A Vision-Based System for Monitoring Eating Behaviors and Musculoskeletal Function

Muhammad Ahmed Raza[1*] and Robert B. Fisher[1]

[1*]School of Informatics, The University of Edinburgh
Edinburgh, EH8 9AB, State, United Kingdom.

*Corresponding author(s). E-mail(s): m.a.raza@ed.ac.uk;
Contributing authors: rbf@inf.ed.ac.uk;

Abstract

Camera-based systems offer a comprehensive and inconspicuous approach to monitoring the well-being of individuals within the comfort of their homes. This study introduces a vision-based, fully autonomous pipeline for assessing eating behaviors and detecting musculoskeletal changes. The system captures eating activities and provides detailed insights such as hand-to-mouth motion duration and bite count. These indicators are vital for understanding behavioral and physiological influences on food consumption and their associated changes. The system integrates pose estimation and a temporal action localization network to classify actions and generate behavior profiles. Evaluated on the EatSense dataset and a supplementary test set, the system achieves strong performance, including a mean average precision (mAP) of 74% at 0.10 IoU for micro-action detection and a posture anomaly detection accuracy of over 76%. These results demonstrate the system's ability to detect subtle trends such as slower hand movements under increased wrist weights and changes in chewing behavior. Additionally, comparisons against Gemini-2.5-Pro, a state-of-the-art multimodal model, reinforces the system's accuracy. So, by successfully capturing trends aligned with ground truth data, the pipeline shows promise for long-term health monitoring, early detection of musculoskeletal decline, and behavioral changes in dietary habits—offering potential applications in elderly care and remote health assessment. The new test dataset is released on https://groups.inf.ed.ac.uk/vision/DATASETS/EATSENSE/.

Keywords: EatSense, Change in movement detection, Eating Behaviour monitoring, sVideo to report Generation

1 Introduction

Understanding eating behaviors and their physiological foundation is vital for promoting health and detecting early signs of disordered eating or physical decline. Eating is a routine activity that offers a wealth of observable patterns, including the number of bites, chewing duration, and hand-to-mouth motions. These metrics can reveal insights into dietary habits, musculoskeletal function, and broader health outcomes.

Digital home health monitoring systems can be broadly classified into two categories, 1) vision-based, and 2) wearable sensor-based. While wearable sensors or multi-sensor systems are effective for identifying acute conditions, their acceptance is limited due to their intrusive nature and the possibility of users forgetting to wear or recharge them. On the other hand, camera/vision-based systems can be less intrusive and help detect important situations and trends.

Vision-based monitoring systems for behavioral health informatics can potentially identify and monitor minor signs, thus enabling earlier identification and intervention. Clinical systems are costly and require a human operator in the loop; consequently, they risk human error due to misinterpretation or inattention. Therefore, automated vision-based systems can potentially be valuable aids in physical rehabilitation or the evaluation of conditions like stroke and Parkinson's disease (PD) [1]. In many of these applications, pose estimation plays a crucial role. It has been widely used in gait analysis to identify deviations linked to neurological or musculoskeletal conditions, and in person identification by capturing unique patterns of movement [2], [3]. By estimating joint positions and tracking their trajectories, pose estimation enables non-contact assessment of functional mobility and motor health, which is essential in rehabilitation and continuous home monitoring scenarios [4].

Existing research on vision-based monitoring has predominantly focused on isolated aspects of eating, such as bite detection or chewing analysis, or general activity monitoring in contexts like fall detection or rehabilitation. However, there is a gap in integrating these capabilities into a holistic framework capable of analyzing eating behaviors alongside musculoskeletal function. This integration is particularly relevant in aging populations, where changes in upper-body motion and eating habits may signal underlying health issues. In general, the hope is that trends and abnormalities can be found by analyzing the visual data, enabling healthcare professionals to make more informed decisions.

This paper introduces a vision-based fully autonomous framework with three stages: 1) To promote healthy eating habits, it first analyzes eating behaviors, such as chewing duration and mouthful count. 2) Secondly, to identify possible decreases in muscle activity, it tracks the speed of arm movement. 3) Thirdly, a new classification technique is used to detect eating posture anomalies.

The contributions of this paper are:

1. A multi-purpose, fully autonomous, video-to-report (V2R) pipeline for long-term eating behavior and muscle deterioration monitoring (Section 3.2).
2. The paper introduces a relaxed data augmentation (pre-processing) step, autoencoder-style learnable temporal position encodings (TPE), and a temporal segment soft merge and suppress criteria (post-processing) step for the temporal

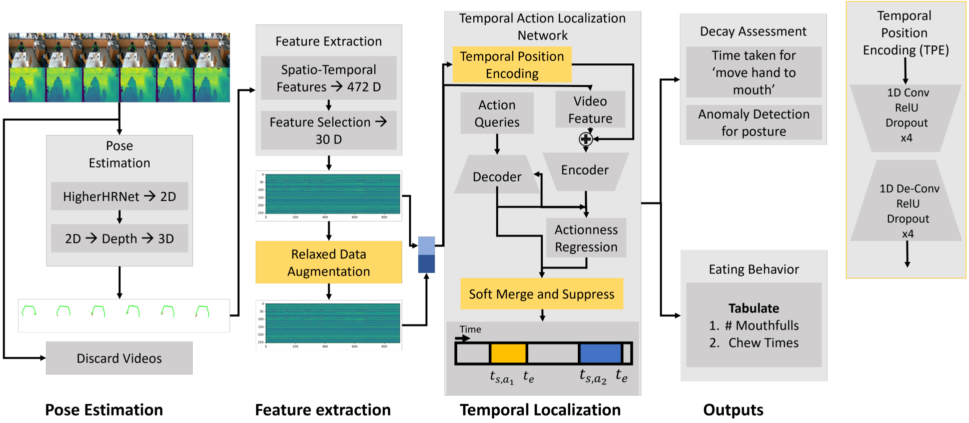

Fig. 1: Block diagram of the proposed system. The proposed pipeline consists of three steps after the video is collected, 1) it finds the poses from the video and estimates features using those poses, 2) uses these estimated features as an input to the temporal action localization to get temporal segments and action classes, and 3) derive insights into the eating behavior and muscular movement of the individual. The yellow blocks highlight the technological contributions to achieve improved temporal action localization on EatSense with a general transformer-based TAL framework.

action localization (TAL) network that helps to more accurately localize the actions in the continuous video (Section 4.1).

3. 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 4.3).
4. A small extension is made to the EatSense dataset [5], where three individual's long-term changes are simulated by adding weights (0, 1kg, and 2.4kg) to the wrists of the subjects (Section 3.1).

2 Literature Review

This research mainly focuses on camera/vision-based sensors. Most vision-based health monitoring systems focus solely on fall detection and its prevention. Although it is important to monitor for falls, it is equally critical to monitor the behavior of individuals for long-term changes. The vision-based health monitoring systems research is categorized into two groups: firstly, research focusing on health-related activities such as classifying or understanding activities of daily living (ADLs) or eating characteristics; and 2) fully autonomous monitoring systems that utilize a video as an input and output a meaningful summary of results useful to a health-care worker.

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

2.1 Vision-based Health-Related Research

Vision-based health-related research in general considers various aspects of an individual's health including vision-based gait analysis for neurological disease detection [6], etc. However, the research presented here focuses on analyzing an individual's upper body motions; hence, the scope of this literature review is limited to only upper-body or full-body related health research. This discussion is further divided into two categories, activities of daily living (ADL) monitoring and eating behavior monitoring. ADL monitoring involves observing an individual throughout the day and drawing insights on the data for various applications including rehabilitation or action quality assessment, etc. The latter on the other hand, focuses more strictly on eating behavior, but can also be useful for further analysis, such as movement decay assessment or eating disorder detection, etc.

2.1.1 ADL Monitoring

To offer personalized services or treatments, it is crucial to have a comprehensive understanding of the daily activities of an individual. For instance, accurately detecting ADL can yield numerous advantages, such as analyzing human lifestyle, monitoring diet, facilitating active rehabilitation, and more.

Much research has been done on ADL monitoring for the physical rehabilitation of individuals in the past decade [7],[1]. In [8] and [9] the authors used a publicly available dataset captured with a Kinect V2 sensor to automatically assess the physical ADLs for individuals suffering from Parkinson's or who have had a stroke. Deb et al. [10] combined two publicly available datasets for automatically assessing ADLs with attention modules in the deep network for better explainability of the model.

