subjects. Secondly, since working with humans is challenging, as eating characteristics vary from person to person, we propose a two-step approach to obtain a generalized model using distinguishable micro-movements.

5.1 Introduction

EatSense, as introduced in Chapter 3, is a human-centric, upper-body-focused dataset that supports the modeling of eating behavior and the investigation of changes in motion/motor decline (i.e., quality of motion assessment). Four weights (0,1,1.8 and 2.4kg) are put on the volunteers' wrists while they eat to simulate a change in motion and performance decline. These subjects belong to various ethnicities, genders, and age groups.

In this chapter, we explore three main research questions:

1. Does using weights effectively simulate a decline in performance over time?

2. Can we distinguish minor changes in motion induced by the four different weight levels?

3. Can we develop generalized models over all age groups for performance decline classification/regression as there might not be any consistent pattern to exploit across all subjects, i.e., can we develop a model that can be applied to new volunteers?

To answer these questions, firstly, we demonstrate through trunk stability and speed of movement tests that decline in performance is observable and the weights can indeed be used to simulate performance decline, i.e., when weights of different levels are attached to the wrists of the subjects (Section 5.4). Secondly, we demonstrate through a t-SNE plot that we can indeed distinguish minor changes in motion (Section 5.5.2.2). Lastly, we explore the generalization ability of existing models with strictly explainable features across various subjects in all age groups (Section 5.5).

The contributions of this chapter are:

- The first computer vision-based quality of motion assessment quantitative approach solely based on the eating behavior of individual subjects.
- We address the most common problem of lack of generalizability when it comes to modeling human behavior (using existing models, limited to the performance of eating assessment using only EatSense in our case). (Section 5.5).
- Demonstrate that 4 weight classes simulate performance decline in the upper-body movements.

5.2 Literature Review

A brief review of past clinical and sensor-based techniques for performance decline assessment and behavior analysis is presented.

5.2.1 Performance Decline Assessment Tests

Several studies [147], [148], have proposed a series of tests in a clinical context to observe degradation in functional motor movements. Alonso et. al. [149] provide an overview of computerized techniques like 'Equitest' and 'Force Platforms' as well as clinical tests like 'timed up and go' and 'Functional Reach Test' for evaluating balance.

Studies investigating motion tracking and assessment methods based on magnetometers or inertial measurement units (IMUs) have been conducted in non-clinical settings. In [150], Filippeschi et al. compare IMU-based human motion tracking methods, concentrating on upper-body limbs that may be relevant for motion evaluation. Carnevale et al. [151] examined the evaluation of shoulder kinematics using wearable sensors following musculoskeletal injuries or neurological trauma. Meng et

al. [152] recently demonstrated an IMU-based upper limb motion assessment model with promising outcomes.

Numerous findings in the area of vision-based healthcare have also been reported in non-clinical settings: (1) motion tracking; (2) fall detection [153]; (3) anomaly in gait detection [154], [155]; and (4) exercises that support the rehabilitation of patients recuperating from illnesses that adversely affect their level of activity [156], [157], [158].

To illustrate the efficacy of their suggested method, Nalci et al. [159] compared their computer vision-based functional balance alternative test to the BTrackS Balance Assessment Board, which is utilized in clinical examinations, and concluded that the proposed system achieves a high degree of correlation with BTrackS. A single camera was utilized by Yang et al. [160] to track joint markers on upper-body limbs, perform data analytics for the computation of rehabilitation parameters, and offer a reliable classification appropriate for home healthcare. The device was designed to be portable and affordable. A real-time risk assessment rapid upper-body limb assessment tool utilizing cameras (depth or RGB) to detect abnormal postures in real-time and offline analysis was proposed by the authors of [161], [162], and [163].

Recently, a vision-based balance assessment test while sitting was proposed by Barlett et al. [164]. To the best of our knowledge, no vision-based study has specifically examined performance decline in relation to upper-body limb movement in the human pose.

5.2.2 Behavior Analysis

The phrase ‘human behavior analysis’ covers a wide range of topics, including activity recognition, facial expression analysis, and gesture recognition. Activity recognition-based behavior analysis algorithms, according to Onofri [96], require knowledge that falls into two categories: contextual knowledge and prior knowledge. The environment in which an activity is occurring, including the objects involved and the time and location, is referred to as contextual knowledge. Prior knowledge refers to the recognition system’s awareness of historical information, such as the fact that event C usually occurs after event B and that there is very little chance that event C will occur after event A.

Research on human motion in sports games has been done extensively [165], [166], [167], as well as in other contexts [168], [54], [169]. It is common practice to quantify

and assess behaviors by combining attributes of the human body such as position, distance, speed, acceleration, motion type, and time. Tennis trainees' primary poses' spatial, rotational, and temporal properties were retrieved by Oshita et al. [166], who then compared the trainees' workout regimens with those of experts.

In [170], Yordanova et al. introduced a technique for identifying human behavior called Computational Causal Behavior Models (CCBM) to track an individual's everyday kitchen activities. This examined a person's behaviors, the kind of food they are cooking, and any possible health consequences by combining a symbolic description of the person's behavior with probabilistic inference. Using inertial signals from a smartwatch, Kyritsis et al. [93] presented an algorithm that can automatically identify cycles in food consumption that take place throughout a meal. They model the sequence of actions needed in eating by using five distinct wrist micro-movements: 'pick food', 'upward', 'downward', 'mouth', and 'other movements'.

Prior studies using eating actions, such as [171] and [138], were primarily conducted to categorize eating and drinking actions to better understand individual actions. Conversely, Tufano et al. provide a thorough comparison study of 13 frameworks, including optical flow-based and deep learning frameworks, in [172]. For this comparative analysis, they focused on three distinct eating behaviors, i.e., bites, chews, and swallows.

Nevertheless, we are not aware of any prior research that evaluates the quality of motion based on behavioral traits and addresses the generalization problems across eating actions.

5.3 Features Useful for Behavior Quality Assessment

5.3.1 Hand-Crafted Features

The goal of investigating manually designed feature-based methods is to gain a comprehensive grasp of each subject's health. Complex features make it difficult for medical experts to identify the underlying causes of patients' health issues.

All of the individual frames are used to extract the suggested features. The features include instantaneous spatial characteristics such as (1) angles between the elbows and shoulders, (2) product of all joints, (3) distance from the table of all joints, and (4) relative distances of all joint positions with respect to the chest. Additionally, temporal features like (1) velocity, (2) acceleration, and (3) lags are included (past instantaneous

joint position; for example, if the joint position at time t is denoted as $\mathbf{x}_t$ and the joint position in the previous frame taken at time $t-n$ is denoted as $\mathbf{x}_{t-n}$ which is the n^{th} lag.) (4) A weighted sum of the last three lags. Section 4.5.1 discusses the mathematical formulation of each of these features.

5.3.2 Deep Features

A Spatial-Temporal Graph Convolutional Network (ST-GCN) [127] was used to extract deep features from the EatSense videos. Unlike the manually created features, in this method, we only need the subjects' 3D poses. In contrast to the frame-by-frame feature extraction, we take into account a whole action spanning multiple frames to utilize both temporal and spatial properties to build a graph. Graph convolutions are then used to estimate high-level feature mappings on the constructed graph.

5.4 Performance Decline Simulation

This section demonstrates the effectiveness of adding weights to the individuals' wrists to simulate performance decline. For this purpose, experimentally proven tests such as the balance assessment and speed of motion tests are used. These tests were modified in light of the necessity to investigate degradation in an eating setting. The plots and these minor modifications are described in the following sections.

5.4.1 Balance Assessment Test

The Balance Assessment Test [164], [173], also referred to as the postural sway or trunk stability test [174], measures how well the participant keeps their body's center of mass within its base support. In clinical trials, this is done while the subject is standing; however, in this case, the test is conducted while the subject is sitting for a complete meal for approximately 6 to 10 minutes. In every video, the participant is recorded wearing weights ranging from 0 to 2.4 kg.

We estimate the feature 'the distance of the chest from the table' (explained in Section 5.3.1) for each frame based on the subject's 3D pose to identify any sway in the subject's posture. We temporally stack results from the videos by the increasing value of the weights as they are captured with subjects wearing weights. For uniformity, the data from the left-handed volunteers was rotated around the y-axis.

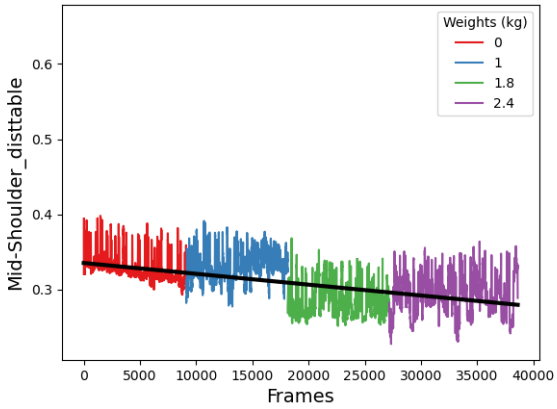


Figure 5.1: Demonstrates negative slope (i.e. increasing slouch) in the context of the proposed experiment. 20% of the recorded video frames were sampled randomly from each of the 4 weight cases of subject no. 4. These frames are then subsequently arranged in order of increasing weights.

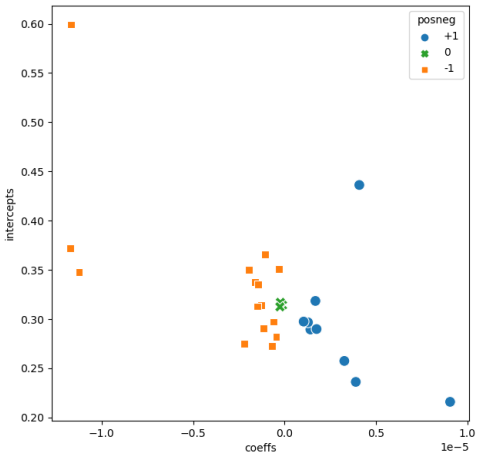


Figure 5.2: Balance Assessment Test. +1 (blue) represents subjects with positive slopes, -1 (orange) represents subjects with negative slopes, and 0 (green), indicates a change in their trunk positions, i.e., the subjects started with an upright posture but over time as the weight is increased, their chest position changed. See the text for more discussion.

Figure 5.1 shows the mid-shoulder to table distance from one subject. The colors (red, blue, green and purple) highlight the four videos stacked one after another. A line is fitted by linear regression through the temporal data, which consists of videos in the increasing order of the weights. The expected line, which is highlighted in black, has a negative slope. A negative slope indicates a reduction in distance from the table when weights are increased. As a result, the experiment's negative slope shows that performance decline when weights are added.

A positive slope ought to suggest that the posture improved over time, while a negative slope suggests that the core/trunk position deteriorated over time. The association between slope coefficients and intercepts is plotted in (Fig. 5.2), where 0 (green) indicates no discernible change in the trunk position, +1 (blue) indicates positive slopes, and -1 (orange) indicates negative slopes. Here, visible change is quantified and denoted by an orange or blue indicator depending on whether the coefficients exceed or fall below $\pm 0.03 \times 10^{-5}$. The plot shows that 15 out of 27 subjects have negative slopes. The fact that they were unable to remain upright suggests a weakening of the core. However, some subjects show an upward trend, which makes us believe that this is what happens when they try to compensate for the weights by re-adjusting their balance.

5.4.2 Speed of Motion Test

The speed of motion test measures the rate at which an individual completes a task as part of their daily routine in order to track the aging-related deterioration of muscle. Sarcopenia is the term for the age-related loss of muscle function [175], [176]. In this study, the speeds at which the upper body limbs move are monitored to measure the gradual loss of muscle strength over time, i.e., performance decline, which is simulated using varying weight levels.

Initially, since the dataset comprises several sub-actions, most of which entail unpredictable motion orders, only the sub-action 'move hand towards mouth' is examined since it is the primary micro-movement involving motion against gravity. To do this, we use the wrist joint position relative to the chest distance (explained in Section 5.3.1) to estimate the dominant hand's velocity (via inter-frame position differences). (For uniformity, the left-handed volunteers were rotated around the y-axis.) The wrist velocities are measured in the increasing sequence of the weights, much like in the postural sway test. Using linear regression, a line is fitted through each subject's speed

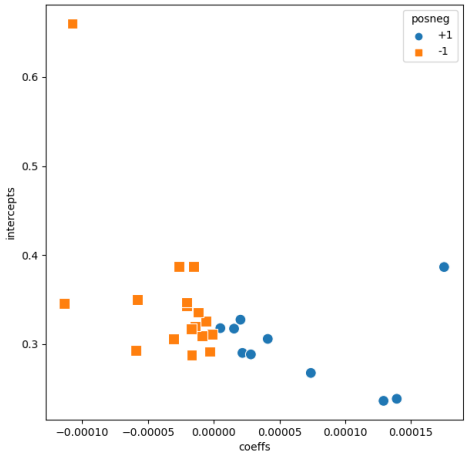


Figure 5.3: Speed of Motion Test. (blue) represents subjects with an increase in hand speed (positive slope) as the weight is increased and (orange) represents subjects with a decrease in hand speed (negative slope). See the text for more discussion.

vs. weight curve.

It is anticipated that the slopes will be negative in order to show that the upward movement speed is decreasing. Fig. 5.3, presents a scatter plot showing the relationship between slope coefficients and intercepts. It shows that 17 out of 27 volunteers show a decrease in motion speeds throughout different weight classes. On the other hand, the participants who report leading an active lifestyle are primarily those who exhibit either positive or neutral trends in the data.

5.4.3 Effect of weights by Age and Gender

In this section, we utilize the results of the balance assessment (BAT) and speed of motion tests (SMT) to explore the effect of weights different age groups and gender (male and female).

Table 5.1 provides a summary of the performance decline trends in two physical assessment tests conducted across two demographic groups, categorized by gender (male and female) and divided into various age groups. The table specifically highlights the number of individuals who demonstrated a negative slope trend in their performance, indicating a decline in their physical abilities.

If we look at the subjects over the age of 50, predominantly (3 out of 4) female subjects show a decline in their performance in both tests. Also, if we look at subjects

Table 5.1: The statistics for both balance assessment and speed of motion tests, on two different demographics, i.e., gender (male (M) and female (F)) and age groups, of the subjects involved in the EatSense dataset. The numbers in each of the columns in 'Balance Test' and 'Speed Test' are the numbers of subjects who showed negative slope trends, i.e., a performance decline, followed by the percentage of subjects showing that effect.

	Balance Test		Speed Test		Total	
	Decline		Decline		Subjects	
Age Gr.	M:%	F:%	M:%	F:%	M	F
<30	3:38	3:100	5:63	2:66	8	3
30-39	2:50	1:50	2:50	0:0	4	2
40-49	0:0	2:100	1:50	2:100	1	2
50-59	0:0	2:67	0:0	2:67	0	3
60+	1:33	1:100	2:67	1:100	3	1
Total	6:38	9:82	10:63	7:64	16	11

aged below 50, more than 50% (i.e., 85% and 57%) of females show a decline in both tests whereas a maximum of 61% of males show a decline only in SMT. When analyzing the total performance decline across all age groups, it is evident that males and females are affected differently by the two tests. For the Balance Test, 6 males and 9 females experienced a decline, indicating that females are more likely to have challenges maintaining balance over time.

5.4.4 Discussion and Conclusion

The results from the tests in the two sub-sections 5.4.1 and 5.4.2 above show that the performance declines with heavier weights for over 60% of the subjects. After dividing the subjects into age groups and genders (male and female), we see higher percentage of females experiencing a performance decline. Collectively, in these tests, 12 volunteers show a decline in both tests evidencing that the weights are effectively simulating performance decline.

The analysis could be extended by clustering using different demographics such as ethnicity, height, and weight (not currently available in the EatSense dataset). This

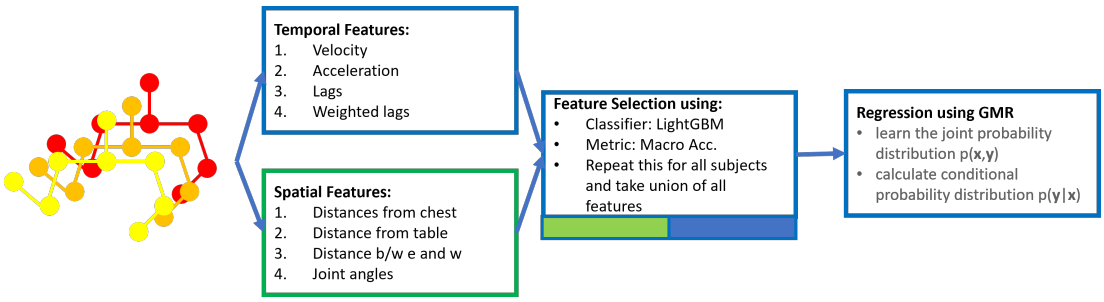


Figure 5.4: The complete pipeline of the proposed regression approach.

could be interesting future work but is out of the scope of this thesis.