Elkholy et al. [11] monitored a subset of ADLs (sitting, standing up, walking, etc) and designed a multi-head deep network for the classification of normal/abnormal and assessing the efficiency of the action performed.

2.1.2 Eating Behavior Monitoring

Eating behaviors can generally be classified as how the person is eating, such as mindfully or rapidly, based on feeding motions, bites, chews, and swallows, etc, and what actions they commonly perform while they eat or drink such as mixing sugar into their tea. These eating behavior monitoring systems can be divided into four categories based on their underlying application [12]: 1) Eating/drinking activity recognition, 2) bite/chews/swallows detection, 3) portion-size estimation, and 4) sensor location, i.e., placement of a specific sensor on the body part or the object under observation. However, this literature review is limited to vision-based applications that monitor eating behaviors.

Bi et. al. [13] developed a head-mounted camera system to detect eating and non-eating activity. Nour et. al. [14] proposed an eating detection algorithm using pose-based action recognition for elderly people with dementia. Similarly, most vision-based eating monitoring systems, such as [15], [16], [17], and [18] focus on recognizing eating/drinking actions rather than developing deeper behavioral insights through the data.

In [19] Lasschujit et al. proposed a tray equipped with a camera that monitored bites and chews alongside weight sensors to monitor the instantaneous amount the food eaten by the individual. In [20], [21], [22] the researchers focused on developing pipelines for automatic bite detection and their counts for a full meal. Similarly, in [23] [24] the focus was detecting and counting the chewing activity.

Tufano et al. [25] carried out an extensive review of thirteen video-based techniques with outcomes including intake gesture detection, meal duration, bite counts, and the number of chews, etc. They also highlight the lack of research for the understanding of eating behaviors in uncontrolled environments. Recently, Raza et al. [26] utilized eating videos to assess performance levels and presented a general eating behavior state diagram. They also proposed an uncertainty-aware algorithm to obtain a generalized model to regress performance changes across multiple subjects since the indicative features of decay vary significantly across people with different lifestyles.

To summarize, detecting bites/chews and eating actions are important aspects of eating behavior monitoring. For this purpose, numerous vision-based pipelines for monitoring eating behavior exist, though most of them target only specific aspects of eating behavior assessment. Typically, these systems focus on either detecting eating or drinking actions or on estimating the number of bites and the duration of chewing. Our proposed automated pipeline, however, not only tracks the number of mouthfuls (bites) and chewing duration but also evaluates how the individual's performance changes over time while eating. To the best of our knowledge, no existing pipeline provides insights into both eating behavior and performance changes following video-based temporal action localization.

2.2 Fully Autonomous Monitoring Systems

Recent years have seen increased interest in fully autonomous systems for elderly monitoring. Luo et. al. [27] proposed a fully autonomous system that used two modalities (an infrared and a depth sensor) to monitor elderly individuals, generating time logs of their daily activities. The system utilized a frame-by-frame temporal activity detection algorithm coupled with smoothing windowed filtering. Although the framewise temporal action localization helps in understanding various aspects of an elderly individual's routine, solely logging activities does not provide valuable insights into the patterns or anomalies in the lifestyle of the individual.

Recently, Huang et. al. [28] proposed a similar system that comprised frame-wise temporal action localization followed by facial analysis, activity detection, and subjects' interaction with the environment. This system performs meaningful analysis of the videos for a deeper understanding of an individual's long-term behavior.

Recent advancements in large multimodal models (LMMs), such as Gemini-2.5-Pro, have shown promise in understanding complex visual tasks when prompted with natural language instructions. While these models are primarily designed for general-purpose reasoning, their ability to process unstructured visual data and extract temporal patterns makes them a valuable tool for cross-validating domain-specific pipelines. In this work, we used Gemini-2.5-Pro as a benchmarking tool for post-hoc comparison of eating behavior statistics, demonstrating that such LMMs can serve as flexible validators for temporal and behavioral data extraction.

In summary, past research provides valuable insights into the activities of daily life and behaviors of individuals using full-body or gait motion analysis pipelines. However, there's a gap in research focusing specifically on fully autonomous vision-based systems to analyze health statistics, concentrating strictly on upper body motions with holistic evaluation. The research presented here introduces a new fully autonomous system aimed at monitoring and analyzing eating behavior and musculoskeletal degradation. This is an important contribution because eating is typically undertaken regularly and in a standard location and thus is amenable to observation by a fixed camera.

3 Methodology

3.1 EatSense and Test Set

For this research, the EatSense [5] dataset was used[1]. EatSense was collected in dining environments equipped with an RGB-D Intel RealSense D415 camera, where both RGB and depth information is recorded. RealSense uses infrared (IR) projection that helps it capture depth data accurately in low-light conditions, making it somewhat invariant to lighting conditions. The camera was directed toward the dining table, ensuring that each frame captured only a single subject's frontal view from an oblique angle. Recordings were conducted at various locations, featuring diverse camera-to-subject distances (depending on the location of the dining table from the wall), different backgrounds, and no control over the subject movements. The dataset contains 135 videos of 27 healthy subjects (with faces obfuscated to protect their identity [29]) from different ethnicities and age groups (varying from below 30 to over 60), and with different eating styles. [29].

EatSense contains both gesture-based ('chewing', etc.) and velocity-based ('move hand towards mouth', etc) micro actions while a person eats - 16 sub-action classes in total. The most frequent actions are "eat it" (2,630 instances), "move hand towards mouth" (2,851 instances), and "move hand away from mouth" (2,792 instances), reflecting the core focus on hand-to-mouth activity. Other common actions include "pick food from dish with tool in one hand" (1,548 instances), "chewing" (795 instances), and "other" actions (2,057 instances), capturing additional context. Less frequent but still relevant actions include "drink" (247 instances), "pick food with one hand" (440 instances), and "pick food with both hands" (282 instances), along with actions involving utensils and cups, such as "pick up a cup/glass" (213 instances), "put the cup/glass back" (214 instances), and "put one tool back" (253 instances). Rare classes include "no action" (64 instances) and "pick up tools with both hands" (65 instances). Overall, the dataset offers rich coverage of micro-actions that are crucial for understanding fine-grained eating behaviors.

However, the research presented in this paper only utilizes two micro-actions ('move hand towards mouth' and 'move hand away from mouth') both of which last less than one second. As the videos are recorded at 15 fps, that means, on average, 'move hand towards mouth' and 'move hand away from mouth' span over 12.7 and 9.4 frames respectively. There are collectively 5643 instances of these actions in EatSense. On the

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

other hand, it also simulates musculoskeletal deterioration by tying weights of different magnitudes to the wrists of the subject. These characteristics make EatSense a perfect choice for training the proposed autonomous system.

EatSense is an imbalanced dataset with more videos for some subjects and fewer for others. To avoid introducing any bias due to different behavioral characteristics across subjects, for this study the dataset was balanced by using 4 videos each from 24 subjects (there are 27 subjects in total, but three have only two videos), hence 96 videos in total. Then, the dataset is divided into five parts by splitting the dataset into four groups of five subjects each and one group of 4 subjects.

We captured a new supplementary test set strictly for the holistic evaluation of the pipeline with characteristics and settings similar to those of the EatSense dataset. However, in this case, three videos were recorded for each of the three subjects. In these recordings, two of the subjects were instructed to eat from the same bowls of three different sizes — while wearing weights of 0kg, 1kg, and 2.4kg on each wrist, respectively, in a distraction-free environment, i.e., no chatting and no phone. This is usually the case for elderly individuals. Moreover, to test how the proposed pipeline holds when the assumptions are not met, the third subject was purposely requested to chat continuously whenever they could chat and use a mobile phone while they ate, hence a distraction-filled environment. This setup aimed to simulate changes in eating habits resulting from musculoskeletal deterioration, such as a reduced number of mouthfuls or slower arm movements with increased decay (heavier weights). The new test dataset is released on https://groups.inf.ed.ac.uk/vision/DATASETS/EATSENSE/.

3.2 Proposed System

The recorded videos consist of untrimmed, full-length footage of subjects eating. To extract information such as the name of each action and their start and end times in the video, a temporal action localization (TAL) framework was employed. Refer to Figure 1 for an illustration of the process, with both an overview and more details below.

Many TAL networks heavily rely on separately estimating video encodings (i.e. frame-wise activity and context descriptions) as a step in the pipeline, as end-to-end training demands substantial computational power. Typically, these video embeddings are estimated using deep learning networks like I3D [30] and TSP [31]. However, in this research, we leveraged features engineered using domain knowledge to make the model more intuitive and understandable for healthcare experts.

After obtaining temporal segment information, the output activity segments are analyzed to count the number of mouthfuls and estimate chewing duration. Lastly, to monitor musculoskeletal decay, the time taken for the ‘move hand towards mouth’ action was tracked. The whole pipeline of the proposed system is shown in the block diagram in Figure 1. The details of each of these steps are discussed below.

3.2.1 Pose Estimation and Feature Extraction

Initially, the 2D poses of the person eating are estimated using HigherHRNet [32], as it is identified as the most accurate pose estimation algorithm for the EatSense dataset, as reported in [5]. As an RGB-D camera was employed for recording, the 2D points of each of the eight visible joints are projected into 3D space utilizing basic computer vision techniques along with depth information from the RGB-D camera. This gives 3D coordinates for each joint. Given that only the upper body of the subject is visible, eight joints ($\vec{j}_i^{[t]}, i = 1 : 8$) were selected for analysis: head, chest, left shoulder, right shoulder, left elbow, right elbow, left wrist, and right wrist.

Following the footsteps of the EatSense dataset [5] conventions and mathematical details for feature estimation, the 24-dimensional vector (8×3) containing the absolute location of eight 3D joints is used to estimate various spatial and temporal features.

These include instantaneous positions of the joints relative to the chest ($\vec{r}_i^{[t]} = \vec{j}_i^{[t]} - \vec{j}_2^{[t]}$), the instantaneous distance between the chest and the table ($\vec{c}^{[t]}$), past three lags, velocity ($\vec{v}_i^{[t]}$), acceleration ($\vec{a}_i^{[t]}$), etc. Lastly, a forward sequential feature selection (FSFS) algorithm was used to identify the most contributing features (out of $\{\vec{j}_i^{[t]}, \vec{r}_i^{[t]}, \vec{c}_i^{[t]}, \vec{v}_i^{[t]}, \vec{a}_i^{[t]}, \dots\}$, , i = 1:8) for the frame-wise classification of the 'background action' versus two micro-actions ('move hand towards mouth' and 'move hand away from mouth'). Using FSFS, the top 30 features were selected (illustrated in Fig. 2, the curve starts to get flat beyond 30 features), forming a 30-D video feature embedding ($\vec{f}^{[t]}$). This embedding serves as the input to the TAL networks.

3.2.2 Temporal Action Localization Network

The model was trained to use the feature vectors ($\vec{f}^{[t]}$) localize the distinction between 'move hand towards mouth' and 'move hand away from mouth' (i.e. find the times t_s for the start and end t_e of instances of the two action primitives). Given that these actions last less than a second (less than 15 frames), and to minimize any discrepancies introduced due to human labeling error, the ground truth action instances were temporally extended or cropped. Firstly, the exact hand-labelled boundaries were used. Secondly, we redefined the start t_s^g and end t_e^g of the temporal segment with a relaxed boundary threshold $\epsilon \in \{-2 \times \frac{1}{15}, -1 \times \frac{1}{15}, 0 \times \frac{1}{15}, 1 \times \frac{1}{15}, 2 \times \frac{1}{15}\}$. This represents ±2 frames (chosen randomly), multiplied by 1/15 (because videos are recorded at 15 fps), and added to t_s^g and t_e^g. This relaxed data augmentation process is illustrated in Figure 3.

We introduced temporal positional embedding (TPE) using an Autoencoder-style architecture that processes temporal data using 1D convolutions, capturing information over sequential time steps. This can be represented mathematically as follows.

Let the input be a sequence of feature vectors extracted from a video, representing either hand-crafted or learned features over time:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T], \quad \mathbf{x}_t \in \mathbb{R}^d \tag{1}$$

where T is the number of time steps (frames), and d is the dimensionality of each feature vector. To inject temporal information, we add positional encodings to each frame. These can be either fixed (e.g., sinusoidal as in transformers) or, as in our case,

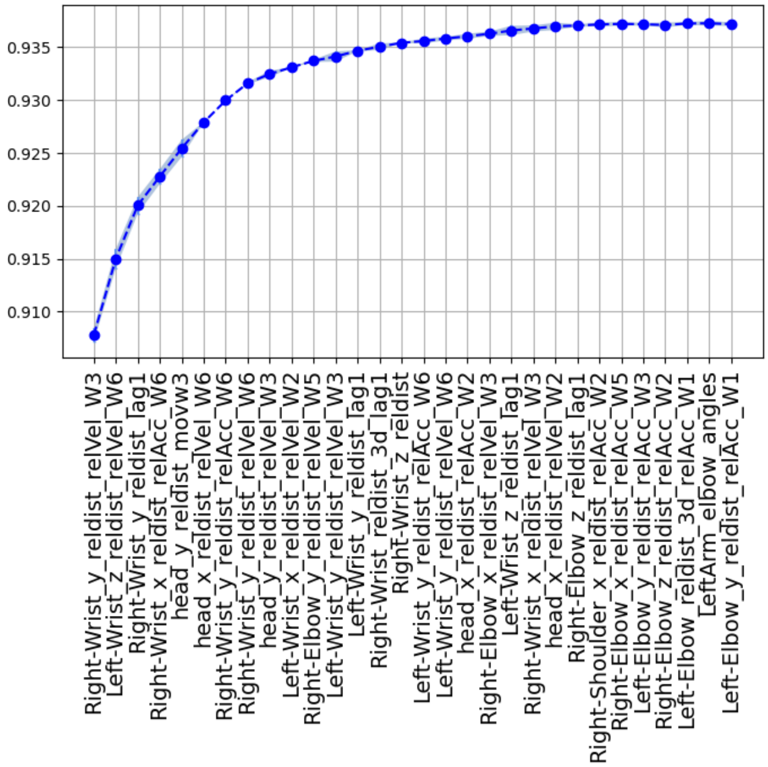

Fig. 2: Top 30 most contributing features selected using forward sequential feature selection. The vertical axis shows accuracy achieved on the frame-wise classification of 'move hand towards mouth' and 'move hand away from mouth'.

learned positional embeddings denoted as:

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T], \quad \mathbf{p}_t \in \mathbb{R}^d \tag{2}$$

The positionally encoded input becomes:

$$\mathbf{Z} = \mathbf{X} + \mathbf{P} \tag{3}$$

where positional encodings $\mathbf{p}_t$ are learned as model parameters. The encoded sequence $\mathbf{Z} \in \mathbb{R}^{T \times d}$ is passed through a stack of 1D convolutional layers to learn temporal dependencies. Each layer applies a 1D convolution with increasing dilation to capture longer temporal context:

$$\mathbf{h}^{(i)} = \text{ReLU}\left(\text{Conv1D}(\mathbf{h}^{(i-1)}; r_i, k)\right) \tag{4}$$

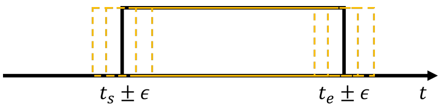

Fig. 3: Relaxed Data Augmentation. t_s^g and t_e^g refer to the start and end of action in the ground truth whereas ϵ denotes the ± 2 frames i.e, $\epsilon \in \{-2 \times \frac{1}{15}, -1 \times \frac{1}{15}, 0 \times \frac{1}{15}, 1 \times \frac{1}{15}, 2 \times \frac{1}{15}\}$ relaxation for augmentation.

where $\mathbf{h}^{(0)} = \mathbf{Z}$, r_i is the dilation rate at layer i (e.g., $r_i = 2^i$) and k is the kernel size. After L such layers, the temporally encoded representation is:

$$\mathbf{H} = \mathbf{h}^{(L)} \in \mathbb{R}^{T \times d'} \tag{5}$$

A decoder reconstructs the input using transposed convolutions:

$$\hat{\mathbf{Z}} = \text{Decoder}(\mathbf{H}) \tag{6}$$

The TPE is trained using a reconstruction loss, i.e., the mean squared error (MSE) as an auxilliary loss fucnction alongside the main loss function as shown in equation 8:

$$\mathcal{L}_{\text{rec}} = \left\| \hat{\mathbf{Z}} - \mathbf{Z} \right\|_2^2 \tag{7}$$

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{TALN}} + \lambda \mathcal{L}_{\text{rec}} \tag{8}$$