5.5 Generalized Regression

The EatSense dataset has weights (0 kg, 1 kg, 1.8 kg, and 2.4 kg) on the individuals' wrists to simulate performance decline. To predict performance as the weights are increased, a motion model with a shared set of characteristics should ideally already exist. However, the volunteers appear to respond in different ways if we add weights to their wrists, some appear to slouch more, while others alter their posture. This section discusses the problem statement with an experiment to depict subjective bias in the parameters of performance decline and the experiments to design and test a pipeline that caters to this problem of generalized regression. Moreover, for comparison purposes, we split our pipeline tests into two sub-experiments, namely deep features-based and hand-crafted features-based regression, to predict how performance varies with weight level.

5.5.1 Problem Statement

To understand the underlying problem of generalization in EatSense, we pick out the most contributing features for the classification (using LightGBM) of the performance into the four performance decline assessment levels (0: 0kg, 1: 1kg, 2: 1.8kg, 3: 2.4kg) using forward sequential feature selector. The most contributing features are shown on the right side of Fig. 5.5. Since these features are engineered using domain knowledge, each of these features characterizes the performance of individuals. Since the plot shows the performance over the test set containing multiple individuals, we can conclude that these features are the most important to characterize performance generically across individuals.

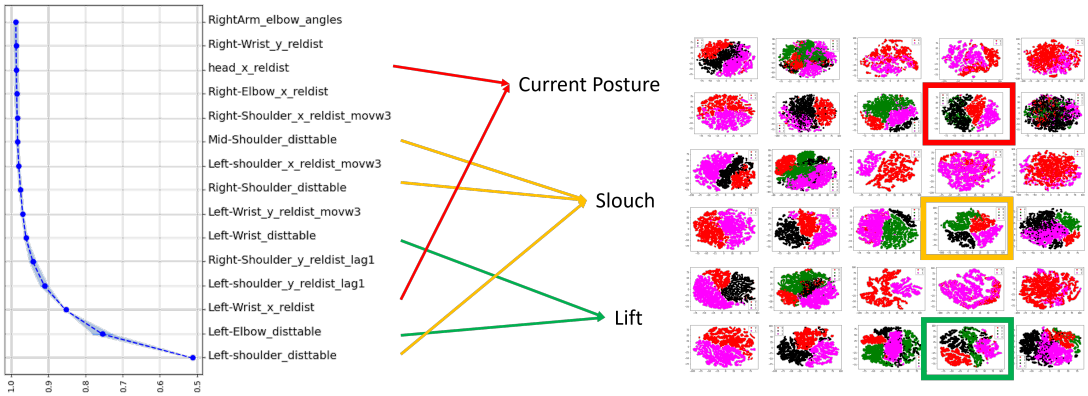


Figure 5.5: Features that parametrize performance decline. The left figure shows the features selected through the forward sequential feature selection process by posing four weights as a four-class classification problem. The plot shows the increase in accuracy as additional classification features are added. On the right, we show the t-SNE plots for the first ten subjects for each of the three parameters that characterize performance decline such as lift and slouch. These parameters are deduced from the selected set of features using domain knowledge (see text for more details). The colors in the t-SNE represent four weight categories, i.e., red:0, black:1, pink:2 and green:3.

If we translate these features back to real-world descriptions, we can understand what parameterizes the performance decline in the subject's performance. For example: 1) Features such as distance from the table for the left wrist and left elbow account for the **lift of the non-dominant arm**. 2) When a person eats or has difficulty in taking their arms to the mouth, they will potentially slouch. So, features such as distance from the table of shoulders (either left-shoulder, right-shoulder, or chest) can show **slouch**. 3) Features such as the instantaneous position of the limbs make up the **current posture** at time t. Similarly, other features listed in Fig. 5.5 also correspond to other metrics such as stretch and posture at time $t-1$.

To further illustrate, we estimated t-SNE plots after we combined features such as the distance from the table of left elbow and left wrist that make up for lift (shown with green arrows in the figure). We did the same for other parameters (i.e., slouch and current posture are shown by amber and red arrows respectively) listed above. The 10 t-SNE plots on the left represent the 10 subjects with t-SNE in 2D using only the features tagged by the respective arrows. The t-SNE plots in Fig. 5.5 show that all 10 individuals behave differently under the 3 different performance decline parameters.

Separability in a t-SNE shows how effectively we might be able to differentiate

between different categories. Fig. 5.5 shows some subjects have separable t-SNE for one or two but not all three characterizations. For example, subject 9 (highlighted with a box around the t-SNE plot), shows a separable t-SNE plot for 'slouch' (i.e. all 4 colors are visible) but is not separable with the other parameters. Hence, parameters for quality assessment are highly subject-dependent. This suggests that each of the subjects experience performance decline differently and a common set of features that parameterize the decline does not exist. Therefore, it is challenging to identify a set of characteristics and a model that parametrizes the process of performance change without overfitting on a subset of participants.

5.5.2 Hand-Crafted Features-Based Regression

From joint positions, both temporal and spatial characteristics were estimated. These were briefly discussed in Section 5.3.1. The main goal of exploring manually designed feature-based methods is to obtain a thorough grasp of each subject's health in an explainable manner. Through the application of these methodologies, researchers and medical practitioners can acquire a comprehensive understanding of diverse facets of an individual's well-being. Deep features, on the other hand, have not yet shown to be as effective in offering understandable explanations, despite their strength in representing complex patterns and relationships in data. In the realm of health, it can be difficult for medical personnel to gain a thorough understanding of the elements affecting a patient's health due to the intricacy of deep features.

5.5.2.1 Feature Selection

A forward sequential feature selector (FSFS) was employed [177] with LightGBM [178] serving as the classifier of the four classes of varying weights in subsets of the dataset to choose a common subset of features across all subjects. Assume that the data consisting of the participants' joint locations with respect to their chests and the rest of the features is represented by D. Based on the highest macro-accuracy, a set f_i of the top 12 contributing features for each subject i was chosen. Next, a pool of features was created by taking the union of the top most contributing two features from f_i for each of the individual subjects. Lastly, to choose the best subset of features for regression, we pose the feature selection as a regression problem to predict weights in kg instead of classes, and select the top 8 most contributing features (F) from the unioned pool of features using the forward sequential feature selector (FSFS) method.

FSFS for this regression task used mean squared error as the loss function and GMR as the regressor. These 8 selected features were used in LightGBM, GMR, and MLP regression experiments in section 5.5.4.

Equations 5.1 and 5.2 describe a two-step feature selection process applied across all subjects. In the first step, the FSFS method uses a classifier (denoted by the subscript C) to select a subset of features f_i for each subject i where $d_i \subset D$ represents the feature set for the i^{th} subject ($i = 1, ..., 27$). This step is repeated for all subjects individually. In the second step, the selected feature sets f_i from all subjects are combined, i.e., a union of all selected features, and subjected to a regressor-based FSFS procedure (denoted by the subscript R) to generate the final optimal set of features F, from which the best features are chosen in order.

$$f_i = FSFS_C(d_i)_{i=1}^{27} \tag{5.1}$$

$$F = FSFS_R(\cup_{i=1}^{27} f_i) \tag{5.2}$$

The following eight features have been shortlisted in the order of their contribution for regression: (1) left-wrist distance from table; (2) left-wrist x-component position at time t; (3) left-wrist y-component position at time $t-1$; (4) left-wrist distance from table to right-elbow; (5) left-wrist y-component position at time t; (6) left-shoulder distance from table; (7) left-shoulder velocity of x-component computed with window-size of ± 2; and (8) left-elbow distance from table to left-elbow.

The selected features include both temporal (position at time $t-1$, velocity, etc.) and spatial (instantaneous distance from the table, location at time t, etc.) aspects. The fact that most of the selected features are associated with the left arm is one clear pattern. This illustrates the significance of the non-dominant arm in weight identification. This could imply that the non-dominant arm's movement is more visibly impacted than the dominant arm's when using weights of different amounts. This could be explained by the fact that people typically use their dominant arm for eating because it is stronger and more accustomed to performing fine motor tasks than their non-dominant arm, which is weaker because it is used less frequently.

5.5.2.2 Feature Visualization

We use T-SNE to project the 8-dimensional data to 2 dimensions to show how the data looks with the 8 most contributing features. The information is shown in Fig. 5.6. Although there aren't four distinct groups for each of the four weights, there is a slight

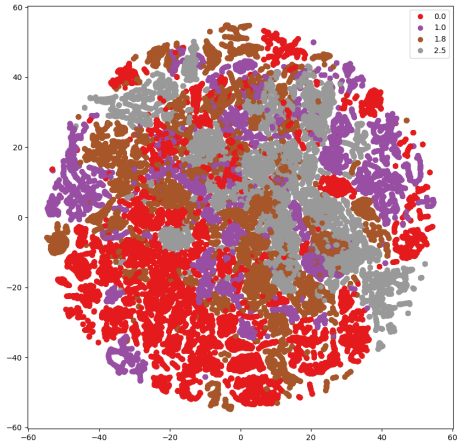


Figure 5.6: T-SNE plot for the best performing 8 features mapped to 2D plane for all 27 subjects.

clustering that implies some modeling may be feasible, particularly for the no-weight (shown in red) class.

5.5.2.3 Gaussian Mixture Regression

The regression technique known as Gaussian mixture regression (GMR) [179],[180] is a modified form of Gaussian mixture modeling (GMM). GMR is a probabilistic approach that assumes a finite number of Gaussian mixtures can effectively represent all the data points in the input $\times$ output space. It can model multi-modal mappings since it works with probabilistic distributions instead of functions. The regression approach's entire pipeline is displayed in Fig. 5.4. Below is a quick summary of the GMR's training and prediction details. For more information, readers are referred to [179] and [181].

The Gaussian mixture model (GMM) is fitted across the feature set F (eq. 5.2) in an unsupervised format using the Expectation Maximization (EM) technique to train the GMR regressor. The vectors $\mathbf{z}_n = [\mathbf{x}_n^T \mathbf{y}_n]^T$ can be concatenated together since there is no difference between the input $\mathbf{x}_n$ and the target $\mathbf{y}_n$. Equation 5.3 illustrates how the GMM models the probability density function of the vector $\mathbf{z}_n$ by representing a weighted sum of E Gaussian functions.

$$p(z_n) = \sum_{e=1}^{E} \pi_e \mathcal{N}(\mathbf{z}_n; \mu_e, \sigma_e), \quad with \sum_{e=1}^{E} \pi_e = 1 \tag{5.3}$$

For inference, with regression we are interested in predicting $\hat{\mathbf{y}} = E(\mathbf{y}|\mathbf{x})$ i.e., the expected value of $\mathbf{y}$ given $\mathbf{x}$. For this purpose, μ_e and σ_e can be separated into input and output components as follows:

$$\mu_e = [\mu_{e,X}^T, \mu_{e,Y}^T]; \quad \sigma_e = \begin{bmatrix} \sigma_{e,X} & \sigma_{e,XY} \\ \sigma_{e,YX} & \sigma_{e,Y} \end{bmatrix} \tag{5.4}$$

Given the decomposition in eq. 5.4, the expected value of $\mathbf{y}$ given $\mathbf{x}$ can be calculated by,

$$\hat{\mathbf{y}} = \sum_{e=1}^{E} h_e(\mathbf{x})(\mu_{e,Y} + \sigma_{e,YX}\sigma_{e,X}^{-1}(\mathbf{x} - \mu_{e,X})); \tag{5.5}$$

where,

$$h_e(\mathbf{x}) = \frac{\pi_e \mathcal{N}(\mathbf{x}; \mu_{e,X}, \sigma_{e,X})}{\sum_{l=1}^{E} \pi_l \mathcal{N}(\mathbf{x}; \mu_{l,X}, \sigma_{l,X})} \tag{5.6}$$

Since probabilistic models are inherently flexible and can successfully describe complicated problems while accounting for uncertainty, we propose employing GMR to model the regression problem across different subjects. The experiments demonstrate the good performance of GMR, as indicated in section 5.5.4.

5.5.2.4 Multi-Layer Perceptron Regression

The ability of a Multilayer Perceptron (MLP) to learn and recognize complicated (non)linear patterns in data has made it a popular type of artificial neural network (ANN). Several linked layers of neurons make up this supervised algorithm, each of which processes and modifies the input to match an output.

The problem of degradation (weight) estimation tends to overfit to a subset of training subjects, meaning that it does not generalize to other subjects. As a result, lasso ($\mathcal{L}_1$) and ridge ($\mathcal{L}_2$) regularisation are both included in the joint loss function. The loss function is eq 5.7 where y is the ground truth label or weight, and $\hat{y}$ represents the regression's predicted output. 5.7. The feature set F (eq. 5.2) was used for training.

$$\mathcal{L} = \alpha|y - \hat{y}|_2^2 + (1 - \alpha)|y - \hat{y}| \tag{5.7}$$

where α was set to 0.5.

5.5.3 Deep Features-Based Regression

High-level representations of data that deep neural networks (DNN) develop to capture intricate patterns and correlations in data are known as deep features. Compared to

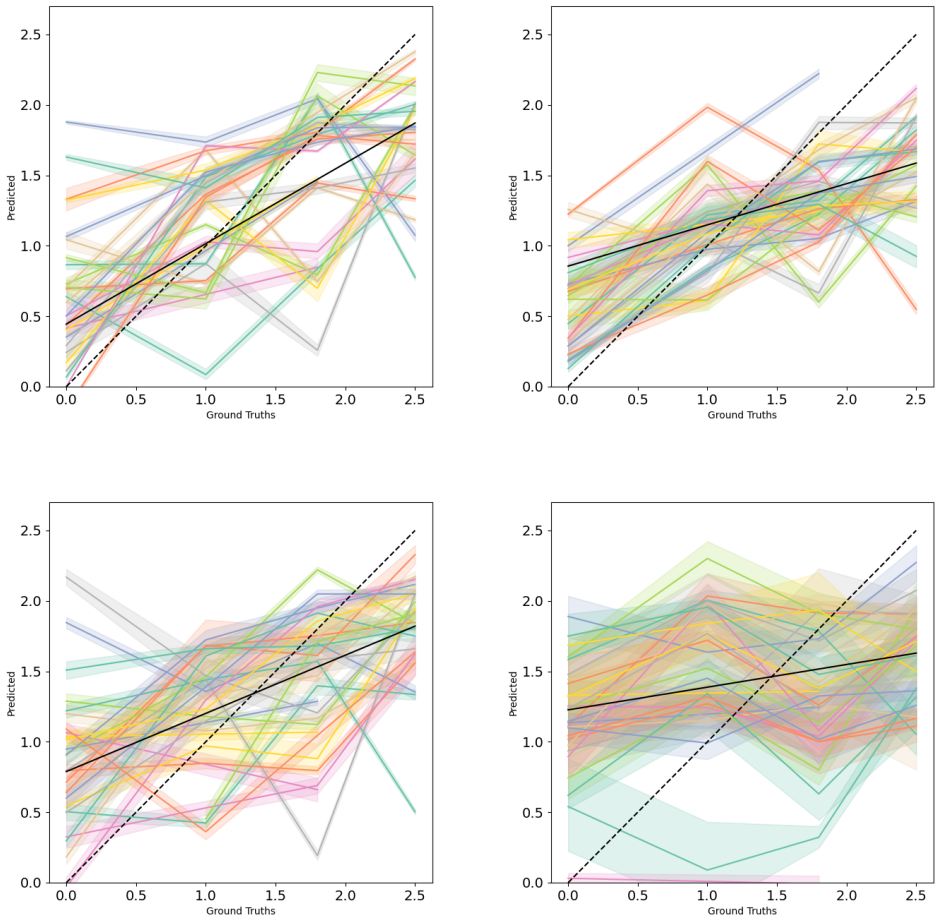


Figure 5.7: The plots show the expected weight against the ground truth weight. The solid line represents the RANSAC-based least square fit of the data, while the dashed black line shows the perfect correlation. GMR (top-left), MLP (top-right), LightGBM (bottom-left), and ST-GCN Regression (bottom-right) are the four regressors that were evaluated. Every colored curve represents the outcome of a separate leave-one-out model. The solid-colored curves show the average of these predictions because each micromovement in the test set consists of multiple frames or clips, and the shading around each curve shows the range of one standard deviation.

handcrafted features or shallow representations, deep features have several advantages. Their automatic inference from the data, which enables the network to dynamically adapt to the particular task, is one of its main advantages.