$\mathcal{L}_{\text{TALN}}$ is defined in eq. 9 where λ are weights for balancing each term and $(\hat{c}_i, \hat{t}_{s,i}, \hat{t}_{e,i}, \hat{s}_i), \quad i = 1, \ldots, N$ are the the predicted class, start time, end time, and confidence for the i^{th} query.

$$\begin{aligned}
\mathcal{L}_{\text{TALN}} = \frac{1}{N} \sum_{i=1}^{N} \Big[& \lambda_{\text{cls}} \cdot \text{CrossEntropy}(\hat{c}_i, c_{\sigma(i)}) \\
& + \lambda_{\text{reg}} \cdot \left(\left| \hat{t}_{s,i} - t_{s,\sigma(i)} \right| + \left| \hat{t}_{e,i} - t_{e,\sigma(i)} \right| \right) \\
& + \lambda_{\text{iou}} \cdot \left(1 - \text{IoU} \left([\hat{t}_{s,i}, \hat{t}_{e,i}], [t_{s,\sigma(i)}, t_{e,\sigma(i)}] \right) \right) \\
& + \lambda_{\text{act}} \cdot \text{SmoothL1} \left(\hat{s}_i,\ \text{IoU} \left([\hat{t}_{s,i}, \hat{t}_{e,i}], [t_{s,\sigma(i)}, t_{e,\sigma(i)}] \right) \right) \Big]
\end{aligned} \tag{9}$$

Overall, this architecture allows the model to learn temporally contextualized features that include positional information, making it well-suited for tasks like action localization, segmentation, and anomaly detection in sequential video data.

Additionally, the model uses a mask to ignore irrelevant or padded portions of the data. This masking is applied during the final stage, ensuring that only meaningful parts of the sequence contribute to the output. The architecture is designed to learn positional information from temporal data efficiently.

The outputs of the TAL network consist of temporal segments (t_s^i and t_e^i), action class labels, and a confidence score. Since transformers work on a fixed number of queries, the number of output detections of each of the test video sequences is fixed.

TadTR inherently uses an action matching module and does not require NMS while training, however, it still produces redundant predictions during inference which are suppressed, and merged using NMS. However, instead of suppressing predictions below a certain threshold, we define a soft merge and suppress criteria. Let the output of the Temporal Action Localization (TAL) transformer consist of a fixed number of predictions:

$$\mathcal{D} = \left\{ \left(t_s^i, t_e^i, c^i, s^i \right) \right\}_{i=1}^{N} \tag{10}$$

where $t_s^i, t_e^i \in \mathbb{R}$ are the predicted start and end times for segment i, $c^i \in \{1, 2, \ldots, C\}$ is the predicted class label and $s^i \in [0, 1]$ is the *actionness score*, i.e., the confidence score predicted by the actionness regression head.

We define a filtered detection set $\mathcal{D}' \subseteq \mathcal{D}$ based on the following conditions:

$$s^i > \tau_s \quad \text{where } \tau_s = 0.3 \tag{11}$$

$$t_e^i - t_s^i > \delta_t \quad \text{where } \delta_t = 0.1\,\text{seconds} \tag{12}$$

Next, we define a soft merge mechanism for overlapping predictions. Let $d_i = (t_s^i, t_e^i, c^i, s^i)$ and $d_j = (t_s^j, t_e^j, c^j, s^j)$ be two detections from $\mathcal{D}'$ with the same class label $c^i = c^j$. Their temporal overlap is given by:

$$\text{Overlap}(i, j) = \max\left(0, \min(t_e^i, t_e^j) - \max(t_s^i, t_s^j)\right) \tag{13}$$

The detections d_i and d_j are merged if:

$$\text{Overlap}(i, j) > \delta_o \quad \text{where } \delta_o = \frac{2}{15}\,\text{seconds} \tag{14}$$

The resulting merged segment $d_m = (t_s^m, t_e^m, c^m, s^m)$ is computed as:

$$t_s^m = \min(t_s^i, t_s^j) \tag{15}$$

$$t_e^m = \max(t_e^i, t_e^j) \tag{16}$$

$$c^m = c^i = c^j \tag{17}$$

$$s^m = \max(s^i, s^j) \tag{18}$$

All predictions not satisfying the score, duration, or overlap conditions are discarded. This approach avoids hard non-maximum suppression (NMS) and allows temporally close predictions with sufficient confidence to be softly merged, improving robustness for short-duration actions.

3.2.3 Anomaly Detection

Anomaly detection (AD), also known as one-class classification, identifies patterns that deviate significantly from the norm. It's an essential method for finding outliers or odd behavior in data analysis. For detecting anomalies in the instantaneous posture of a person during the 'move hand towards mouth' micro-action, an anomaly detection

(AD) / one-class classifier was used. To emphasize the posture, we utilize 30 postural features intuitively chosen, including 21 postural features (derived from the positions of eight 3D upper-body joints relative to the chest), the instantaneous distance between wrists, and the distance of each of the 8 joints from the table. The chest serves as the origin and is therefore excluded from the features.

Let the training dataset consist of posture features extracted from videos with no wrist weights (i.e., normal data):

$$\mathcal{X}_{\text{normal}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \tag{19}$$

We train a one-class Support Vector Machine (SVM) on $\mathcal{X}_{\text{normal}}$ to learn a boundary that captures the region of high-density (normal) data in the feature space. The one-class SVM solves the following optimization problem:

$$\min_{\mathbf{w},\rho,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu N}\sum_{i=1}^{N} \xi_i - \rho \tag{20}$$

subject to:

$$(\mathbf{w} \cdot \phi(\mathbf{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, N \tag{21}$$

where $\phi(\cdot)$ is a nonlinear feature mapping to a high-dimensional space, $\nu \in (0, 1]$ controls the trade-off between the fraction of outliers and the decision boundary tightness, ρ is the offset and ξ_i are slack variables allowing violations of the margin for soft decision boundaries. The decision function is defined as:

$$f(\mathbf{x}) = \text{sign}((\mathbf{w} \cdot \phi(\mathbf{x})) - \rho) \tag{22}$$

During inference, for a new feature vector $\mathbf{x}'$, the model computes an anomaly score:

$$\text{score}(\mathbf{x}') = (\mathbf{w} \cdot \phi(\mathbf{x}')) - \rho \tag{23}$$

A negative score implies that the posture is anomalous, while a positive score indicates it is similar to the normal data distribution.

In this work, the model is trained on postural data collected without wrist weights and evaluated on data collected with 2.4 kg wrist weights, under the hypothesis that the added weight induces posture changes that deviate from the norm. The classification performance of the anomaly detection system is reported in terms of standard metrics such as accuracy and F-score.

4 Experiments

The proposed pipeline was evaluated in multiple ways. Firstly, we train and evaluate the temporal localization network (see section 4.1). Secondly, we assess if the dataset under evaluation itself has any behavioral trend to explore (see section 4.2). Thirdly, we analyze the pipeline as a whole with videos as input and three different statistics (characterizing eating behavior and musculoskeletal changes - (see section 4.3)). Please note, the first two experiments are on individual blocks of the pipeline and utilized the test set presented in the EatSense dataset for its evaluation. However, the extension of EatSense was only used for validation in the third experiment, i.e., in the holistic analysis of the pipeline.

In addition to evaluating individual components and the overall system performance, Gemini 2.5 Pro was used as an external comparative model, given the methodological heterogeneity among prior eating behavior analysis pipelines. This state-of-the-art multimodal foundation model was guided using prompt-based instructions that described the temporal structure of eating actions. Its output, comprising per-instance action intervals and summary statistics, was used to compare against and validate the predictions generated by the proposed pipeline. The exact text prompt used for all Gemini experiments is provided in A.

4.1 Temporal Action Localization Network (TALN) Tests

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

The experiments on the TAL network were divided into two sub-experiments, TAL using 1) deep learning-based video encodings (TSP [31]) and 2) hand-crafted (HC) video encodings. For both sub-experiments, 5-fold cross-validation was used, with the data splits as described in the previous paragraph. Mean average precision (mAP) is used as a metric for the performance evaluation. 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 temporal intersection over union (tIoU) thresholds. tIoU measures the overlap between the predicted action segment and the ground truth segment in terms of their temporal boundaries. 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.

Table 1 presents the mean and standard deviation of the achieved mean average precision (mAP) at 10%, 30%, and 50% temporal intersection over union (tIoU). The last column displays the average mAP across all thresholds between 0 and 0.95 (incremented by 0.05 at each step). The results indicate that both hand-crafted and deep learning-based video encodings achieve very similar performance at identifying action instances and segmenting them from the continuous video. The high standard deviation represents high variability in the dataset due to the difference in parameters that make up each of the individual's motion profiles. Hence, splitting them subject-wise for 5-fold CV causes a domain shift [26].

The baseline architecture (TadTR + TSP) achieves competitive results, with mAP@0.10 and mAP@0.30 scores of 69.4 and 61.4, respectively. When paired with HC features instead of TSP, TadTR shows an improvement at mAP@0.50, reaching 34.1 compared to 22.9 with TSP. However, the best performance comes from the proposed model, TALN + TSP + Mods and TALN + HC + Mods. These models significantly outperform the baseline, with mAP@0.10 scores of 71.7 and 73.8, and average scores across IoU thresholds (0.05–0.95) of 41.8 and 41.2, respectively.

Table 1: The mean (μ) and standard deviation (σ) of the 5-fold CV results from the two TAL networks at various intersection over union thresholds. HC represents hand-crafted features, TSP shows the results on deep video-encoded features and Mods stand for the proposed modifications in the baseline architecture.

Architecture	**mAP@0.10**		**mAP@0.30**		**mAP@0.50**		**0:0.05:0.95**	
	μ	σ	μ	σ	μ	σ	μ	σ
Tridet + HC	52.1	3.1	42.0	1.9	22.4	4.13	27.1	1.18
TadTR + TSP (baseline)	69.4	12.3	61.4	13.0	42.9	14.1	40.8	8.9
TadTR + HC	63.1	4.2	53.7	6.5	30.3	6.2	34.1	3.8
TALN + TSP + Mods (Ours)	71.7	8.5	64.1	9.8	**49.4**	12.9	41.2	7.2
TALN + HC + Mods (Ours)	**73.8**	10.5	**65.6**	13.7	43.2	16.6	**41.8**	8.9

The modifications introduced in the proposed models (TALN + Mods) demonstrate notable improvements in precision and robustness, as indicated by both the mean and standard deviation across all metrics. This also highlighting the impact of integrating domain specific hand-crafted features with modifications.

For the remaining experiments, such as the end-to-end evaluation, the TAL model trained with hand-crafted features was used.

It's worth noting that mAP estimation relies on an intersection over union (IoU) threshold, which is highly sensitive when applied to very short actions. In particular, one or two frames on either side can significantly impact the mAP estimation at any threshold. Therefore, we chose to compare mean average precision (mAP) at lower temporal intersection over union (IoU) thresholds, given that the actions ('move hand towards mouth' and 'move hand away from mouth') last less than a second at 15 frames per second (fps). In any case, the mAP@0.10 score means that about 74% of all micro-actions were detected with a temporal overlap with the ground truth of 10%, i.e. 1 frame.

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 provides an effective test bed for musculoskeletal change detection as it simulates a change in upper-body movement (by attaching weights to the wrists of the subjects) which was demonstrated by balance assessment speed tests explored by fitting linear models in [26]. Questions such as 'Is there any observable change in performance as a function of the four weights?' were masked by the linear model and remained unanswered.

We investigate whether the weights affect eating speed by fitting a piecewise constant function across the normalized average time to complete the 'move hand towards mouth' micro-action for all 27 subjects across four different weights (i.e. normalized by person, as people move at different basic speeds). Figure 4 illustrates a normalized duration versus weight plot with a piecewise constant function shown in black and a least square fit in blue. Normalization involves percentage normalization with respect to the no-weight case for each subject (i.e., when no weight was attached to the wrists),

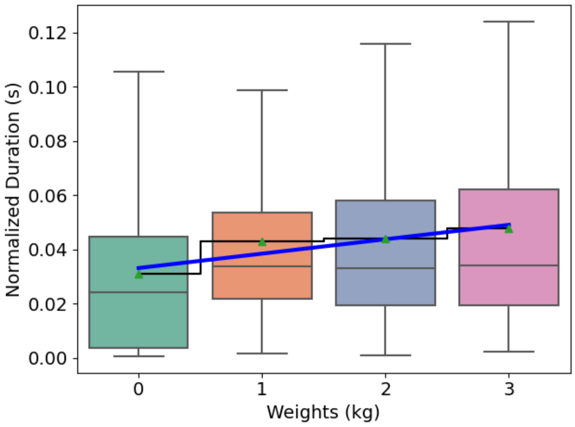

Fig. 4: Normalized duration (time taken to 'move hand towards mouth' for the 24 people in EatSense) versus weight plot with a piecewise constant function for each weight class shown in the black color and the least square fit shown in blue.

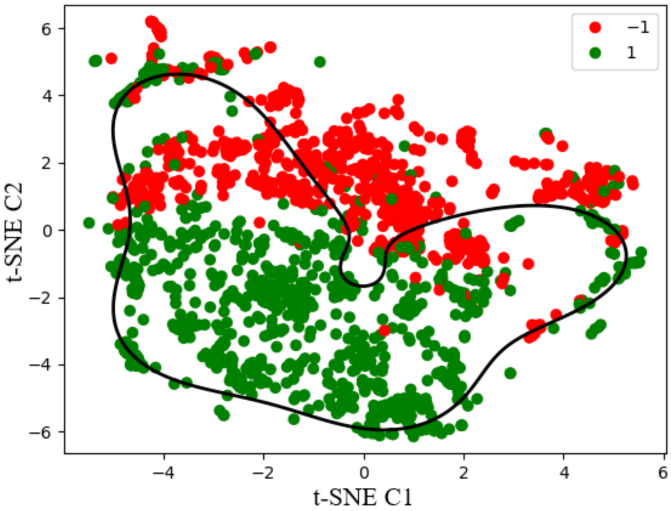

Fig. 5: The decision boundary estimated by an anomaly detector (SVM) with radial basis function as the kernel. 1 (green) indicates normal samples and -1 (red) shows anomalous samples. The black line is the decision boundary estimated by an SVM trained on the 30-D features.

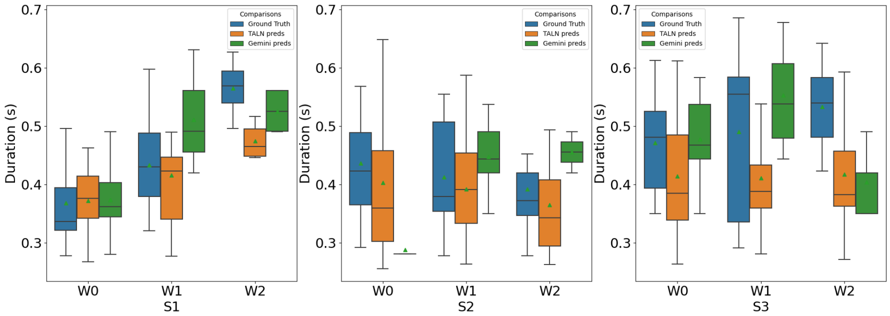

Fig. 6: Time taken for 'move hand towards mouth' micro-action (from left to right: Subject 1 (S1), Subject 2 (S2), and Subject 3 (S3)). Blue represents ground truth durations, orange indicates predictions from the proposed pipeline, and green shows estimates produced by Gemini-2.5-Pro using prompt-based video analysis. The light green triangle in each box plot marks the mean; the black line shows the median; the boxes represent the interquartile range (25th to 75th percentiles).

which aims to remove any scale bias across different subjects. The piecewise constant function (black step line) fitted is

$$f(w) = mean_p(mean_{i_p}(\frac{D_{w,p,i_p}}{\frac{1}{n_p}\sum_{j=1}^{n_p} D_{0,p,j}})) \tag{24}$$

where $D_{w,p,i}$ refers to actual time taken by subject $p \in \{0, ..., 23\}$ while performing action instance i_p with weight $w \in \{0, 1.0, 1.8, 2.4kg\}$, where each person p has n_p action instances with weight $w = 0$. Figure 4 shows a rising trend as the weights are increased on the wrists indicating gradually slowing arm movement (but 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.

To support this claim, we use the balanced dataset described in Section 4.1, randomly split into 80% training and 20% testing samples. The one-class SVM model, introduced in Section 3.2.3, is trained on the 2D output of the t-SNE projection and employs a radial basis function (RBF) kernel to learn a decision boundary. The model achieves an F1-score of 64.6% on the 2D data, suggesting that a distinguishable anomaly exists between the no-weight and maximum-weight conditions. For visualization, the postural feature vector corresponding to the last frame of each action instance is projected into a 2D space and plotted using red and green points. The learned decision boundary is illustrated as a black contour in Fig. 5. If we use the 30-D vector (postural features) directly instead of projecting into the 2D space, an SVM classifier with an RBF kernel achieves 76.2% accuracy.