A Spatial-Temporal Graph Convolutional Network (ST-GCN) was utilized to explore generalized regression with deep features [127]. According to [8], ST-GCN was the optimal action recognition algorithm for the EatSense action classification, which

is why it was selected for this task.

5.5.3.1 ST-GCN

In ST-GCN [127], a spatial-temporal graph is created with joints as graph nodes, inter-joint connections, and temporal connections (e.g., joint j at time t and $t+1$) as graph edges given the sequence of the body joints (3D in our example). High-level feature maps are produced by subjecting the input data to spatial-temporal graph convolution procedures. A classification head is utilized in generic ST-GCN to complete the classification work.

The original approach as described in the previous paragraph was used for extracting high-level deep features for ST-GCN (without the classification head). However, since we want to pose the problem as a regression task rather than a classification, two significant changes were made to the ST-GCN framework. First, a regression head took the place of the classification head. Second, as given in eq. 5.8, the mean squared error was used in place of the default cross-entropy loss function used for classification by ST-GCN.

$$\mathcal{L} = \|y - \hat{y}\|_2^2 \tag{5.8}$$

5.5.4 Experiments

The generalized regression experiments consist of two sub-experiments, as was previously indicated, *i.e.* comparing deep features-based regression and manually constructed feature-based regression. Prior to any of these sub-experiments, hyper-parameter adjustment is done. The following section discusses the sub-experiments and the hyper-parameter selection techniques used in each. The experimental question under discussion is: How accurately can the weight on the wrist be predicted for individual subjects (as a proxy for modeling decline in the subjects)?

5.5.4.1 Hyperparameter Tuning

The number of Gaussians E required to accurately describe the input $\times$ output space is the most crucial hyperparameter for GMR. A recursive method consisting of conducting 26-vs-1 cross-validation across subjects and looking for E was employed. One participant was left out for testing while the other 26 were used for training and validation in the 26-vs-1 cross-validation method and the average was reported. Bayesian

optimization was utilized to search for the optimum E and identify the configuration with the lowest mean squared error across individuals between the anticipated labels and ground truth.

The number of layers, the number of neurons in each layer, the learning rate, the drop-out rate, and the batch size in MLP were selected using empirical methods. Similar to GMR, only features chosen according to the sub-section 5.5.2.1 were employed for MLP. Additionally, empirical selection was used to determine other ST-GCN hyperparameters, such as learning rate.

5.5.4.2 Estimating the Weight Level using Regression

Following the selection of the optimal configurations, the average mean squared errors (MSE) and actual error for GMR, MLP, LightGBM, and ST-GCN regression were determined using leave-one-out cross-validation. The leave-one-out method involves training the model on a major subset of the available data (in this case, 26 subjects), except for one subject. The model's performance is then assessed on the subject that was left out. Every subject goes through this process as a test set, and the model's total performance is the mean of the individual subjects' performances.

Only the two most distinguishing micro-movements (actions), 'move hand towards mouth' and 'move hand from mouth', were used in each of the two sub-experiments. Since these two behaviors entail working against or with gravity, they were selected because they appear to be the most susceptible to the effects of different weights. The features were set frame-by-frame for MLP, LightGBM, and GMR. In contrast, feature maps were extracted from vectors that contained the 3D postures of a single complete motion for the purpose of training ST-GCN. These feature maps then pass through the regression head to determine the weights. The weight that each individual wears is predicted using their respective regression models.

We show both quantitative and visual statistics to illustrate the performance of the suggested regression model. Fig. 5.7, displays the predictions made by the 27 distinct models that were trained with the leave-one-out strategy.

Every curve represents the result of the single subject who did not participate in the training procedure. These predictions are averaged across time since each subject makes several motions of the hand towards and away from the mouth during a single eating session, which makes up the test set. This mean is shown by the solid line, and the ± 1 standard deviation of the predictions is shown by the shading around it. To summarize, we fit a RANSAC [182] linear regression model across the predicted

weights of all 27 regression models. [1] The RANSAC linear regression fit line over the predicted weights is shown by the black solid line in figure 5.7, while the perfect correlation between the predicted and ground truth weights is shown by the black dashed line.

The mean squared error (MSE) and actual error are the two measurements used to produce results for quantitative analysis. The discrepancy between the expected and true values, expressed as the ($\mathcal{L}_2$)-norm, is known as the mean square error (MSE). Similarly, we also estimate the actual error that shows the deviation from the original weights in kilograms. Equations 5.9 and 5.10 describe the process for calculating the actual error in the context of weight prediction. The actual error for each subject p denoted as M_p is determined by taking the average of the differences between the predicted and true values of weight for all N_p samples of that subject, as shown in Equation 5.9. Here, M_p represents the deviation between the predicted and true weights for the p^{th} subject. Equation 5.10 then computes the overall mean error by averaging the actual errors M_p across all 27 subjects. This final value summarizes the model's performance in terms of prediction accuracy across the entire test set.

Results are given in Table 5.3.

$$M_p = \frac{1}{N_p} \sum_{n=1}^{N_p} (predicted_{p,n} - true_{p,n}) \tag{5.9}$$

$$mean = \frac{1}{27} \sum_{p=1}^{27} M_p \tag{5.10}$$

5.5.5 Discussion of Results

Both MLP and ST-GCN perform well in various scenarios and can manage a broad variety of data distributions. An MLP, for instance, is capable of modeling complex non-linearities in high-dimensional data, while ST-GCN is more suited for jobs involving both spatial and temporal dimensions. In contrast, GMR models data distributions as combinations of Gaussian mixtures using a probabilistic approach. When using high-dimensional feature space for a problem, LightGBM functions as an ensemble of decision trees and is appropriate.

In figure 5.7, we can visually compare the performance level of three hand-crafted feature-based methods - GMR, MLP, and LightGBM, and deep feature-based ST-GCN

[1] RANSAC is an estimator that is robust against outliers in which it estimates the model parameters by randomly sampling the observed data.

Table 5.2: Mean squared error for GMR, MLP, LightGBM, and ST-GCN as a result of Leave-one-out regression. The last row shows the average of these errors. Values closer to zero are better, and GMR has the best average performance.

S#	GMR	MLP	LightGBM	ST-GCN
0	0.977	0.680	**0.658**	1.540
1	**0.691**	1.856	0.724	1.274
2	**0.189**	1.369	0.845	1.160
3	**0.805**	1.010	1.382	1.210
4	**0.592**	1.363	0.669	0.705
5	**0.404**	1.269	0.859	0.961
6	0.643	0.939	**0.396**	0.663
7	**0.291**	0.581	0.618	0.708
8	0.613	**0.519**	1.674	1.299
9	**0.398**	1.190	0.760	1.069
10	0.931	1.235	1.229	**0.872**
11	**0.597**	0.787	0.738	0.703
12	**0.635**	1.275	0.544	0.975
13	0.629	0.833	**0.420**	1.172
14	0.788	0.627	**0.345**	0.961
15	0.760	0.884	1.279	**0.750**
16	**0.288**	0.432	0.433	1.034
17	**0.598**	0.631	0.629	0.910
18	0.599	0.463	**0.383**	1.290
19	**0.140**	0.313	0.127	0.967
20	**0.327**	0.887	0.329	0.989
21	0.586	1.260	**0.442**	0.814
22	0.284	0.371	**0.177**	0.685
23	**0.328**	0.395	0.452	1.120
24	**0.645**	1.467	0.810	1.327
25	**0.337**	0.538	0.852	1.041
26	0.267	0.834	**0.258**	0.872
Avg.	**0.531**	0.889	0.668	1.003

Table 5.3: Actual error for GMR, MLP, LightGBM, and ST-GCN as a result of Leave-one-out regression. The last row shows the mean of these errors. Values closer to zero are better.

S#	GMR	MLP	LightGBM	ST-GCN
0	-0.256	-0.186	**-0.164**	-0.912
1	0.470	0.641	**0.464**	1.019
2	**0.179**	0.898	0.755	0.746
3	-0.391	0.144	**-0.017**	0.055
4	-0.259	-0.296	**-0.064**	-0.064
5	0.125	0.127	0.072	**0.005**
6	-0.346	-0.233	**-0.009**	0.142
7	-0.153	-0.328	-0.093	**-0.009**
8	-0.422	-0.337	**-0.240**	0.817
9	**-0.345**	-0.496	-0.612	-0.801
10	0.187	-0.516	**-0.183**	-0.222
11	-0.180	**-0.082**	-0.542	0.399
12	-0.246	-0.497	**0.049**	-0.181
13	**-0.053**	-0.356	-0.256	-0.124
14	-0.185	-0.193	**-0.023**	0.411
15	**0.031**	-0.234	0.044	-0.137
16	0.146	-0.244	-0.224	**-0.087**
17	0.402	-0.137	0.526	**-0.001**
18	0.215	-0.192	0.261	**-0.113**
19	0.141	0.051	**0.027**	0.101
20	**-0.039**	-0.600	-0.488	-0.094
21	**0.009**	-0.460	0.254	-0.200
22	0.200	**-0.038**	-0.049	-0.141
23	0.134	-0.183	-0.225	**-0.120**
24	0.091	-0.405	0.295	**0.011**
25	**-0.003**	-0.483	-0.374	-0.230
26	**0.031**	-0.444	0.092	-0.199
Avg.	-0.019	-0.188	-0.026	**0.002**
std.	**0.233**	0.333	0.312	0.404

- using line plots that compare the predicted weights to the ground truth. The more the level of the predicted weights (solid black line) matches the actual values (dashed black line), the more accurate the predicted weights from the respective regression models.

Examining the findings shown in the top-left image, it is evident that GMR performs well because the plot shows a discernible increasing trend, suggesting that the weights (0, 1, 1.8, and 2.4 kg) were correctly predicted. However, given their lower correlation between ground truth and predicted values, the top-right (MLP) and bottom-left (LightGBM) results imply that these models do not generalize as well on the data. It is evident from the bottom-right figure (ST-GCN regressor) that the model is not a good fit for the data. Two possible explanations for this could be (1) not enough data to train a regression model with only two micro-movements, or (2) not enough temporal context and restricted discriminative features because the micro-movements under consideration extend across fewer than 10 frames.

When comparing these methods quantitatively, GMR performs better, as evidenced by the average MSE displayed in Table 5.2. GMR achieved a mean squared error of 0.53 i.e., lower than MLP, LightGBM, and ST-GCN.

In real scenarios, it is unlikely to have data from various stages of performance decline to train a model. Instead, one would have to use one of the generic regression models trained in section 5.5.4.2. Therefore, relying solely on MSE to quantify the error may seem to be complicated or not intuitive in a physical sense and may not be the most appropriate metric for selecting the best model. Table 5.3 addresses this by showing the actual error, which measures the average amount in kilograms that the predictions are off by $\frac{1}{N} * \sum_{n=1}^{N} (predicted - true)$ for each row. The data indicates that the lowest standard deviation, 0.233, corresponds to a mean difference for GMR of about 19 grams. ST-GCN, on the other hand, has the lowest mean and a slightly higher standard deviation.

The data has numerous modes, as shown by the t-SNE visualization shown in Figure 5.6. When considering eight dimensions, we expect more distinct boundaries. Rather than trying to fit a single line or curve across all data, Gaussian Mixture Regression (GMR) works well in this situation intuitively by clustering data points and expressing each mode with its own Gaussian component.

As a result, when it comes to modeling the underlying distributions, GMR performs better than other regression techniques.

5.6 Conclusion

We examined the eating habits of the participants in this chapter, along with the techniques for measuring their performance and degree of performance decline. Two sets of experiments were carried out to quantify the performance levels during eating: one set involved hand-crafted features using the uncertainty-aware algorithm GMR, which was compared against MLP and LightGBM, and the other involved deep features-based regression using ST-GCN.

According to the results, GMR outperformed other regression models by a small margin, meaning that it can be used to predict an individual's wrist weight level or degree of performance decline based on a generically trained model, i.e., a model that has been sufficiently trained with data from subjects other than the test subject.

In a perfect scenario, we would gather long-term data from elderly participants to confirm the performance decline model. However, this would be highly unethical, as interventions should be made at the earliest signs of performance decline. Therefore, the experiments discussed here are confined to using wrist weights worn by healthy volunteers.

Chapter 6

V2R: A Fully Autonomous Vision-Based System

In this chapter, we introduce a fully automated pipeline that takes in a video of an individual and outputs a report that considers both eating behavior statistics and muscle degradation.

6.1 Introduction

Digital home health monitoring systems can be broadly categorized into two main types: vision-based systems and wearable sensor-based systems. Wearable sensor-based systems, or multi-sensor systems, are effective for identifying acute health conditions. However, their acceptance is limited due to their intrusive nature and the possibility of users forgetting to wear or recharge them regularly. In contrast, camera or vision-based systems offer a non-intrusive alternative. They can assist in the detection of crucial situations and trends without requiring users to wear any devices. This non-intrusive nature makes them more acceptable and potentially more user-friendly for continuous monitoring in home healthcare settings.

Vision-based monitoring systems facilitate early identification and intervention by making it simpler to track subtle indicators in behavioral health informatics. Clinical systems raise the risk of human error because they are expensive and need a human operator to be in the loop. Consequently, automated vision-based systems can be useful tools for physical rehabilitation or the assessment of diseases such as Parkinson's disease (PD) and stroke [183]. Analyzing the visual data generally reveals trends and abnormalities, allowing medical professionals to make better decisions.

We explored three research questions in this chapter:

1. How does the TAL network for two actions localization perform when compared between hand-crafted and deep-learning-based features?

2. Is the change in motion speeds detectable by increasing weights on the wrists?

3. Can we capture the general behavioral trends, present in the ground-truth, through a fully autonomous pipeline?

This chapter presents a vision-based fully autonomous framework comprising three stages. The first stage focuses on promoting healthy eating habits by analyzing eating behaviors, such as chewing duration and mouthful count. In the second stage, the framework tracks the speed of arm movement to identify potential decreases in muscle activity, which could indicate changes in motor function or health status. The third stage introduces a novel classification technique designed to detect anomalies in eating posture. This stage aims to identify deviations from typical eating postures that may indicate discomfort, impairment, or other issues requiring attention.

The contributions of this chapter are:

- A multi-purpose, fully autonomous, video-to-report (V2R) pipeline for long-term eating behavior and performance decline monitoring (Section 6.3.2).

- The chapter introduces a relaxed data augmentation (pre-processing) step and the temporal segments output merging (post-processing step) for the temporal action localization (TAL) network that helps to accurately localize the atomic-actions in the continuous video (Section 6.4.1).

- The proposed pipeline can capture trends and generate valuable insights on changes in eating behavior and upper-body muscular movements. This is demonstrated by carrying out a holistic analysis of the proposed pipeline (Section 6.4.3).

- A small extension is made to the EatSense dataset (Ch. 3), where long-term changes of an individual's behavior are simulated by adding weights (0, 1kg, and 2.4kg to each wrist) to the wrists of the subjects (Section 6.3.1).

6.2 Literature Review

The majority of vision-based health monitoring systems are only concerned with preventing falls and detecting them. While keeping an eye on people falling is important, it's just as important to keep an eye on people's behavior for any long-term changes. Research on vision-based health monitoring systems can be divided into two categories: 1) studies that concentrate on health-related tasks like categorizing or comprehending activities or eating habits; and 2) fully autonomous monitoring systems that use a video as an input and provide a meaningful summary of findings that can be helpful to a healthcare professional.

This section presents the literature review on non-clinical, home-based health monitoring systems.

6.2.1 Vision-based Health-Related Research

In general, vision-based health research takes into account several factors related to an individual's health, such as vision-based gait analysis for the detection of neurological diseases [184]. However, the research presented in this chapter focuses on examining an individual's upper body motions; as a result, this literature review's scope is restricted to studies on the health of the upper body or full body. This discussion is further broken down into the monitoring of eating behavior and general activities. ADL monitoring (appendix B) entails keeping an eye on a person all day long and analyzing the data for a variety of purposes, such as action quality evaluation or rehabilitation. While the latter is more focused on eating behavior, it can also be helpful for additional analysis, such as evaluating performance decline and eating disorder detection, etc.

6.2.1.1 Activity Monitoring

A thorough understanding of a person's daily activities is essential to providing individualized services or treatments. Accurately identifying activities has several benefits, including the ability to analyze lifestyle, track diet, support active rehabilitation, and more.

Over the past ten years, a lot of research has been conducted on ADL monitoring for physical rehabilitation of individuals [185],[183]. The aim of [186] and [187] was to automatically evaluate the physical activities of daily living (ADLs) for people with Parkinson's disease or stroke using a publicly available dataset that was taken with

a Kinect V2 sensor. To improve the explainability of the model, Deb et al. [188] combined two publicly accessible datasets for automatically evaluating ADLs using attention modules in the deep network.