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.

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 (system-wide) evaluation of the proposed pipeline is undertaken, in addition to the tests on the individual network and classifier components, as discussed in the previous subsections. The system-wide test uses the test set that consists of the nine new videos recorded similar to EatSense. They were discussed briefly in the last paragraph of section 3.1. In addition to the original pipeline, Gemini-2.5-Pro was used for an external evaluation of the same set of videos. The model was provided with a detailed natural language prompt specifying the detection task for "move hand towards mouth" actions and chewing intervals. The prompt instructed Gemini to identify action intervals, count them, and compute descriptive statistics such as mean, median, Q1, and Q3, all in milliseconds. It also estimated chewing intervals using logical assumptions regarding action ordering. This served as an independent baseline to verify the results from the proposed system.

Similarly to the inference pipeline, data preparation was: 1) extracting poses and 2) estimating the same 30 previously selected features, to get a 30-D feature vector per frame. The temporal sequence of these features is used to temporally localize actions in the video utilizing the TAL network, which outputs the two action class labels alongside the temporal segment for each of the predictions. The temporal segments are then used to estimate statistics such as the number of mouthfuls, chewing duration, and time taken for the 'move hand towards mouth' micro-action.

For end-to-end tests, we utilize the videos in the new test set, and ground truth (GT) labelled actions to estimate metrics such as accuracy for anomaly detection algorithm, tabulating the number of mouthfuls, bar chart for GT versus predicted chewing duration and estimated time taken for performing action 'move hand towards mouth'.

4.3.1 Duration of Hand-to-Mouth Actions

The 'move hand towards mouth' micro-action is important because the subject has to move their hand against gravity and changes in action duration can potentially indicate decay in strength or control. With the temporal segment extracted alongside class predictions, the duration of the 'move hand towards mouth' micro-actions can be extracted directly from the TAL network results.

Results and Discussion: Figure 6 shows the time taken to move the hand from the plate (where the food is) to the mouth. The left and the middle figure in 6 show the statistics extracted from subjects 1 and 2. They show two very interesting effects of weights on individuals. Firstly, the predicted output follows the pattern of the ground-truth values indicating the effectiveness of the proposed pipeline with the highest mean difference of 0.1 seconds approximately.

Secondly, it shows opposite trends for the time taken by subjects 1 and 3 as compared to subject 2. The trend is the change in performance (average time taken to complete the ‘move hand towards mouth’ action). As the subjects wore three different weights on their wrists, one might predict a systematic change in action duration. Subjects 1 and 3 take a longer time to complete the action with higher weights indicating a slower movement (possibly due to muscular degradation or trying harder to maintain hand-to-mouth coordination).

Subject 2, on the other hand, shows the opposite trend, indicating that they take a shorter time to complete this action with higher weights which is counter-intuitive. Since every subject shows decay differently, this leads to the question: Does subject 2 show any other parameters for decay? Upon further investigation of the videos, it was found out that subject 2 was slouching more to offset the weight on their wrists which resulted in shorter movement distances (as the food was now closer to the person’s mouth) and thus shorter time intervals for eating with higher weights. Hence, the weights altered the overall posture of subject 2.

For subject 3, it was observed that the predicted segments had low confidence scores (no predictions more than 60% confidence) and an extremely gradual increase in the mean value of the boxplots that show the time taken for ‘move hand towards mouth’ action in predictions. This is likely because the trained TAL network does not generalize well on the EatSense dataset on entirely new subjects potentially due to the lack of diversity in the dataset. This issue was also raised by Raza et. al. in [26] about EatSense. However, even if we use less confident predictions as shown in the right box plot in Fig. 6, the highest difference between the mean of the predicted and the ground truth values is still less than 0.11 seconds.

For a comparative analysis of the pipeline’s predictions, hand-to-mouth durations were also computed using Gemini-2.5-Pro. While Gemini’s estimates followed the expected trends for some videos, its performance varied considerably across conditions. In the Subject 2, weight 0 case, Gemini exhibited a catastrophic failure by predicting unrealistically low durations with almost no variance, effectively collapsing to near-constant outputs. This indicates a breakdown in which the model failed to detect normal variability in the action’s execution. Such a failure substantially distorts the statistical profile for that video and undermines the interpretability of its outputs. For Subject 3—recorded in a distraction-filled environment—Gemini’s predictions also showed higher variance and reduced alignment with ground truth, though without the severe collapse observed for Subject 2 (w0).

Overall, while Gemini occasionally captured general patterns, its predictions displayed larger discrepancies in mean values and variability for most videos. These results demonstrate that the proposed pipeline is not only more consistent but also significantly more robust against such failure cases.

4.3.2 Posture Anomaly Detection

The 3D instantaneous posture features corresponding to the temporal segments predicted by the TAL network from the new 3-person dataset are input for classification using the SVM model previously trained on the full EatSense dataset (as described in Section 4.2 claim 3). The purpose is to detect changes in posture.

Results and Discussion: The SVM for postural anomalies achieve a 64.1% F1-score. Figure 7 shows the decision boundary on the test set with six colors for

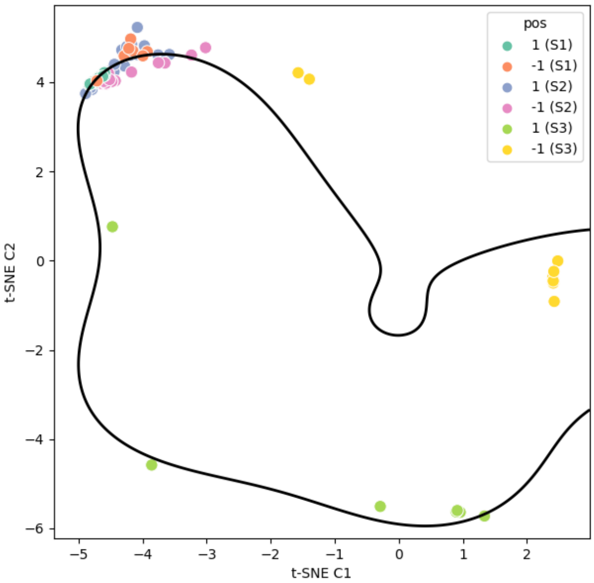

Fig. 7: The decision boundary estimated by anomaly detector (SVM) with radial basis function as the kernel. -1/1 (Sx) in the legend indicates anomalous and normal data with respect to the subject.

two weights and three subjects. For subject 1, most of the points with no weights (turquoise) show normal posture including the ones where the subject was wearing weights (orange). This is consistent with the speed tests from Fig. 6, i.e., increased time for 'moving hand to mouth action'. For subject 2, more abnormal posture points (cyan) are beyond the boundary line and normal posture points (pink) are inside the boundary, hence indicating a postural change. This is consistent with the deductions made earlier about the time taken to complete the 'move hand towards mouth' micro-action (in Section 4.3.1).

For subject 3, most of the points are marked as a normal posture that includes both weighted (yellow) and unweighted (green) cases. This shows that subject 3 showed mixed traits in their posture. This was also visible from the videos, that subject 3 was continuously trying to re-align their posture every time showing symptoms of degraded motion due to the weight on their wrist. Overall, the results of this experiment indicate that the anomaly detection algorithm can quantify postural anomalies.

Table 2: The number of mouthfuls (i.e., "move hand towards mouth" actions), comparing ground truth, the proposed pipeline (shown in column Est.), and Gemini-2.5-Pro estimates (shown in column Est. by Gemini). Gemini was given a structured natural language prompt describing the action and chewing interval definitions. W represents the weight (in Kg) the subject S1/S2/S3 was wearing in corresponding videos and Dur. shows the duration of each of the videos in 'minute:seconds' format.

				Mouthfuls		
Subject	Video ID	Dur.	W (kg)	GT	Est.	Est. by Gemini
S1	20240118_144635	4:25	0	17	18	17
	20240118_145455	2:23	1	10	11	10
	20240118_150211	1:51	2.4	6	7	6
S2	20240118_152302	2:11	0	16	17	17
	20240118_152736	2:04	1	12	11	13
	20240118_153325	1:26	2.4	8	7	8
S3	20230620_143046	5:06	0	14	16	14
	20230620_143736	5:43	1	13	12	15
	20230620_144531	5:43	2.4	13	9	13

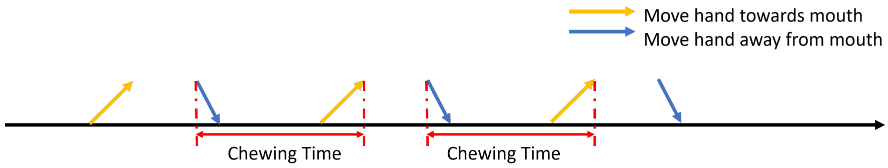

Fig. 8: Chewing duration is estimated when the 'move hand away from mouth' starts until 'move hand towards mouth' ends. This is done under the assumption that there is no distraction such as a phone conversation, another person to talk to, or other distractions.

4.3.3 Mouthfuls

To count the number of mouthfuls, the number of times a subject performs the 'move hand towards mouth' action is counted. The number of mouthfuls tracks the individual's long-term nutritional adequacy.

Results and Discussion: The number of mouthfuls is tabulated in Table 2 which shows how many mouthfuls both subjects ate in reality (the ground truth) against that estimated by the proposed pipeline and gemini-2.5-pro. The table also shows the duration of the videos and the weight worn by the subjects with their corresponding video ID.

Table 2 presents the results for both the proposed pipeline, which yields a mean absolute error (MAE) of ±1.4, and Gemini-2.5-Pro, which achieves an MAE of ±0.4. Both methods produce predictions that closely approximate the ground truth. The

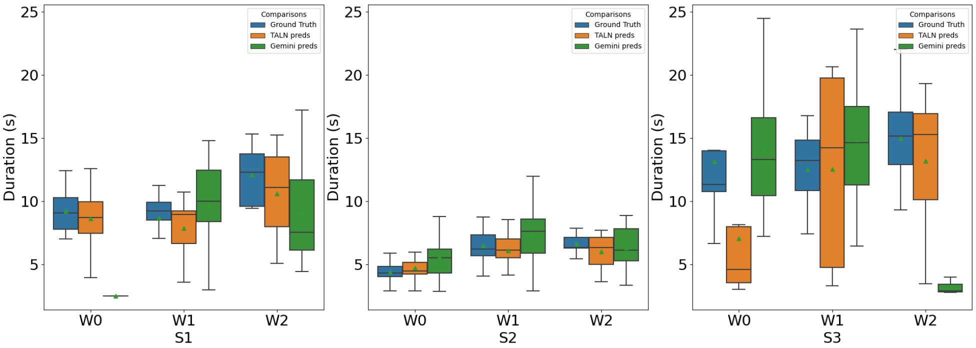

Fig. 9: Estimated chewing durations for Subjects 1 (S1), 2 (S2), and 3 (S3). Blue shows the ground truth (based on manual annotations), orange indicates predictions from the proposed pipeline, and green represents Gemini-2.5-Pro outputs derived from prompt-based inference. Each box plot shows the mean (green triangle), median (black line), and interquartile range (25th to 75th percentiles).

MAE is computed as:

$$\frac{1}{N}\sum_{i=1}^{N}|Est_i - GT_i| \tag{25}$$

where N is the total number of videos Est_i is the estimated value for the i^{th} video, and GT_i is its corresponding ground truth value. Gemini-2.5-Pro matched or slightly exceeded the proposed pipeline's accuracy in several cases. Overall, both of these not only predict the number of mouthfuls with decent accuracy but also follow patterns in the ground truth, such as fewer predicted mouthfuls for videos involving lesser mouthfuls in reality. The similarity of Gemini's counts and statistical summaries supports the correctness of the pipeline and strengthens confidence in its deployment in real-world scenarios.

4.3.4 Chewing Duration

Assuming that the environment is distraction-free and the person does not perform any other action between two consecutive 'move hand towards mouth' actions, the time taken for chewing is from the start of the 'move hand away from mouth' and the end of the subsequent 'move hand towards mouth'. This is shown in Fig. 8. Our proposed test set includes six videos involving two subjects, i.e., subjects 1 and 2 where we made sure the eating session was recorded in a distraction-free environment. We also demonstrate our proposed technique on subject 3 (marked as S3) where the environment was not distraction-free and the subject was talking and using their phone at times.

The predictions of the TAL network, due to processing errors, even after merging the overlapping temporal segments, do not have the same number of 'move hand towards mouth' and 'move hand away from mouth' action instances, which in reality should be identical. To overcome this problem, the chewing duration is estimated only if 'move hand away from mouth' is followed by 'move hand towards mouth'.

Results and Discussion: These assumptions mentioned above are reasonable for most elderly people who are living independently and have a distraction-free environment. Figure 9 shows the chewing duration for the nine videos where subject 1 (S1) is shown in the boxplot on the left, subject 2 (S2) in the middle, and subject 3 (S3) on the right.

Notably, for subjects 1 and 2, there are similar distributions of chewing time across all videos, which is expected as the same food was eaten in all six cases (possibly the chewing time distribution for subject 1 in the large weight condition was longer because this was eaten last and the volunteer was getting full). The predicted chewing time follows the pattern of the ground truth chewing time with the highest mean difference of 1.25 seconds.

However, for subject 3, the 0.25 and 0.75 quartiles of the boxplot indicate wider (low kurtosis) distribution, i.e., the subject was taking more time to chew. This is certainly because the environment was distraction-full and our proposed method of estimating chewing times requires a distraction-free environment and hence fails to get accurate estimations. However, even though accurate estimation is not possible, the boxplots still show a rising trend indicating longer chewing times which could mean fatigue/tiredness for both the ground truth and predicted segments with higher weights. On the other hand, the predictions seem somewhat erroneous because the model struggles to generalize to new subjects.

Additionally, chewing duration statistics extracted using Gemini-2.5-Pro were overlaid for comparative analysis. These values, derived from the model's structured textual output, demonstrated partial alignment with both the ground truth and the pipeline's predictions. However, Gemini's outputs generally showed greater mean differences from the ground truth, even for subjects recorded in distraction-free environments. In the Subject 1, weight 0 case, Gemini exhibited a catastrophic failure by predicting unrealistically low chewing durations with almost no variance, effectively collapsing to near-constant outputs and missing the natural variability present in the ground truth. A similar failure occurred for Subject 3, weight 2, where the model again produced nearly constant, underestimated durations, leading to a distorted statistical profile for that video. For Subject 3 more broadly, Gemini's estimates showed increased variability and larger mean errors in other conditions, reflecting reduced temporal alignment. These cases illustrate that while Gemini can approximate coarse chewing-time trends, it is more prone to collapse and large-magnitude errors than the proposed pipeline, which maintained consistent and plausible estimates across all weight and environmental conditions.

To summarize, the predicted chewing times follow the general trend of the ground truth chewing times hence indicating the effectiveness of the pipeline as a stepping stone to draw further insights such as identifying potential eating disorders.

Note: The charts and tables showing performance are just indicative of the overall effectiveness of the pipeline and a simulation of deterioration. In a real scenario, data is likely to be collected every week, which would be tracked over months or years. The weekly timescale will allow better averaging and removal of outliers.

5 Limitations and Future Work Directions

While the proposed system shows strong performance in controlled settings, its generalizability might be limited as it is tested on a relatively small dataset. While it performs well on data resembling the training set, results from new individuals showed slightly reduced confidence and temporal consistency in predictions. This suggests a potential domain shift and highlights the need for improved model robustness across diverse movement patterns and eating styles. The system also assumes a distraction-free environment for accurate estimation of chewing duration and behavioral patterns. This might be true for elderly living independently, but in real-life scenarios, distractions such as conversations or phone use can affect chewing time estimation accuracy.

Future work should focus on expanding the dataset to include more diverse participants and naturalistic environments. Integrating additional sensing modalities and deploying the system for long-term, in-home use could improve its robustness. Real-time feedback and more advanced anomaly detection methods may also enhance its clinical utility and support early intervention.

6 Conclusion

This paper presents a fully autonomous vision-based pipeline for analyzing eating behaviors and monitoring musculoskeletal function, offering a novel approach to understanding the interplay between dietary habits and physical health. The system combines pose estimation, temporal action localization, and advanced data augmentation techniques to assess key metrics such as hand-to-mouth motion duration, bite counts, and chewing times. These indicators offer insights into eating behaviors and physical performance, supporting the potential use of the pipeline for long-term health monitoring and early detection of changes in musculoskeletal and eating patterns.