In order to classify certain activities (such as sitting, standing up, walking, etc.) as normal or abnormal and evaluate the effectiveness of the action taken, Elkholy et al. [189] monitored a subset of these activities and created a multi-head deep network.

6.2.1.2 Eating Behavior Monitoring

Eating behaviors can be broadly categorized by the way an individual bites, chews, swallows, and so on; they can also be classified by the actions they typically take while eating or drinking, such as adding sugar to their tea. Instead of gaining deeper behavioral insights from the data, the majority of vision-based eating monitoring algorithms, like [190], [191], and [192], concentrate on identifying eating and drinking activities.

An extensive review of thirteen video-based techniques was conducted by Tufano et al. [193], with results including the number of chews, meal duration, bite counts, and intake gesture detection. They also draw attention to the paucity of studies on eating behaviors in unregulated settings. In [194], Lasschujit et al. presented a tray with weighing sensors to track the amount of food consumed instantly and a camera to record bites and chews.

Chapter 5 presented a general eating behavior state diagram and used eating videos to evaluate performance levels. Since the indicative features of performance decline differ significantly amongst people with different lifestyles, we also proposed an uncertainty-aware algorithm to obtain a generalized model to regress performance changes across multiple subjects. This algorithm was discussed in Section 5.5.

6.2.2 Fully Autonomous Monitoring Systems

Interest in fully autonomous systems for monitoring the elderly has grown in recent years. Luo et al. [195] used an infrared and a depth sensor to monitor elderly people and create time logs of their daily activities. The system used smoothing windowed filtering in conjunction with a frame-by-frame temporal activity detection algorithm. While framewise temporal action localization helps understand different aspects of an elderly person's routine, activity logging alone is not a useful way to identify patterns or abnormalities in the person's lifestyle.

A similar system was recently proposed by Huang et al. [196] and included frame-

wise temporal action localization, followed by activity detection, facial analysis, and subjects' interaction with the environment. This system analyzes the videos in a meaningful way to gain a better understanding of a person's long-term behavior.

In conclusion, prior research using full-body or gait motion analysis pipelines offers important insights into people's everyday activities and behaviors. Research on completely autonomous vision-based systems to evaluate health statistics is lacking, and there isn't any research that focuses only on upper body motions with a comprehensive assessment. In this chapter, a novel fully autonomous system for tracking and evaluating eating habits and musculoskeletal performance decline is presented.

6.3 Methodology

6.3.1 EatSense and New Test Set

The EatSense dataset discussed in chapter 3 was used in this study. EatSense was gathered in dining settings that had RGB-D Intel RealSense D415 cameras installed.

EatSense includes sixteen action classes that encompass both gesture-based (such as 'chewing') and velocity-based (such as 'move hand towards mouth') micro actions during eating. Additionally, it simulates the performance decline of the musculoskeletal system by fastening weights of different sizes to the subject's wrists. Our proposed dataset, EatSense, is the ideal option for training the suggested autonomous system because it focuses solely on the upper body of individuals and has detailed sub-action labels. Nevertheless, the study described in this chapter only makes use of two micro-actions, 'move hand towards mouth' and 'move hand away from mouth', each lasting roughly a second. Given that the videos are being recorded at a frame rate of 15 frames per second, the average duration of 'move hand towards mouth' and 'move hand away from mouth' is 12.7 and 9.4 frames, respectively. EatSense contains a total of 5643 instances of these actions.

We recorded an additional test set with properties and configurations similar to the EatSense dataset. Three videos for each of the two subjects were captured in this recording session. During these recordings, participants were requested to eat from bowls that were large, medium, and small in size (containing roughly 1000, 625, and 250 ml volume of fluid when full), all while wearing weights on their wrists that were 0 kg, 1 kg, and 2.4 kg, respectively. We used three weight classes to simulate easy, medium, and hard cases on the subjects and to reduce the labeling time required for

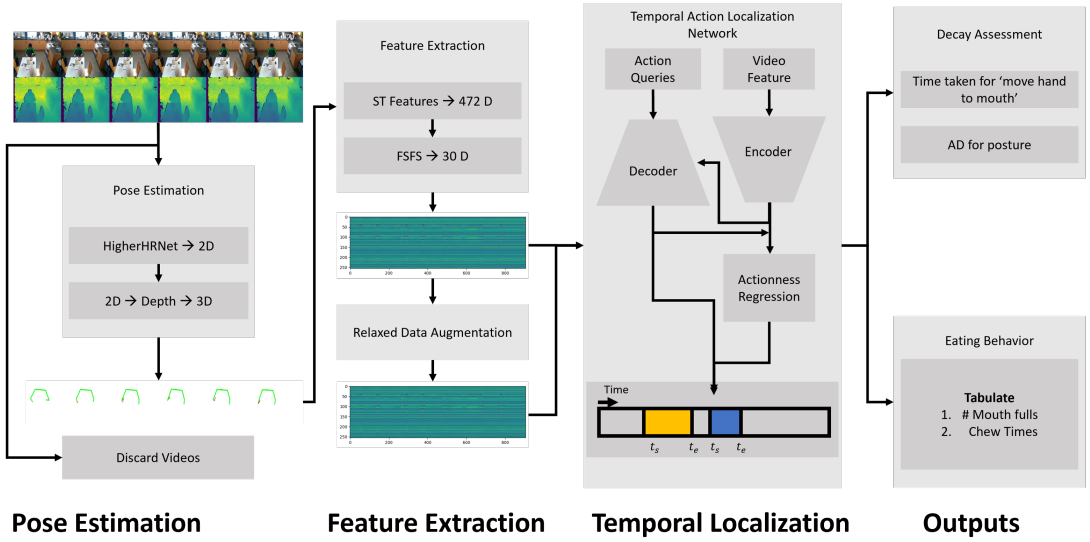


Figure 6.1: Block diagram of the proposed system. The proposed pipeline consists of three steps after the video is collected, 1) it finds the poses from the video and estimates features using those poses, 2) uses these estimated features as an input to the temporal action localization framework to get temporal segments and action classes, and 3) derive insights into the eating behavior and muscular movement of the individual.

the researchers involved. With the use of heavier weights, the setup attempted to replicate changes in eating behaviors brought on by performance decline, such as fewer mouthfuls or slower arm movements. This test set is made available on the original EatSense website as well.

6.3.2 Proposed System

The proposed pipeline is shown in the Fig. 6.1. The pipeline uses a temporal action localization (TAL) framework to extract data from the untrimmed videos in the EatSense dataset. The TAL network outputs the names and the start and end times of each temporal segment.

Because end-to-end training requires a significant amount of computational power, many TAL networks heavily rely on separately extracting video embeddings (i.e., frame-wise activity and context descriptions) as a step in the pipeline. Typically, deep learning networks such as I3D [144] and TSP [197] are used to estimate these video embeddings. However, in this study, we also used domain knowledge to engineer features that can potentially improve the model's explainability for medical professionals.

We only localize two actions out of the 16 sub-actions in the EatSense dataset. The rest are marked as background for detection and localization. The output activity segments are analyzed to count mouthfuls and estimate chewing duration after temporal segment information is obtained. Additionally, posture changes in the upper body due to different weights are detected using an anomaly detection algorithm (one-class classifier). Lastly, the duration of the 'move hand to mouth' action was timed to track musculoskeletal performance decline.

The pipeline can be divided into four stages, denoted by the text under the block diagram in Fig. 6.1. Each of these blocks is discussed in detail in the following subsections.

6.3.2.1 Pose Estimation and Feature Extraction

HigherHRNet [198] is used to estimate the 2D poses and we get the 3D joint location using the depth information from the RGB-D camera placed in front of the person eating in the EatSense dataset's videos. Since the subject's upper body is the only visible part, eight joints were chosen for analysis: the head, chest, left shoulder, right shoulder, left elbow, right elbow, left wrist, and right wrist. The details of the pose estimation process are discussed in chapter 3 (section 3.3.2.1).

Different spatial and temporal features are computed from the estimated 24-D vector (8×3) that contains the absolute location of eight 3D joints. These consist of the joints' instantaneous positions in relation to the chest, the chest's instantaneous distance from the table, the previous three lags, acceleration, and velocity, among other things. The mathematical details of these features were discussed in detail in chapter 4 section 4.5.1. Since we only localize two actions, we use a forward sequential feature selection (FSFS) algorithm to determine which features contributed the most to the frame-wise classification of the 'background' versus the two micro-actions, 'move hand towards mouth' and 'move hand away from mouth'. The top 30 features shown in Fig. 6.2 were chosen using FSFS to create a 30-D video feature embedding and were used as the input for the TAL network.

6.3.2.2 Temporal Action Localization

TadTR [142] served as inspiration for the transformer used for temporal action localization, which is the process of determining when actions begin and end in a video. The Transformer-based framework (TadTR) receives learnable action queries and video

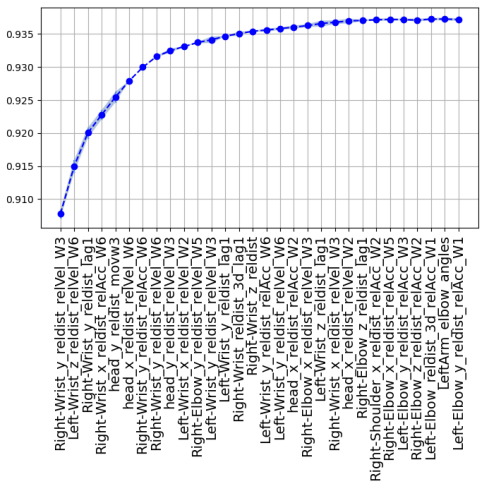


Figure 6.2: Most contributing 30 features selected using forward sequential feature selection. The vertical axis shows the accuracy achieved on the frame-wise classification of 'background', 'move hand to mouth' and 'move hand away from mouth'.

features as input. Using temporal deformable self-attention, the encoder records the temporal context between video clips. Using temporal deformable cross-attention, the decoder retrieves pertinent snippet-level context for every query and employs self-attention to model inter-query relations. Following the decoder, feedforward networks predict classes and action segments.

Using recently extracted context, a segment refinement mechanism iteratively improves the segments. Confidence scores for each of these predicted segments are predicted by a regression module called 'actionness head'. One-to-one targets are generated by the action matching module, allowing for end-to-end training without non-maximal suppression (NMS). Rather than utilizing the intricate pipelines present in earlier techniques, the overall architecture uses a single network to directly predict localized action instances in a video.

The goal of training the model was to localize the 'move hand towards mouth' and 'move hand away from mouth' actions. Since these actions last less than a second (less than 15 frames on average), any error could falsely inflate or deflate the performance of TAL network, so to compensate for the human labeling errors, the ground truth action instances were cropped or temporally extended at random by ± 2 frames on either side. First, the boundaries with precise hand labels were applied. Second, we used a relaxed boundary threshold $\varepsilon \in \{-2 \times \frac{1}{15}, -1 \times \frac{1}{15}, 0 \times \frac{1}{15}, 1 \times \frac{1}{15}, 2 \times \frac{1}{15}\}$ to redefine the temporal segment's start t_s^g and end t_e^g. Since videos are recorded at a frame rate of 15 frames per second, the ± 2 frames are randomly selected, multiplied by $1/15$, and

Figure 6.3: Relaxed Data Augmentation. t_s^g and t_e^g refer to the start and end of action in the ground truth whereas ε denotes the ± 2 frames i.e, $\varepsilon \in \{-2 \times \frac{1}{15}, -1 \times \frac{1}{15}, 0 \times \frac{1}{15}, 1 \times \frac{1}{15}, 2 \times \frac{1}{15}\}$ seconds relaxation for augmentation.

added to t_s^g and t_e^g. Figure 6.3 illustrates this augmentation process.

The temporal segments (t_s^p and t_e^p), action class labels, and a confidence score make up the TAL network's outputs. Each test video sequence has a fixed number of output detections because transformers operate on a fixed number of queries. Based on these outputs, overlapping predicted detections were combined if any of the following conditions were met: 1) the actionness score was more than 0.3; 2) the segments overlapped for more than two frames, or 2*($\frac{1}{15}$) seconds); and 3) each prediction lasted longer than 0.1 seconds. The remaining predicted segments were discarded.

6.3.2.3 Anomaly Detection

One-class classification, another name for anomaly detection (AD), is the process of identifying patterns that exhibit a large deviation from the norm. It's a crucial technique in data analysis for identifying anomalies or strange behavior. Using an anomaly detection (AD) / one-class classifier, abnormalities in a person's instantaneous posture during the 'move hand to mouth' micro-action were identified. We use thirty intuitively selected postural features to highlight the posture, which includes the instantaneous distance between wrists, the distance of each of the eight joints from the table, and 21 postural features (derived from the positions of eight 3D upper-body joints relative to the chest). Since the chest is the origin, it is not included in the features.

EatSense has videos representing four distinct weight classes. We hypothesize that the data recorded with the highest weight (2.4 kg per wrist) is the most anomalous compared to normal data recorded without weights since weights can change an individual's posture and arm movement speeds. As a result, an AD model was trained with the highest weight representing anomalous data and zero weight representing normal data to explore this research question.

6.4 Experiments

The suggested pipeline was assessed through several experiments. Initially, we assess the V2R approach's accuracy for each component independently. Second, we determine whether there is a behavioral trend in the dataset itself. Third, we evaluate the pipeline as a whole, using videos as the input and evaluating each of the four outputs independently.

6.4.1 Temporal Action Localization (TAL) Tests

Claim 1: The TAL network alongside the proposed data pre-processing pipeline can extract most instances of the 2 actions from the continuous video, and accurately estimate the starting and ending frame times.

Using the EatSense dataset, 5-fold cross-validation is used to assess the TAL network. EatSense is an unbalanced dataset where some subjects have more videos than others. In order to mitigate potential bias resulting from varying behavioral traits among subjects, the dataset was rebalanced using four videos from each of the twenty-seven subjects (three of whom have only two videos so these 3 were not used), for a total of $4 \times 24 = 96$ videos. The dataset is then split into four groups of five subjects each and another group of four subjects, resulting in a total of five parts.

Two sub-experiments were conducted on the TAL network: 1) TAL using hand-crafted (HC) video encodings, and 2) TAL using deep learning-based video encodings (TSP [197]). Five-fold cross-validation was applied to both sub-experiments, using the data splits mentioned in the preceding paragraph. The networks utilized the same hyperparameter values including the learning rate and the number of epochs. The mean and standard deviation of the achieved mean average precision (mAP) at 10%, 30%, and 50% intersection over union (IoU) are shown in Table 6.1. The average mAP (incremented by 0.05 at each step) over all thresholds between 0 and 0.95 is shown in the final column.

The results suggest that both hand-crafted and deep learning-based video encodings achieve comparable performance in identifying action instances and segmenting them from continuous video streams. However, the high standard deviation observed indicates significant variability in the dataset, likely stemming from differences in the parameters characterizing each individual's motion profile. Consequently, subject-wise splitting for 5-fold cross-validation leads to a domain shift [10]. For subsequent experiments, such as holistic evaluations, the Temporal Action Localization (TAL) model

Table 6.1: The mean (μ) and standard deviation (σ) of the 5-fold CV results from the two TAL networks at various intersections over union thresholds. HC represents hand-crafted features and TSP shows the results on deep video encoded features.

Features	**mAP@0.10**		**mAP@0.30**		**mAP@0.50**		**0:0.05:0.95**	
	μ	σ	μ	σ	μ	σ	μ	σ
HC	73.8	10.5	65.6	13.7	43.2	16.6	41.8	8.9
TSP	71.7	8.5	64.1	9.8	49.4	12.9	41.2	7.2

trained with hand-crafted features was utilized. This decision was made to ensure consistency and facilitate direct comparisons across different stages of the evaluation process.

It should be noted that the intersection over union (IoU) threshold, which is used for mAP estimation, is extremely sensitive when used for brief actions. Specifically, the mAP estimation at any threshold can be greatly impacted by one or two frames on either side. Since the two actions ('move hand towards mouth' and 'move hand away from mouth') take less than a second at a frame rate of 15 frames per second (fps), we decided to compare mean average precision (mAP) at lower intersection over union (IoU) thresholds. Regardless, the mAP@0.10 score indicates that approximately 74% of all micro-actions were correctly identified with a temporal overlap of 10% with the ground truth.

6.4.2 EatSense Validation

Claim 2: On average, the change in motion speeds caused by the use of weights in the EatSense dataset is detectable.

The EatSense dataset serves as a useful test bed for musculoskeletal change detection because it replicates a change in upper-body movement (by attaching weights to the subjects' wrists). This was demonstrated by balance assessment speed tests, which were investigated by fitting linear models in [10]. The linear model hides answers to questions like 'Is there any observable change in performance as a function of the four weights?'.