The system was validated using the EatSense dataset and a new test set, demonstrating its ability to accurately identify patterns in eating behavior and movement changes under controlled conditions. The paper demonstrates the effectiveness of the method through both component-level and holistic analyses (e.g., TAL network achieves 74% mAP@0.10 and anomaly detection (SVM) achieves 64.2%) and by capturing trends successfully with holistic tests. To complement this evaluation, we compared the pipeline's output statistics with those from Gemini-2.5-Pro, a state-of-the-art multimodal model. While Gemini produced reasonable mouthful-count estimates, it exhibited multiple catastrophic failures in tasks requiring precise temporal segmentation. In these cases, Gemini's predictions collapsed to unrealistically low values with almost no variance, producing near-constant outputs and failing to capture the natural variability present in the actions, thus producing distorted statistical profiles. In contrast, the proposed pipeline maintained stable and plausible predictions across all subjects and weight conditions, avoiding collapse even in distraction-filled environments.

However, the system exhibited limited generalizability to new subjects, so further efforts are needed to improve its robustness across diverse individuals. Future work may focus on expanding the dataset to include more diverse subjects, deploying the system for long-term real-world monitoring, and incorporating real-time feedback mechanisms to support early health interventions. In summary, the results indicate that the pipeline can capture relevant trends, including variations in arm movement speed due to wrist weights and changes in chewing behavior. These findings support the system's potential as a basis for further research into automated monitoring of physical and behavioral health indicators.

Acknowledgment

The authors thank all the volunteers who participated in the data collection process. Ethics approval was obtained for data collection and distribution.

Funding

Muhammad Ahmed Raza received research support from the Higher Education Commission of Pakistan, and Robert B. Fisher. received financial support from the Advanced Care Research Centre (ACRC) at the University of Edinburgh required for the EatSense data collection.

References

[1] Debnath, B., O'brien, M., Yamaguchi, M., Behera, A.: A review of computer vision-based approaches for physical rehabilitation and assessment. Multimedia Systems **28**(1), 209–239 (2022)

[2] Topham, L.K., Khan, W., Al-Jumeily, D., Hussain, A.: Human body pose estimation for gait identification: A comprehensive survey of datasets and models. ACM Computing Surveys **55**(6), 1–42 (2022)

[3] Topham, L.K., Khan, W., Al-Jumeily, D., Waraich, A., Hussain, A.J.: Gait identification using limb joint movement and deep machine learning. IEEE Access **10**, 100113–100127 (2022)

[4] Gao, Z., Chen, J., Liu, Y., Jin, Y., Tian, D.: A systematic survey on human pose estimation: upstream and downstream tasks, approaches, lightweight models, and prospects. Artificial Intelligence Review **58**(3), 68 (2025)

[5] Raza, M.A., Chen, L., Nanbo, L., Fisher, R.B.: Eatsense: human centric, action recognition and localization dataset for understanding eating behaviors and quality of motion assessment. Image and Vision Computing **137**, 104762 (2023)

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

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

[8] Raihan, M.J., Ahad, M.A.R., Nahid, A.-A.: 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), vol. 28, pp. 1–6 (2021). IEEE

[9] Mottaghi, E., Akbarzadeh-T, M.-R.: Automatic evaluation of motor rehabilitation exercises based on deep mixture density neural networks. Journal of Biomedical Informatics **130**, 104077 (2022)

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

[11] Elkholy, A., Hussein, M.E., Gomaa, W., Damen, D., Saba, E.: Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance. IEEE journal of biomedical and health informatics **24**(1), 280–291 (2019)

[12] Hiraguchi, H., Perone, P., Toet, A., Camps, G., Brouwer, A.-M.: Technology to automatically record eating behavior in real life: A systematic review. Sensors **23**(18), 7757 (2023)

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

[14] Nour, M., Gardoni, M., Renaud, J., Gauthier, S.: Real-time detection and motivation of eating activity in elderly people with dementia using pose estimation with tensorflow and opencv. Adv Soc Sci Res J **8**(3), 28–34 (2021)

[15] Okamoto, K., Yanai, K.: 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, pp. 331–335 (2016). Springer

[16] Okamoto, K., Yanai, K.: Real-time eating action recognition system on a smartphone. In: 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6 (2014). IEEE

[17] Zoidi, O., Tefas, A., Pitas, I.: Exploiting the svm constraints in nmf with application in eating and drinking activity recognition. In: 2013 IEEE International

Conference on Image Processing, pp. 3765–3769 (2013). IEEE

[18] Hossain, D., Imtiaz, M.H., Ghosh, T., Bhaskar, V., Sazonov, E.: Real-time food intake monitoring using wearable egocnetric camera. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 4191–4195 (2020). IEEE

[19] Lasschuijt, M.P., Brouwer-Brolsma, E., Mars, M., Siebelink, E., Feskens, E., Graaf, K., Camps, G.: Concept development and use of an automated food intake and eating behavior assessment method. JoVE (Journal of Visualized Experiments) (168), 62144 (2021)

[20] Konstantinidis, D., Dimitropoulos, K., Ioakimidis, I., Langlet, B., Daras, P.: A deep network for automatic video-based food bite detection. In: Computer Vision Systems: 12th International Conference, ICVS 2019, Thessaloniki, Greece, September 23–25, 2019, Proceedings 12, pp. 586–595 (2019). Springer

[21] Hossain, D., Ghosh, T., Sazonov, E.: Automatic count of bites and chews from videos of eating episodes. Ieee Access **8**, 101934–101945 (2020)

[22] Konstantinidis, D., Dimitropoulos, K., Langlet, B., Daras, P., Ioakimidis, I.: Validation of a deep learning system for the full automation of bite and meal duration analysis of experimental meal videos. Nutrients **12**(1), 209 (2020)

[23] Cadavid, S., Abdel-Mottaleb, M., Helal, A.: Exploiting visual quasi-periodicity for real-time chewing event detection using active appearance models and support vector machines. Personal and Ubiquitous Computing **16**, 729–739 (2012)

[24] Alshboul, S., Fraiwan, M.: Determination of chewing count from video recordings using discrete wavelet decomposition and low pass filtration. Sensors **21**(20), 6806 (2021)

[25] Tufano, M., Lasschuijt, M., Chauhan, A., Feskens, E.J., Camps, G.: Capturing eating behavior from video analysis: A systematic review. Nutrients **14**(22), 4847 (2022)

[26] Raza, M.A., Fisher, R.B.: Vision-based approach to assess performance levels while eating. Machine Vision and Applications **34**(6), 124 (2023)

[27] Luo, Z., Hsieh, J.-T., Balachandar, N., Yeung, S., Pusiol, G., Luxenberg, J., Li, G., Li, L.-J., Downing, N.L., Milstein, A., et al.: Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring. Machine Learning for Healthcare (MLHC) **2**(1) (2018)

[28] Huang, X., Wicaksana, J., Li, S., Cheng, K.-T.: Automated vision-based wellness analysis for elderly care centers. In: Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence, pp. 321–333 (2022)

[29] Raza, M.A., Lochhead, C., Fisher, R.B.: Effect of face obfuscation methods on pose-based action recognition. In: International Conference on AI in Healthcare (2024)

[30] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)

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

[32] Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5386–5395 (2020)

A

Appendix A: Gemini-2.5-Pro Prompt for Holistic Evaluation Comparisons

For the external comparative experiments described in Section 4.3, we used the following natural language prompt with Gemini-2.5-Pro for each video in the evaluation set:

Prompt: *Please analyze the video and count the number of times the action "move hand towards mouth" occurs (move hand to mouth: as soon as the hand moves away from the plate towards the mouth until just before it reaches the mouth). For each instance, identify the start and end times of the action. Then, calculate the average duration, median duration, 25th percentile (Q1), and 75th percentile (Q3) of these actions. In addition, estimate the durations of chewing throughout the video. Provide the original intervals, average chewing time, median, Q1, and Q3 values for these intervals as well. Provide all these statistics in milliseconds.*

The model's responses were parsed to extract:

- Per-instance start and end times for "move hand towards mouth" actions.
- Descriptive statistics (mean, median, Q1, Q3) for these actions.
- Estimated chewing durations with corresponding descriptive statistics.

These results were used to compare Gemini's performance against the proposed TALN-based pipeline for mouthful count, hand-to-mouth duration, and chewing duration estimation.