By fitting a piecewise constant function across the normalized average time taken for all 27 subjects to complete the 'move hand towards mouth' micro-action across four different weights (i.e., normalized by person, as people move at different basic speeds), we further investigate whether the weights affect eating speed. A normalized duration

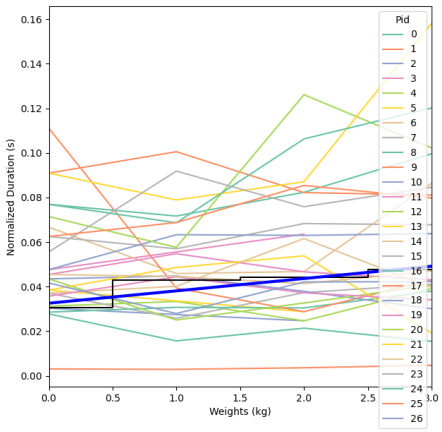


Figure 6.4: Normalized duration (time taken to move hand towards mouth for the 24 people in EatSense) versus weight plot with a piecewise constant function for each weight class shown in the black color and the least square fit shown in blue.

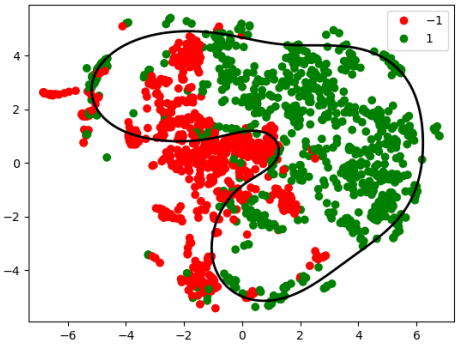


Figure 6.5: The decision boundary estimated by anomaly detector (SVM) with radial basis function as the kernel. 1 (green) indicates normal samples and -1 (red) shows anomalous samples. The black line is the decision boundary found by an SVM trained on the 30-D features.

versus weight plot is displayed in Figure 6.4, where a least square fit is displayed in blue and a piecewise constant function is displayed in black. To eliminate scale bias amongst subjects, normalization entails percentage normalization with regard to the no-weight case for each subject (i.e., when no weight was attached to the wrists). The piecewise constant function fitted is as shown in Equation 6.1.

$$f(w) = mean_p(mean_{i_p}(\frac{D_{w,p,i_p}}{\frac{1}{n_p}\sum_{j=1}^{n_p} D_{0,p,j}})) \tag{6.1}$$

where each person p has n_p action instances with weight $w = 0$. The actual time taken by subject $p \in \{0,...,26\}$ while performing action instance i_p with weight $w \in \{0,1,1.8,2.4kg\}$ is denoted by $D_{w,p,i}$. As the weights are gradually increased on the wrists, Figure 6.4 illustrates a rising trend that indicates a gradually slowing arm movement (though there is a lot of variation in the data due to measurement noise, individual eating instance variations, etc).

Claim 3: Using 30 postural features, an anomaly detection algorithm can effectively detect postural changes.

By comparing data from sessions without weights to those with the 2.4kg weights, we examine whether anomaly detection algorithms can detect changes in body (joint) pose using the balanced dataset as outlined in Section 6.4.1. The EatSense data, which is randomly divided into 80% training and 20% testing samples, is fed into a one-class Support Vector Machine (SVM) classifier with the assumption that data without weights represents normal posture and data with 2.4 kg weights may indicate anomalous posture. Note that we did not measure posture explicitly; rather the SVM is trained to distinguish between the weight 0 and weight 2.4 simply based on the 30 postural features.

For visualization purposes, only the last frame (a 30-dimensional vector) of each action instance's postural features is projected to a 2D t-SNE plot in Fig. 6.5.

The 30-D data (as chosen from Fig 6.2) is used to train an SVM one-class classification model. When classifying between normal and altered posture, the SVM with an RBF kernel achieves a 76.2% F1-score quantitatively. The F1-score is reported here because it provides a balanced measure between normal and abnormal data, being the harmonic mean of precision and recall. To visualize these results, we projected the 30-D data to 2D using t-SNE and re-estimated the boundary in 2D. Fig. 6.5 shows the 2D t-SNE plot and the decision boundary represented by the black line.

6.4.3 Holistic Weakness Detection Evaluation

Claim 4: The proposed pipeline can effectively capture the general temporal behavior trends and produce valuable insights about patterns by solely observing eating activity and tracking upper-body motion.

A holistic assessment of the suggested pipeline is carried out, alongside the examination of the individual network and classifier elements, as covered in the earlier subsections. Six new videos were shot like EatSense and are used as the test set for the system-wide evaluation. The final paragraph in Section 6.3.1 included a brief discussion of them. To summarize the key features of the videos: each person had 3 videos each with a different weight (0, 1, 2.4 kg), and where each video recorded eating food from 3 different-sized bowls (large, medium, small).

To prepare the data, the same 30 features that were previously chosen were estimated and poses were extracted, creating a 30-D feature vector for each frame, much like in the inference pipeline. The TAL network, which outputs the two action class labels along with the temporal segment for each prediction, is used to temporally localize actions in the video based on the temporal sequence of these features. Next, statistics like the number of mouthfuls, chewing time, and 'move hand towards mouth' micro-action are estimated using the temporal segments.

6.4.3.1 Duration of Hand-to-Mouth Actions

The 'move hand towards mouth' micro-action is significant because it requires the subject to work against gravity while moving their hand, and variations in the duration of the action may signal a loss of strength or control. The duration of the 'move hand towards mouth' micro-actions can be directly extracted from the TAL network results once the temporal segment has been extracted along with the class predictions.

Results and Discussion: Figure 6.6 illustrates the time taken to move the hand from the plate (where the food is) to the mouth. The first three videos correspond to subject 1, while the last three represent subject 2.

This figure highlights two notable phenomena. Firstly, the predicted output closely aligns with the pattern of the ground-truth values, indicating the effectiveness of the proposed pipeline. The highest mean difference observed is 0.07 seconds, suggesting a high level of accuracy in the predictions compared to the ground truth.

Additionally, the figure illustrates contrasting trends in the time taken by subject 1 and subject 2 to complete the 'move hand towards mouth' action. This trend reflects

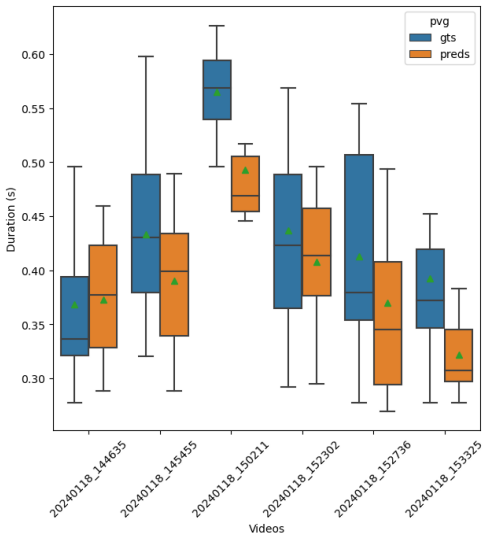


Figure 6.6: Time taken for 'move hand towards mouth' micro-action. Blue shows the ground truth values and orange shows the predicted outputs. The green triangle in the box plot shows the mean value of the distribution, the black line is the median value, and the error bars indicate 25 and 75% quartiles.

the change in performance, specifically the average time taken to complete the action, across different conditions. Given that the subjects wore three different weights on their wrists, one might expect to observe a systematic change in action duration. Subject 1 takes longer to complete the action with higher weights, indicating slower movement, possibly due to muscular inadequacy or increased effort to maintain hand-to-mouth coordination.

In contrast, subject 2 exhibits the opposite trend, taking shorter time intervals to complete the action with higher weights, which is counter-intuitive. Further investigation of the videos revealed that subject 2 tended to slouch more to offset the weight on their wrists. This resulted in shorter movement distances, as the food was now closer to the person's mouth, leading to shorter time intervals for eating with higher weights. Consequently, the weights altered the overall posture of subject 2, influencing the observed trends in action duration.

6.4.3.2 Posture Anomaly Detection

The SVM model that was previously trained on the EatSense dataset is used to classify the 3D instantaneous posture features from the new 2-person test dataset that correspond to the temporal segments predicted by the TAL network. The goal is to pick up

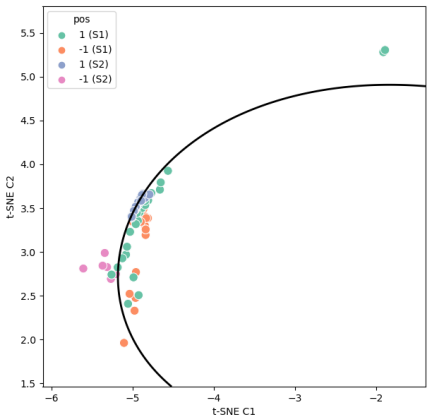


Figure 6.7: The decision boundary estimated by anomaly detector (SVM) with radial basis function as the kernel. -1/1 (Sx) in the legend indicates anomalous and normal data with respect to the subject.

any changes in posture.

Results and Discussion: The SVM model for detecting postural anomalies achieves a commendable F1-score of 71.6%. Figure 6.7 illustrates the decision boundary on the test set, with four colors representing two weights and two subjects.

For subject 1, the majority of points are classified as exhibiting normal posture, including instances where the subject wore weights (blue) and instances without weights (orange). This consistency aligns with the findings from the speed tests shown in Figure 6.6.

Conversely, for subject 2, more points classified as abnormal posture (pink) lie beyond the boundary line, while points classified as normal posture (grey) are situated within the boundary. This observation suggests a significant postural change for subject 2.

These findings corroborate the deductions made earlier regarding the time taken to complete the 'move hand to mouth' micro-action, as discussed in subsection 6.4.3.1. In summary, these results indicate that the anomaly detection algorithm effectively quantifies postural anomalies, demonstrating its utility in assessing and monitoring eating behaviors.

Table 6.2: The number of mouthfuls, ground truth versus estimated for the 'move hand to mouth' actions. W represents the weight (in Kg) the subject S1/S2 was wearing in corresponding videos and V represents the volume of the bowl in milliliters.

				Mouthfuls	
Subject	Video ID	V (ml)	W (kg)	GT	Est.
S1	20240118_144635	1000	0	17	19
	20240118_145455	625	1	10	11
	20240118_150211	250	2.4	6	6
S2	20240118_152302	1000	0	16	15
	20240118_152736	625	1	12	12
	20240118_153325	250	2.4	8	9

6.4.3.3 Mouthfuls

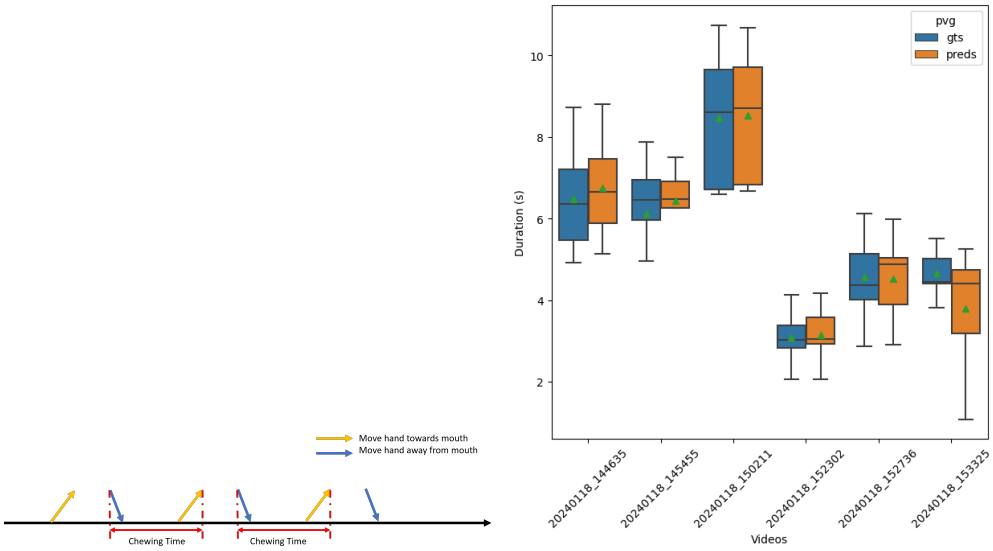
The number of times a subject makes the motion 'move hand towards mouth' is counted to determine the number of mouthfuls. The mouthful count can provide a long-term estimate of a person's nutritional sufficiency.

Results and Discussion: Table 6.2 displays the number of mouthfuls consumed by both subjects, comparing the actual amount consumed (the ground truth) to that estimated by the proposed pipeline. Remarkably, the estimated number of mouthfuls closely matches the ground truth number. This observation indicates that the proposed pipeline not only predicts the number of mouthfuls with an accuracy within ± 2 but also captures patterns in eating behaviors, such as fewer mouthfuls for smaller portions. Overall, these findings underscore the effectiveness and accuracy of the proposed pipeline in estimating mouthful counts and recognizing patterns in eating behaviors.

6.4.3.4 Chewing Duration

The time spent chewing is measured from the beginning of the 'move hand away from mouth' and to the end of the subsequent 'move hand towards mouth', assuming that there are no outside distractions and that the person does not engage in any other activity in between the two consecutive 'move hand towards mouth' actions. Fig. 6.8a illustrates this. However, even after merging the overlapping temporal segments, the TAL network's predictions—which should, in theory, have the same number of 'move hand towards mouth' and 'move hand away from mouth' actions, do not match up due

to processing errors. To overcome this issue, the estimation of the chewing duration is done if and only if 'move hand away from mouth' is followed by 'move hand towards mouth'.



(a) (b)

Figure 6.8: (a) Chewing duration is estimated when the 'move hand away from mouth' starts until 'move hand towards mouth' ends. This is done under the assumption that there is no distraction such as a phone conversation, another person to talk to, or other distractions. (b) Time taken for chewing. The green triangle shows the mean value of the distribution and the error bars indicate 25 and 75% quartile.

Results and Discussion: The assumptions outlined above are generally applicable to most elderly individuals who live independently and have a distraction-free environment. Figure 6.8b(b) depicts the chewing duration for the six videos. It is noteworthy that similar distributions of chewing time are observed across all videos. This consistency is expected since the same food was consumed in all cases. The slight variation observed, such as the potentially longer chewing time for subject 1 in the heaviest weight condition in a small bowl, may be attributed to factors like satiety, as this bowl was consumed last and the volunteer may have been nearing fullness. Importantly, the predicted chewing time closely aligns with the ground truth chewing time, underscoring the effectiveness of the pipeline. This accuracy serves as a foundation for further insights, including the identification of potential eating disorders or abnormalities in chewing behavior.

Note: The charts and tables demonstrating performance serve as indicative mea-

sures of the overall effectiveness of the pipeline and simulate performance decline under controlled conditions. However, in real-world scenarios, data collection is likely to occur every week, with tracking spanning months or even years. Operating on a weekly timescale offers several advantages. Firstly, it enables better averaging of data, smoothing out fluctuations and providing a clearer picture of trends over time. Additionally, it allows for the removal of outliers, enhancing the accuracy of the analysis. By collecting data over an extended period, the pipeline can more accurately monitor changes and trends in eating behaviors, providing valuable insights into long-term health and well-being. This approach facilitates early detection of potential issues and enables proactive interventions to support individuals' health and lifestyle management.

6.5 Conclusion

This chapter introduces a vision-based pipeline, referred to as the video-to-report (V2R) system, designed to monitor individuals' health and behaviors during eating activities, with a focus on the upper body. V2R incorporates temporal action localization to identify two key actions: 'move hand towards mouth' and 'move hand away from mouth'. Following temporal segmentation, V2R conducts two primary analyses. Firstly, it assesses eating behavior by tabulating the number of mouthfuls consumed and logging the time required for chewing. Secondly, V2R identifies musculoskeletal changes by analyzing trends in the duration of the 'move hand to mouth' micro-action and detecting postural anomalies using a one-class SVM classifier.

The effectiveness of the proposed pipeline is demonstrated through various metrics. For instance, individual components of the pipeline achieve notable accuracies, such as the Temporal Action Localization (TAL) network achieving a mean Average Precision (mAP) of 74% at IoU threshold 0.10, and the anomaly detection module (SVM) achieving 71% accuracy. Moreover, holistic tests conducted on a newly presented test set, an extension of EatSense, show the pipeline's capability to successfully capture trends and patterns. Overall, these results underscore the effectiveness and potential utility of the V2R system in monitoring and analyzing eating behaviors and associated health parameters.

Chapter 7

Conclusions

In this thesis, we explored non-intrusive methods to monitor the changes in upper-body activity levels while eating, to help support healthy aging and living independently. The overall contribution of this thesis is motion capture and analysis strictly based on upper-body movements using eating activities. We investigated and contributed towards four major issues for upper-body motion analysis using eating activities: (1) a new vision-based dataset that models eating behavior and simulates performance decline in the upper-body movements (see chapter 3), (2) eating sub-action recognition and localization (see chapter 4), (3) across-subject generalization under subjective bias (see chapter 5), and (4) a fully autonomous pipeline for motion analysis 6.

7.1 Summary of contributions

In **chapter 3**, we presented a new vision-based dataset named EatSense for both the healthcare and computer-vision communities. It is a moderately large database, recorded using an Intel RealSense RGB-D camera, and comprises 27 subjects spanned over 135 videos, each 11.4 minutes long on average with 16 sub-actions based on limb motion and hand gestures. EatSense has several characteristics unlike many publicly available datasets: (1) Multiple layers of abstraction of labels including, dense atomic labels challenging for both sub-action recognition and temporal action localization in untrimmed videos, and four weight classes (based on the weight attached to their wrists, 0, 1 kg, 1.8 kg, and 2.4 kg) for performance decline classification and regression. (2) Behavioral understanding capability, i.e., the capacity to model the complete eating behavior of individuals. (3) Simulating performance decline in the upper body of an individual over time using different weight classes. To the best of

our knowledge, EatSense is the first of its kind multi-disciplinary dataset to have all the above qualities.

In **chapter 4**, to understand the eating behavior of individuals, we explored two categories of action understanding techniques: action recognition (AR) and temporal action localization (TAL). Through various experiments using both deep-learning-based end-to-end AR techniques (section 4.4) and hand-crafted features-based AR (section 4.5) across different modalities, we demonstrated that most common deep learning based AR algorithms achieve over 75% Top-1 accuracy and common hand-crafted features based algorithms achieve over 42% top-1 accuracy using EatSense. On the other hand, for TAL algorithms EatSense posed to be a more challenging dataset (section 4.7.1). We further explored this issue and highlighted the shortcomings of the current TAL networks (section 4.7.2): (1) currently, TAL networks are unable to efficiently handle hugely varying lengths of actions within a video, (2) the current metric to measure accuracy of TAL networks (i.e., mean average precision) is very sensitive, (3) if we remove noisy action classes in EatSense, we can improve the performance of the network.

In **chapter 5**, we utilize the weight labels (0,1,1.8 and 2.4 kg) available in EatSense to assess changes in performance levels of the subjects. Firstly, to quantify the effects of these weight-induced changes, this chapter explores tests, such as balance assessments (trunk stability) and speed of motion assessments, both of which are modified to fit the vision-based eating scenario. Secondly, we presented a pipeline for assessing the changes in the performance decline of the subjects and posed it as a regression problem with a focus on only two micro-actions 'move hand towards mouth' and 'move hand away from mouth'. This performance decline is measured through multiple machine/deep learning models, including Gaussian Mixture Regression (GMR), Multilayer Perceptron (MLP), LightGBM, and one deep learning model ST-GCN. However, due to limited data and subjective bias (different subjects show a decline in performance in different ways), the outputs seemed to overfit. To tackle this problem, the proposed pipeline utilized the most contributing features of each subject and leveraged an uncertainty-aware Gaussian mixture regression algorithm. GMR is effective because it is a probabilistic model that can handle uncertainty and multi-modal data distributions—meaning it can better model the variability in how different people react to added weights. Through various experiments, this chapter also highlighted the effectiveness of EatSense for assessing the performance levels of individuals while they eat.

In **chapter 6**, we presented a vision-based fully autonomous pipeline called V2R (video to report) that takes a video of the subject under observation and generates various health statistics such as musculoskeletal performance decline and eating behavioral logs. This was done from only the upper body and hand-crafted features extracted using the domain knowledge to ensure the explainability of the inputs/outputs of the models. Its primary functions include analyzing eating behaviors like chewing and mouthfuls, detecting changes in arm movement speed, and identifying postural anomalies. This system is designed to offer healthcare workers a comprehensive report, enabling them to monitor long-term changes in an individual's motor abilities. We propose a pipeline that uses the video recorded in a similar setting as the videos in EatSense: (1) extracts 2D poses and lifts them into the 3D space using the depth maps, (2) uses 3D poses to extract features, and then uses forward sequential feature extractor to select most contributing features, (3) use transformer based TAL network to localize two important atomic actions and (4) generate statistics and draw insights using the temporal segments estimated by TAL. We demonstrated the technique's effectiveness by extensive experiments component-wise and over the entire proposed pipeline (section 6.4.3).

7.2 Limitations

This project was conducted as part of the Advanced Care Research Center's (ACRC) work package 6 (WP6), a center established by the University of Edinburgh to promote independent living for the elderly. Through consultations with geriatric experts such as Dr. Bruce Guthrie and Dr. Susan Shenkin from the University of Edinburgh, it became clear that measuring activity levels and assessing eating behaviors would be highly beneficial for individuals suffering from Parkinson's disease and for patients undergoing rehabilitation. These discussions played a significant role in shaping the direction of our thesis. Unfortunately, due to time and resource constraints, we were unable to include participants specifically diagnosed with Parkinson's disease.

Nevertheless, we engaged with care facilities like Braid Care Home, where several individuals expressed interest and volunteered to participate in real-world testing of the project. One major limitation we encountered in involving participants from the Braid Center was the ongoing process of obtaining ethical approvals. WP6 mandated collective ethical approval for all projects under the Schools of Engineering and Informatics, which is still pending and has delayed several aspects of the testing phase of this thesis

and other ongoing research in WP6 as well.

The EatSense dataset generated interest from researchers in Germany and New Zealand, who are interested in utilizing it for their studies, such as facial expression detection for swallowing and taste recognition in food. However, due to our project's limited resources and timeline, we were unable to provide the necessary data labels to support these specific use cases. Addressing these requirements has been earmarked as part of future work, as outlined in Section 7.3. The limitations of the contributions in the thesis in a chapter-wise format are discussed below:

- **Chapter 3** presents a dataset called EatSense. Although EatSense is a large-scale dataset, in its current state, it has several limitations:

    1. Limited data for sub-actions: when trained and tested on a subset of the data, the model seems to overfit due to a limited amount of data and diversity.
    2. Limited cultural diversity: Currently, EatSense contains only 27 subjects from 12 nationalities which could make it challenging to get a generalized model.
    3. Complexity of Sub-Action Annotations: The dense labeling of atomic actions is advantageous for providing insights into complex activities by breaking them down into smaller components. However, achieving consistent and accurate annotations may be challenging due to the large number of actions present in a video, particularly when scaling the dataset.
    4. Fixed Camera Setup: The research presented in this thesis uses 3D spatial information to extract features and is independent of small changes in camera angle and distance w.r.t. the person. However, the limited range of data capture positions may limit the application, which may not generalize well to large changes in the camera angle, distance, and other settings.

- **Chapter 4** explores action recognition and temporal action localization to understand eating behaviors. The analysis uses both deep learning models and hand-crafted features in an effort to balance explainability. Nevertheless, there may be a trade-off: hand-crafted feature-based explainable models may perform worse than more intricate, difficult-to-understand models that may offer greater accuracy but less transparency.

- **Chapter 5** presents a pipeline to explore the generalization ability of existing models on EatSense, to assess the performance levels of the subjects. The proposed pipeline has several limitations:

    1. Forward sequential feature selector used in the system is not exhaustive as it only checks at most $\frac{1}{2}n^2$ combinations out of 2^n, where $n = 472$ is the number of features. Two key implications of this partial search are: 1) the risk of suboptimal feature selection since it selects features that provide the best improvement in each step, 2) the potential for feature redundancy as it does not account for the possible interactions between features that are only evident when they are selected together.
    2. Adding weights has not been previously validated to cause the same effect as ageing. Weights might be one way to simulate changes in motor function, but they may not adequately capture the intricacy of natural motor performance decline, especially in older adults or people with certain medical conditions.
    3. Individual differences exist in eating habits and motor performance levels because of things like age, health, and cultural customs. Without customized baselines, the system could misjudge performance levels or label normal behavior as problematic—especially for people with slower natural eating.
    4. This chapter explored detecting performance decline within two demographics of the dataset (age groups and gender). However, the dataset didn't record other demographics such as height and weight of the subjects which could be explored to draw other interesting insights into the eating characteristics. Especially the balance assessment and speed tests could have benefited by finding intersections of affected/non-affected groups.
    5. Although the balance assessment and speed of motion tests showed that there is possibly a performance decline, it is not validated with any other sensor (e.g., a camera at the side of the subject could show slouch very clearly).

- **Chapter 6** presents a vision of a fully autonomous system. Some limitations of the proposed system are:

    1. V2R is still a TRL-4 system. Hence, it needs more testing in the end-user

environments or similar to understand how elderly people will adopt the technology.

2. V2R currently does not work in real-time since it relies on accurate pose estimation that requires bigger models such as HigherHRNet for accurate pose estimation.

3. V2R estimates several useful statistics using a single camera setup. Some of these can potentially be validated if another sensor such as an IMU was attached to the wrist of the subject or another camera at some angle w.r.t. the first camera was placed to see the movement of the subject.

4. The proposed pipeline was tested on EatSense dataset only. It may have benefited more if we could have validated our claims through another dataset. Unfortunately, no other strictly vision-based dataset(s) exists for our very specific use case.

7.3 Future Works

Future work using EatSense could explore several promising directions. One future work could be to extend the capabilities of EatSense by introducing new label abstractions such as bounding box marked tools (fork/knife/spoon etc.) and utensils (plate/bowl etc.), and facial expressions to indicate dislike or satiety, etc. This could target multiple computer vision problems such as: (1) how to detect objects such as tools in a video that are small and are ambiguous because they look similar from various angles, (2) the detection of plates when they are occluded with food, (3) use facial expressions to understand behavioral characteristics of the individual. These labels also open up pathways to use human object interaction and facial expression understanding to further improve the performance of action recognition and temporal action localization techniques.

In chapter 4 we used temporal action localization to better understand the actions as it exploits temporal relationships in the whole video. However, the temporal action localization algorithms did not perform well enough. Another potential future work direction from a computer vision perspective could be developing algorithms to better handle the complex and unstructured nature of eating activities captured in untrimmed videos. This could involve incorporating multi-scale temporal modeling and attention mechanisms to improve the detection and segmentation of eating actions and designing

scenario-based less sensitive metrics for evaluating TAL networks.

Vision-based algorithms need image/video as an input and this raises concerns over the privacy of the individual. In this thesis, we briefly explored the effects of facial obfuscation techniques on pose-based action recognition to preserve the privacy of individuals in EatSense, and the results show near to no observable effect on the accuracy. However, this analysis was limited because we only used the EatSense dataset. This analysis can be extended by including other publicly available datasets to solidify this hypothesis.

Currently, due to limited data, attaining models that are generalizable across subjects is a problem. We proposed to solve it using an uncertainty-aware algorithm and achieved sub-par performance. Expanding the dataset with more diverse subjects and settings could help in creating generalized models that apply to a wider population, addressing the variability in eating habits and physical conditions. Another viewpoint to explore this problem could be few-shot learning approaches to enable the model to generalize better from limited labeled data, making it more adaptable to new environments and subjects. Lastly, another potential future direction worth exploring could involve analyzing observational differences by categorizing subjects based on different aspects of the gathered demographic data.

Lastly, researchers could also investigate real-time applications of EatSense in monitoring and supporting elderly individuals in their homes, potentially integrating the dataset with smart home systems for continuous health assessment. These real-time systems could alert caregivers or medical professionals about irregularities in eating patterns, potentially preventing health issues.

Appendix A

Action Recognition and Temporal Action Localization

A.1 Naming Convention of the features

This section includes the table A.1 that elaborates the nomenclature and dimensions spanned by the entire feature space.

A.2 Action Recognition with Hand-Crafted Features

This section includes the experiment briefly discussed in the text to cater to human action labeling errors.

To account for human error and overcome the frame-wise classifier offset in the labels, a windowed search of sizes ranging from ± 0 to ± 21 (equivalent to $\pm$ 0 seconds to $\pm$ 1.5 seconds) frames was applied to look for the correct label in that particular range of frames, between the ground truth and the predictions. Fig. A.1 shows the graph with varying window size as the x-axis and accuracy as the window size is increased. It shows that with a tolerance of ± 2 frames, 44.9% accuracy can be achieved for the unseen videos when using the LightGBM (focal loss) classifier.

The figure also shows better performance of nearest neighbors when used with window sizes higher than ± 5 (i.e., the full window size of 11). However, the results with anything over the window size of ± 3 might be unreliable and severely noisy since some of the actions such as 'move hand towards mouth' and 'move hand away from mouth' last about 9-13 frames on average (at 15 fps). We only report the results at ± 2 window size in the table 4.3.

Table A.1: Characteristics of the feature space such as dimensions (dims), vector size, and the nomenclature in this thesis are as follows. For vector size, m represents the number of frames. In the fourth column, $\{joints\}$ represent the name of the joint such as left-shoulder, $\{x/y/z\}$ show their corresponding coordinate axis, $W\{1/2/3/4/5/6\}$ represents various window sizes (k) set for the calculation of acceleration and velocity.

Feature	Vector Size	Dims.	Notation in data
Absolute Positions	$m \times 8 \times 3$	24	$\{joints\}\_\{x/y/z\}$
Relative Positions (w.r.t. 'c')	$m \times 7 \times 3$	21	$\{joints\}\_\{x/y/z\}\_reldist$
Euclidean Distance $j \in \{w,e\}$	$m \times 2$	2	$\{joints\}\_dist\_B$
Polar Coordinate Positions	$m \times 7$	7	$\{joints\}\_reldist\_3d$
Product $j \in \{n,s,e,w\}$	$m \times 1$	1	$joints\_prod$
Angles $j \in \{s,e\}$	$m \times 4$	4	$\{left/right\}Arm\_\{joints\}\_angles$
Velocity	$m \times 7 \times 4 \times 6$	168	$\{joints\}\_\{x/y/z\}\_relvel\_W\{1,...,6\}$
Acceleration	$m \times 7 \times 4 \times 6$	168	$\{joints\}\_\{x/y/z\}\_relAcc\_W\{1,...,6\}$
lags	$m \times 7 \times 4 \times 1$	28	$\{joints\}\_\{x/y/z\}\_lag1$
Weighted Moving window	$m \times 7 \times 3 \times 1$	21	$\{joints\}\_\{x/y/z\}\_reldist\_movw3$
Magnitude	$m \times 7 \times 1 \times 1$	8	$\{joints\}\_mag$
Distance from table	$m \times 8 \times 1 \times 1$	8	$\{joints\}\_disttable$

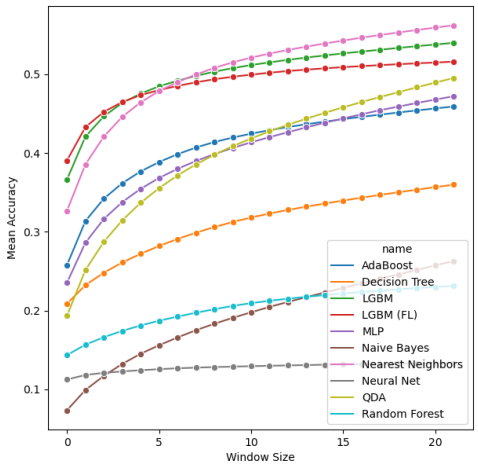


Figure A.1: The effect of window sizes (varying from ± 0 to ± 21) on the accuracy of the algorithms used for frame-wise action recognition with hand-crafted features.

Appendix B

Activities of Daily Living

B.1 List of Activities of Daily Living

Activities of daily living are broadly divided into two categories, i.e., activities of daily living (ADLs) or instrumental ADLs (IADLs) [199]. The term ADL is a collective term of all the basic skills needed to perform daily tasks. Here we list the common activities of daily living.

- **Ambulating:** This encompasses activities such as walking, sitting, standing, lying down, getting up, and climbing stairs indoors and outdoors.
- **Grooming:** This includes the activities essential for maintaining personal hygiene, such as brushing teeth, shaving, bathing, and caring for your hair/nails.
- **Toileting:** This involves managing bladder and bowel functions (continence), safely using the toilet, and maintaining personal hygiene afterward.
- **Dressing:** This entails dressing oneself appropriately, including the capability to effectively use buttons and zips.
- **Eating:** This includes the ability to use cutlery such as forks/knives/spoons and feed themself.

On the other hand, IADLs are more complex activities that are important to live independently. Here is a list of 12 essential IADLs.

- Handling personal finances, such as paying bills, budgeting, and utilizing banking services.

- Managing personal health, including medical appointments and following prescribed treatments.
- Performing personal shopping for groceries and other essentials.
- Preparing and cooking meals for oneself.
- Arranging transportation, including driving, using taxis, and public transit.
- Using communication tools like phones, mail, email, and other devices.
- Completing household chores such as cleaning, gardening, and laundry.
- Taking care of pets.
- Providing childcare.
- Assisting others, which may include overseeing caregivers.
- Engaging in hobbies, or other personal interests.
- Knowing emergency procedures, contacts, and responses.

Bibliography

[1] W. H. Organization. (2022) Ageing and health. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[2] M. Javdan, M. Ghasemaghaei, and M. Abouzahra, "Psychological barriers of using wearable devices by seniors: a mixed-methods study," *Computers in Human Behavior*, vol. 141, p. 107615, 2023.

[3] A. Maier, A. G. Özkil, M. M. Bang, and B. H. Forchhammer, "Remember to remember: A feasibility study adapting wearable technology to the needs of people aged 65 and older with mild cognitive impairment (mci) and alzheimer's dementia," in *20th International Conference on Engineering Design: Design for Life*. Design Society, 2015, pp. 331–340.

[4] T. Hagendorff, "Linking human and machine behavior: A new approach to evaluate training data quality for beneficial machine learning," *Minds and Machines*, vol. 31, no. 4, pp. 563–593, 2021.

[5] X. Ding, Q. Gan, and S. Bahrami, "A systematic survey of data mining and big data in human behavior analysis: Current datasets and models," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 9, p. e4574, 2022.

[6] K. McManus, B. R. Greene, L. G. M. Ader, and B. Caulfield, "Development of data-driven metrics for balance impairment and fall risk assessment in older adults," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2324–2332, 2022.

[7] H. Alwassel, F. C. Heilbron, V. Escorcia, and B. Ghanem, "Diagnosing error in temporal action detectors," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 256–272.

[8] M. A. Raza, L. Chen, N. Li, and R. B. Fisher, "Eatsense: Human centric, action recognition and localization dataset for understanding eating behaviors and quality of motion assessment," *Image and Vision Computing*, vol. 137, p. 104762, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885623001361

[9] M. A. Raza, C. Lochhead, and R. B. Fisher, "Effect of face obfuscation methods on pose-based action recognition," in *International Conference on AI in Healthcare*, 2024.

[10] M. A. Raza and R. B. Fisher, "Vision-based approach to assess performance levels while eating," *Machine Vision and Applications*, vol. 34, no. 6, p. 124, 2023.

[11] G. H. Louie and M. M. Ward, "Sex disparities in self-reported physical functioning: true differences, reporting bias, or incomplete adjustment for confounding?" *Journal of the American Geriatrics Society*, vol. 58, no. 6, pp. 1117–1122, 2010.

[12] M. Tieland, I. Trouwborst, and B. C. Clark, "Skeletal muscle performance and ageing," *Journal of cachexia, sarcopenia and muscle*, vol. 9, no. 1, pp. 3–19, 2018.

[13] G. Anderson, J. Horvath, J. Knickman, D. Colby, S. Schear, and M. Jung, "Chronic conditions: making the case for ongoing care," 2002.

[14] W. H. Organization *et al.*, "Supporting informal caregivers of people living with dementia," *Geneva: World Health Organization*, 2015.

[15] J. Eden and R. Schulz, *Families caring for an aging America*. National Academies Press, 2016.

[16] D. Redfoot, L. Feinberg, and A. N. Houser, *The aging of the baby boom and the growing care gap: A look at future declines in the availability of family caregivers*. AARP Public Policy Institute Washington, DC, 2013.

[17] C. B. Fausset, A. J. Kelly, W. A. Rogers, and A. D. Fisk, "Challenges to aging in place: Understanding home maintenance difficulties," *Journal of Housing for the Elderly*, vol. 25, no. 2, pp. 125–141, 2011.

[18] S. Yeung, N. L. Downing, L. Fei-Fei, A. Milstein *et al.*, "Bedside computer vision-moving artificial intelligence from driver assistance to patient safety," *N Engl J Med*, vol. 378, no. 14, pp. 1271–1273, 2018.

[19] G. Acampora, D. J. Cook, P. Rashidi, and A. V. Vasilakos, "A survey on ambient intelligence in healthcare," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2470–2494, 2013.

[20] T. Van Kasteren, G. Englebienne, and B. J. Kröse, "An activity monitoring system for elderly care using generative and discriminative models," *Personal and ubiquitous computing*, vol. 14, pp. 489–498, 2010.

[21] S. Ahmed, S. Irfan, N. Kiran, N. Masood, N. Anjum, and N. Ramzan, "Remote health monitoring systems for elderly people: a survey," *Sensors*, vol. 23, no. 16, p. 7095, 2023.

[22] B. Ma, J. Yang, F. K. Y. Wong, A. K. C. Wong, T. Ma, J. Meng, Y. Zhao, Y. Wang, and Q. Lu, "Artificial intelligence in elderly healthcare: A scoping review," *Ageing Research Reviews*, vol. 83, p. 101808, 2023.

[23] I. G. Fernandez, S. A. Ahmad, and C. Wada, "Inertial sensor-based instrumented cane for real-time walking cane kinematics estimation," *Sensors*, vol. 20, no. 17, p. 4675, 2020.

[24] E. Ferrari, M. Gamberi, F. Pilati, and A. Regattieri, "Motion analysis system for the digitalization and assessment of manual manufacturing and assembly processes," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 411–416, 2018.

[25] T. Steinebach, E. H. Grosse, C. H. Glock, J. Wakula, and A. Lunin, "Accuracy evaluation of two markerless motion capture systems for measurement of upper extremities: Kinect v2 and captiv," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 30, no. 4, pp. 291–302, 2020.

[26] S. Salisu, N. I. R. Ruhaiyem, T. A. E. Eisa, M. Nasser, F. Saeed, and H. A. Younis, "Motion capture technologies for ergonomics: A systematic literature review," *Diagnostics*, vol. 13, no. 15, p. 2593, 2023.

[27] T. Monnet, M. Samson, A. Bernard, L. David, and P. Lacouture, "Measurement of three-dimensional hand kinematics during swimming with a motion capture system: a feasibility study," *Sports Engineering*, vol. 17, pp. 171–181, 2014.

[28] K. Takayasu, K. Yoshida, T. Mishima, M. Watanabe, T. Matsuda, and H. Kinoshita, "Upper body position analysis of different experience level surgeons during laparoscopic suturing maneuvers using optical motion capture," *The American Journal of Surgery*, vol. 217, no. 1, pp. 12–16, 2019.

[29] G. Pavei, F. Salis, A. Cereatti, and E. Bergamini, "Body center of mass trajectory and mechanical energy using inertial sensors: A feasible stride?" *Gait & posture*, vol. 80, pp. 199–205, 2020.

[30] J. Lebleu, T. Gosseye, C. Detrembleur, P. Mahaudens, O. Cartiaux, and M. Penta, "Lower limb kinematics using inertial sensors during locomotion: Accuracy and reproducibility of joint angle calculations with different sensor-to-segment calibrations," *Sensors*, vol. 20, no. 3, p. 715, 2020.

[31] S. A. Lavender, C. M. Sommerich, S. Bigelow, E. B. Weston, K. Seagren, N. A. Pay, D. Sillars, V. Ramachandran, C. Sun, Y. Xu *et al.*, "A biomechanical evaluation of potential ergonomic solutions for use by firefighter and ems providers when lifting heavy patients in their homes," *Applied ergonomics*, vol. 82, p. 102910, 2020.

[32] S. Chakraborty, A. Nandy, T. Yamaguchi, V. Bonnet, and G. Venture, "Accuracy of image data stream of a markerless motion capture system in determining the local dynamic stability and joint kinematics of human gait," *Journal of biomechanics*, vol. 104, p. 109718, 2020.

[33] A. Sabo, S. Mehdizadeh, K.-D. Ng, A. Iaboni, and B. Taati, "Assessment of parkinsonian gait in older adults with dementia via human pose tracking in video data," *Journal of neuroengineering and rehabilitation*, vol. 17, pp. 1–10, 2020.

[34] E. Parrilla, A.-V. Ruescas, J.-A. Solves, A. Ballester, B. Nacher, S. Alemany, and D. Garrido, "A methodology to create 3d body models in motion," in *Advances in Simulation and Digital Human Modeling: Proceedings of the AHFE 2020 Virtual Conferences on Human Factors and Simulation, and Digital Human Modeling and Applied Optimization, July 16-20, 2020, USA*. Springer, 2021, pp. 309–314.

[35] F. Schlagenhauf, S. Sreeram, and W. Singhose, "Comparison of kinect and vicon motion capture of upper-body joint angle tracking," in *2018 IEEE 14th in-*

ternational conference on control and automation (ICCA). IEEE, 2018, pp. 674–679.

[36] P.-L. Liu, C.-C. Chang, L. Li, and X. Xu, "A simple method to optimally select upper-limb joint angle trajectories from two kinect sensors during the twisting task for posture analysis," *Sensors*, vol. 22, no. 19, p. 7662, 2022.

[37] W. W. Lam and K. N. Fong, "Validity and reliability of upper limb kinematic assessment using a markerless motion capture (mmc) system: A pilot study," *Archives of Physical Medicine and Rehabilitation*, vol. 105, no. 4, pp. 673–681, 2024.

[38] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[39] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 455–472.

[40] L. Qiu, X. Zhang, Y. Li, G. Li, X. Wu, Z. Xiong, X. Han, and S. Cui, "Peeking into occluded joints: A novel framework for crowd pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer, 2020, pp. 488–504.

[41] L. Xu, Y. Guan, S. Jin, W. Liu, C. Qian, P. Luo, W. Ouyang, and X. Wang, "Vipnas: Efficient video pose estimation via neural architecture search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 072–16 081.

[42] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[43] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5386–5395.

[44] G. Wei, C. Lan, W. Zeng, and Z. Chen, "View invariant 3d human pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4601–4610, 2019.

[45] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, "Hemlets posh: Learning part-centric heatmap triplets for 3d human pose and shape estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3000–3014, 2021.

[46] Y. Zhan, F. Li, R. Weng, and W. Choi, "Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 116–13 125.

[47] I. D. Realsense. Projection in intel realsense sdk 2.0. [Online]. Available: https://dev.intelrealsense.com/docs/projection-in-intel-realsense-sdk-20

[48] Intel. Intel realsense cross-platform api. [Online]. Available: https://intelrealsense.github.io/librealsense/doxygen/rsutil_8h.html

[49] Y. Chiba. (2020) Converting 2d image coordinates to 3d coordinates using ros + intel realsense d435/kinect. [Online]. Available: https://medium.com/@yasuhirachiba/converting-2d-image-coordinates-to-3d-coordinates-using-ros-intel-realsense-d435-kinect-88621e8e7

[50] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[51] Y. Cheng, B. Wang, and R. T. Tan, "Dual networks based 3d multi-person pose estimation from monocular video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1636–1651, 2022.

[52] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 2014, pp. 556–571.

[53] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," in *European conference on computer vision*. Springer, 2022, pp. 422–438.

[54] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 304–313.

[55] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4902–4910.

[56] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7919–7928.

[57] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9839–9848.

[58] M. D. Hssayeni, J. Jimenez-Shahed, M. A. Burack, and B. Ghoraani, "Wearable sensors for estimation of parkinsonian tremor severity during free body movements," *Sensors*, vol. 19, no. 19, p. 4215, 2019.

[59] A. Marcante, R. Di Marco, G. Gentile, C. Pellicano, F. Assogna, F. E. Pontieri, G. Spalletta, L. Macchiusi, D. Gatsios, A. Giannakis *et al.*, "Foot pressure wearable sensors for freezing of gait detection in parkinson's disease," *Sensors*, vol. 21, no. 1, p. 128, 2020.

[60] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward unobtrusive in-home gait analysis based on radar micro-doppler signatures," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2629–2640, 2019.

[61] A. Paiement, L. Tao, S. Hannuna, M. Camplani, D. Damen, and M. Mirmehdi, "Online quality assessment of human movement from skeleton data," in *British Machine Vision Conference*. BMVA press, 2014, pp. 153–166.

[62] A. A. Osman, T. Bolkart, and M. J. Black, "Star: Sparse trained articulated human body regressor," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 598–613.

[63] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[64] H. Jeon, W. Lee, H. Park, H. J. Lee, S. K. Kim, H. B. Kim, B. Jeon, and K. S. Park, "Automatic classification of tremor severity in parkinson's disease using a wearable device," *Sensors*, vol. 17, no. 9, p. 2067, 2017.

[65] W. Liu, X. Lin, X. Chen, Q. Wang, X. Wang, B. Yang, N. Cai, R. Chen, G. Chen, and Y. Lin, "Vision-based estimation of mds-updrs scores for quantifying parkinson's disease tremor severity," *Medical Image Analysis*, vol. 85, p. 102754, 2023.

[66] M. B. Shaikh and D. Chai, "Rgb-d data-based action recognition: A review," *Sensors*, vol. 21, no. 12, p. 4246, 2021.

[67] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[68] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 2949–2958.

[69] G. Bertasius, H. Soo Park, S. X. Yu, and J. Shi, "Am i a baller? basketball performance assessment from first-person videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2177–2185.

[70] Z. Lei, B. Y. Tan, N. P. Garg, L. Li, A. Sidarta, and W. T. Ang, "An intention prediction based shared control system for point-to-point navigation of a robotic wheelchair," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8893–8900, 2022.

[71] T. N. Nguyen, H. H. Huynh, and J. Meunier, "3d reconstruction with time-of-flight depth camera and multiple mirrors," *IEEE Access*, vol. 6, pp. 38 106–38 114, 2018.

[72] R. D. Rondinelli, W. Dunn, K. M. Hassanein, C. A. Keesling, S. C. Meredith, T. L. Schulz, and N. J. Lawrence, "A simulation of hand impairments: effects on upper extremity function and implications toward medical impairment rating and disability determination," *Archives of physical medicine and rehabilitation*, vol. 78, no. 12, pp. 1358–1363, 1997.

[73] S. Ishikawa, S. Okamoto, K. Isogai, Y. Akiyama, N. Yanagihara, and Y. Yamada, "Wearable dummy to simulate joint impairment: severity-based assessment of simulated spasticity of knee joint," in *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*. IEEE, 2013, pp. 300–305.

[74] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[75] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[76] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[77] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[78] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *International Conf. on Computer Vision (ICCV)*, Dec. 2013, pp. 3192–3199.

[79] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[80] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2616–2625.

[81] L. Chunhui, H. Yueyu, L. Yanghao, S. Sijie, and L. Jiaying, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.

[82] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[83] L. Tao, T. Burghardt, S. Hannuna, M. Camplani, A. Paiement, D. Damen, M. Mirmehdi, and I. Craddock, "A comparative home activity monitoring study using visual and inertial sensors," in *2015 17th International Conference on E-health Networking, Application & Services (HealthCom)*. IEEE, 2015, pp. 644–647.

[84] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012. [Online]. Available: https://arxiv.org/abs/1212.0402

[85] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 720–736.

[86] Z. Tang and A. Hoover, "A new video dataset for recognizing intake gestures in a cafeteria setting," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 4399–4405.

[87] A. Hoover, "Data description: Clemson cafeteria dataset," *Online, URL: http://cecas. clemson. edu/ahoover/cafeteria*, 2020.

[88] S. Bi and D. Kotz, "Eating detection with a head-mounted video camera," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, 2022, pp. 60–66.

[89] P. A. Neves, J. Simões, R. Costa, L. Pimenta, N. J. Gonçalves, C. Albuquerque, C. Cunha, E. Zdravevski, P. Lameski, N. M. Garcia *et al.*, "Thought on food: A systematic review of current approaches and challenges for food intake detection," *Sensors*, vol. 22, no. 17, p. 6443, 2022.

[90] P. V. Rouast, H. Heydarian, M. T. Adam, and M. E. Rollo, "Oreba: A dataset for objectively recognizing eating behavior and associated intake," *IEEE Access*, vol. 8, pp. 181 955–181 963, 2020.

[91] C. A. Merck, C. Maher, M. Mirtchouk, M. Zheng, Y. Huang, and S. Kleinberg, "Multimodality sensing for eating recognition." in *PervasiveHealth*, 2016, pp. 130–137.

[92] Y. Shen, J. Salley, E. Muth, and A. Hoover, "Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables," *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 599–606, 2016.

[93] K. Kyritsis, C. Diou, and A. Delopoulos, "Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2325–2334, 2019.

[94] Q. Men, H. Leung, and Y. Yang, "Self-feeding frequency estimation and eating action recognition from skeletal representation using kinect," *World Wide Web*, vol. 22, pp. 1343–1358, 2019.

[95] A. Iosifidis, E. Marami, A. Tefas, I. Pitas, and K. Lyroudia, "The mobiserv-aiia eating and drinking multi-view database for vision-based assisted living," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, no. 2, pp. 254–273, 2015.

[96] L. Onofri, P. Soda, M. Pechenizkiy, and G. Iannello, "A survey on using domain and contextual knowledge for human activity recognition in video streams," *Expert Systems with Applications*, vol. 63, pp. 97–111, 2016.

[97] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 9–14.

[98] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[99] J. Ortells, M. T. Herrero-Ezquerro, and R. A. Mollineda, "Vision-based gait impairment analysis for aided diagnosis," *Medical & biological engineering & computing*, vol. 56, no. 9, pp. 1553–1564, 2018.

[100] D. Sethi, S. Bharti, and C. Prakash, "A comprehensive survey on gait analysis: History, parameters, approaches, pose estimation, and future work," *Artificial Intelligence in Medicine*, p. 102314, 2022.

[101] M. Moro, G. Marchesi, F. Hesse, F. Odone, and M. Casadio, "Markerless vs. marker-based gait analysis: A proof of concept study," *Sensors*, vol. 22, no. 5, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/5/2011

[102] Y. Makihara, M. S. Nixon, and Y. Yagi, "Gait recognition: Databases, representations, and applications," *Computer Vision: A Reference Guide*, pp. 1–13, 2020.

[103] Y. Sun, J. S. Hare, and M. S. Nixon, "Detecting heel strikes for gait analysis through acceleration flow," *IET Computer Vision*, vol. 12, no. 5, pp. 686–692, 2018.

[104] L. Wang, G. Zhao, N. Rajpoot, and M. S. Nixon, "Special issue on new advances in video-based gait analysis and applications: challenges and solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 4, pp. 982–985, 2010.

[105] S. Fostinelli, R. De Amicis, A. Leone, V. Giustizieri, G. Binetti, S. Bertoli, A. Battezzati, and S. F. Cappa, "Eating behavior in aging and dementia: the need for a comprehensive assessment," *Frontiers in nutrition*, vol. 7, p. 604488, 2020.

[106] H. Hiraguchi, P. Perone, A. Toet, G. Camps, and A.-M. Brouwer, "Technology to automatically record eating behavior in real life: A systematic review," *Sensors*, vol. 23, no. 18, p. 7757, 2023.

[107] H. Heydarian, M. Adam, T. Burrows, C. Collins, and M. E. Rollo, "Assessing eating behaviour using upper limb mounted motion sensors: A systematic review," *Nutrients*, vol. 11, no. 5, p. 1168, 2019.

[108] E. Curto and H. Araujo, "An experimental assessment of depth estimation in transparent and translucent scenes for intel realsense d415, sr305 and l515," *Sensors*, vol. 22, no. 19, p. 7378, 2022.

[109] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[110] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[111] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[112] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[113] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019. [Online]. Available: https://doi.org/10.1145/3343031.3350535

[114] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2020.

[115] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.

[116] J. Ortells, M. T. Herrero-Ezquerro, and R. A. Mollineda, "Vision-based gait impairment analysis for aided diagnosis," *Medical & biological engineering & computing*, vol. 56, pp. 1553–1564, 2018.

[117] K. Jun, Y. Lee, S. Lee, D.-W. Lee, and M. S. Kim, "Pathological gait classification using kinect v2 and gated recurrent neural networks," *Ieee Access*, vol. 8, pp. 139 881–139 891, 2020.

[118] H. Hukkelås and F. Lindseth, "Does image anonymization impact computer vision training?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 140–150.

[119] J. Jiang, W. Skalli, A. Siadat, and L. Gajny, "Effect of face blurring on human pose estimation: Ensuring subject privacy for medical and occupational health applications," *Sensors*, vol. 22, no. 23, p. 9376, 2022.

[120] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *International symposium on visual computing*. Springer, 2019, pp. 565–578.

[121] H. Hukkelås and F. Lindseth, "Deepprivacy2: Towards realistic full-body anonymization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1329–1338.

[122] L. Xu, Y. Guan, S. Jin, W. Liu, C. Qian, P. Luo, W. Ouyang, and X. Wang, "Vipnas: Efficient video pose estimation via neural architecture search," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 072–16 081.

[123] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5386–5395.

[124] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Computer Vision – ECCV 2016*,

B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 816–833.

[125] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 4263–4270.

[126] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 2969–2978.

[127] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[128] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[129] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[130] ——, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[131] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 13 359–13 368.

[132] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[133] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[134] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, "Tam: Temporal adaptive module for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 13 708–13 718.

[135] M. A. Gul, M. H. Yousaf, S. Nawaz, Z. Ur Rehman, and H. Kim, "Patient monitoring by abnormal human activity recognition based on cnn architecture," *Electronics*, vol. 9, no. 12, p. 1993, 2020.

[136] P. Woznowski., R. King., W. Harwin., and I. Craddock., "A human activity recognition framework for healthcare applications: Ontology, labelling strategies, and best practice," in *Proceedings of the International Conference on Internet of Things and Big Data - IoTBD,*, INSTICC. SciTePress, 2016, pp. 369–377.

[137] S. Sharma and A. Hoover, "Top-down detection of eating episodes by analyzing large windows of wrist motion using a convolutional neural network," *Bioengineering*, vol. 9, no. 2, p. 70, 2022.

[138] K. Okamoto and K. Yanai, "Grillcam: A real-time eating action recognition system," in *International Conference on Multimedia Modeling*. Springer, 2016, pp. 331–335.

[139] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[140] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "Bmn: Boundary-matching network for temporal action proposal generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3889–3898.

[141] C. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," *arXiv preprint arXiv:2202.07925*, 2022.

[142] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai, "End-to-end temporal action detection with transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5427–5441, 2022.

[143] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao, "Tridet: Temporal action detection with relative boundary modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 857–18 866.

[144] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[145] H. Alwassel, S. Giancola, and B. Ghanem, "Tsp: Temporally-sensitive pretraining of video encoders for localization tasks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3173–3183.

[146] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[147] F. B. Horak, "Clinical assessment of balance disorders," *Gait & posture*, vol. 6, no. 1, pp. 76–84, 1997.

[148] J. Gill, J. Allum, M. Carpenter, M. Held-Ziolkowska, A. Adkin, F. Honegger, and K. Pierchala, "Trunk sway measures of postural stability during clinical balance tests: effects of age," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 56, no. 7, pp. M438–M447, 2001.

[149] A. C. Alonso, N. M. Luna, F. N. Dionísio, D. S. Speciali, L. E. G. Leme, and J. M. D. Greve, "Functional balance assessment," *Medicalexpress*, vol. 1, pp. 298–301, 2014.

[150] A. Filippeschi, N. Schmitz, M. Miezal, G. Bleser, E. Ruffaldi, and D. Stricker, "Survey of motion tracking methods based on inertial sensors: A focus on upper limb human motion," *Sensors*, vol. 17, no. 6, p. 1257, 2017.

[151] A. Carnevale, U. G. Longo, E. Schena, C. Massaroni, D. Lo Presti, A. Berton, V. Candela, and V. Denaro, "Wearable systems for shoulder kinematics assessment: A systematic review," *BMC musculoskeletal disorders*, vol. 20, no. 1, pp. 1–24, 2019.

[152] L. Meng, M. Chen, B. Li, F. He, R. Xu, and D. Ming, "An inertial-based upper-limb motion assessment model: performance validation across various motion tasks," *IEEE Sensors Journal*, 2023.

[153] M. Amsaprabhaa *et al.*, "Multimodal spatiotemporal skeletal kinematic gait feature fusion for vision-based fall detection," *Expert Systems with Applications*, vol. 212, p. 118681, 2023.

[154] R. K. Yadav, S. G. Neogi, and V. B. Semwal, "A computational approach to identify normal and abnormal persons gait using various machine learning and deep learning classifier," in *Machine Learning, Image Processing, Network Security and Data Sciences: 4th International Conference, MIND 2022, Virtual Event, January 19–20, 2023, Proceedings, Part I*. Springer, 2023, pp. 14–26.

[155] Z. Yang, "An efficient automatic gait anomaly detection method based on semisupervised clustering," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[156] A. Barzegar Khanghah, G. Fernie, and A. Roshan Fekr, "Design and validation of vision-based exercise biofeedback for tele-rehabilitation," *Sensors*, vol. 23, no. 3, p. 1206, 2023.

[157] A. Kanade, M. Sharma, and M. Muniyandi, "Tele-evalnet: A low-cost, teleconsultation system for home based rehabilitation of stroke survivors using multiscale cnn-convlstm architecture," in *European Conference on Computer Vision*. Springer, 2023, pp. 738–750.

[158] Y. Ren, C. Lin, Q. Zhou, Z. Yingyuan, G. Wang, and A. Lu, "Effectiveness of virtual reality games in improving physical function, balance and reducing falls in balance-impaired older adults: A systematic review and meta-analysis," *Archives of Gerontology and Geriatrics*, p. 104924, 2023.

[159] A. Nalci, A. Khodamoradi, O. Balkan, F. Nahab, and H. Garudadri, "A computer vision based candidate for functional balance test," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 3504–3508.

[160] C. Yang, A. Kerr, V. Stankovic, L. Stankovic, P. Rowe, and S. Cheng, "Human upper limb motion analysis for post-stroke impairment assessment using video analytics," *IEEE Access*, vol. 4, pp. 650–659, 2016.

[161] V. M. Manghisi, A. E. Uva, M. Fiorentino, V. Bevilacqua, G. F. Trotta, and G. Monno, "Real time rula assessment using kinect v2 sensor," *Applied ergonomics*, vol. 65, pp. 481–491, 2017.

[162] L. Li, T. Martin, and X. Xu, "A novel vision-based real-time method for evaluating postural risk factors associated with musculoskeletal disorders," *Applied Ergonomics*, vol. 87, p. 103138, 2020.

[163] G. K. Nayak and E. Kim, "Development of a fully automated rula assessment system based on computer vision," *International Journal of Industrial Ergonomics*, vol. 86, p. 103218, 2021.

[164] K. A. Bartlett and J. D. Camba, "An rgb-d sensor-based instrument for sitting balance assessment," *Multimedia Tools and Applications*, pp. 1–24, 2023.

[165] M. Blomqvist, P. Luhtanen, and L. Laakso, "Validation of a notational analysis system in badminton," *Journal of Human Movement Studies*, vol. 35, no. 3, pp. 137–150, 1998.

[166] M. Oshita, T. Inao, S. Ineno, T. Mukai, and S. Kuriyama, "Development and evaluation of a self-training system for tennis shots with motion feature assessment and visualization," *The Visual Computer*, vol. 35, no. 11, pp. 1517–1529, 2019.

[167] G. Vuckovic, B. Dezman, J. Pers, and S. Kovacic, "Motion analysis of the international and national rank squash players," in *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.* IEEE, 2005, pp. 334–338.

[168] J. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (osats) for surgical residents," *British journal of surgery*, vol. 84, no. 2, pp. 273–278, 1997.

[169] Z. Li, Y. Huang, M. Cai, and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[170] K. Yordanova, S. Lüdtke, S. Whitehouse, F. Krüger, A. Paiement, M. Mirmehdi, I. Craddock, and T. Kirste, "Analysing cooking behaviour in home settings: Towards health monitoring," *Sensors*, vol. 19, no. 3, p. 646, 2019.

[171] O. Zoidi, A. Tefas, and I. Pitas, "Exploiting the svm constraints in nmf with application in eating and drinking activity recognition," in *2013 IEEE International Conference on Image Processing*, 2013, pp. 3765–3769.

[172] M. Tufano, M. Lasschuijt, A. Chauhan, E. J. Feskens, and G. Camps, "Capturing eating behavior from video analysis: A systematic review," *Nutrients*, vol. 14, no. 22, p. 4847, 2022.

[173] V. Khattar and B. Hathiram, "The clinical test for the sensory interaction of balance," *Int J Otorhinolaryngol Clin*, vol. 4, pp. 41–45, 2012.

[174] K. Berg, "Balance and its measure in the elderly: a review," *Physiotherapy Canada*, vol. 41, no. 5, pp. 240–246, 1989.

[175] I. H. Rosenberg, "Sarcopenia: origins and clinical relevance," *The Journal of nutrition*, vol. 127, no. 5, pp. 990S–991S, 1997.

[176] Y. Rolland, S. Czerwinski, G. A. Van Kan, J. Morley, M. Cesari, G. Onder, J. Woo, R. Baumgartner, F. Pillard, Y. Boirie *et al.*, "Sarcopenia: its assessment, etiology, pathogenesis, consequences and future perspectives," *The Journal of Nutrition Health and Aging*, vol. 12, pp. 433–450, 2008.

[177] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[178] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[179] Z. Ghahramani and M. Jordan, "Supervised learning from incomplete data via an em approach," in *Advances in Neural Information Processing Systems*, J. Cowan, G. Tesauro, and J. Alspector, Eds., vol. 6. Morgan-Kaufmann, 1993.

[180] A. Fabisch, "gmr: Gaussian mixture regression," *Journal of Open Source Software*, vol. 6, no. 62, p. 3054, 2021. [Online]. Available: https://doi.org/10.21105/joss.03054

[181] F. Stulp and O. Sigaud, "Many regression algorithms, one unified model: A review," *Neural Networks*, vol. 69, pp. 60–79, 2015.

[182] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981. [Online]. Available: /brokenurl#http://publication.wilsonwong.me/load.php?id=233282275

[183] B. Debnath, M. O'brien, M. Yamaguchi, and A. Behera, "A review of computer vision-based approaches for physical rehabilitation and assessment," *Multimedia Systems*, vol. 28, no. 1, pp. 209–239, 2022.

[184] R. Mehrizi, X. Peng, S. Zhang, R. Liao, and K. Li, "Automatic health problem detection from gait videos using deep neural networks," *arXiv preprint arXiv:1906.01480*, 2019.

[185] S. Sardari, S. Sharifzadeh, A. Daneshkhah, B. Nakisa, S. W. Loke, V. Palade, and M. J. Duncan, "Artificial intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review," *Computers in Biology and Medicine*, p. 106835, 2023.

[186] M. J. Raihan, M. A. R. Ahad, and A.-A. Nahid, "Automated rehabilitation exercise assessment by genetic algorithm-optimized cnn," in *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2021, pp. 1–6.

[187] E. Mottaghi and M.-R. Akbarzadeh-T, "Automatic evaluation of motor rehabilitation exercises based on deep mixture density neural networks," *Journal of Biomedical Informatics*, vol. 130, p. 104077, 2022.

[188] S. Deb, M. F. Islam, S. Rahman, and S. Rahman, "Graph convolutional networks for assessment of physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 410–419, 2022.

[189] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba, "Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance," *IEEE journal of biomedical and health informatics*, vol. 24, no. 1, pp. 280–291, 2019.

[190] S. Bi and D. Kotz, "Eating detection with a head-mounted video camera," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2022, pp. 60–66.

[191] K. Okamoto and K. Yanai, "Grillcam: A real-time eating action recognition system," in *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*. Springer, 2016, pp. 331–335.

[192] O. Zoidi, A. Tefas, and I. Pitas, "Exploiting the svm constraints in nmf with application in eating and drinking activity recognition," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 3765–3769.

[193] M. Tufano, M. Lasschuijt, A. Chauhan, E. J. Feskens, and G. Camps, "Capturing eating behavior from video analysis: A systematic review," *Nutrients*, vol. 14, no. 22, p. 4847, 2022.

[194] M. P. Lasschuijt, E. Brouwer-Brolsma, M. Mars, E. Siebelink, E. Feskens, K. de Graaf, and G. Camps, "Concept development and use of an automated food intake and eating behavior assessment method," *JoVE (Journal of Visualized Experiments)*, no. 168, p. e62144, 2021.

[195] Z. Luo, J.-T. Hsieh, N. Balachandar, S. Yeung, G. Pusiol, J. Luxenberg, G. Li, L.-J. Li, N. L. Downing, A. Milstein *et al.*, "Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring," *Machine Learning for Healthcare (MLHC)*, vol. 2, no. 1, 2018.

[196] X. Huang, J. Wicaksana, S. Li, and K.-T. Cheng, "Automated vision-based wellness analysis for elderly care centers," in *Multimodal AI in healthcare: A paradigm shift in health intelligence*. Springer, 2022, pp. 321–333.

[197] H. Alwassel, S. Giancola, and B. Ghanem, "Tsp: Temporally-sensitive pretraining of video encoders for localization tasks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3173–3183.

[198] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5386–5395.

[199] R. Kunde, “What Are Activities of Daily Living (ADLs)?” https://www.webmd.com/a-to-z-guides/what-are-activities-of-daily-living, 2023, [Online; accessed 18-July-2024].