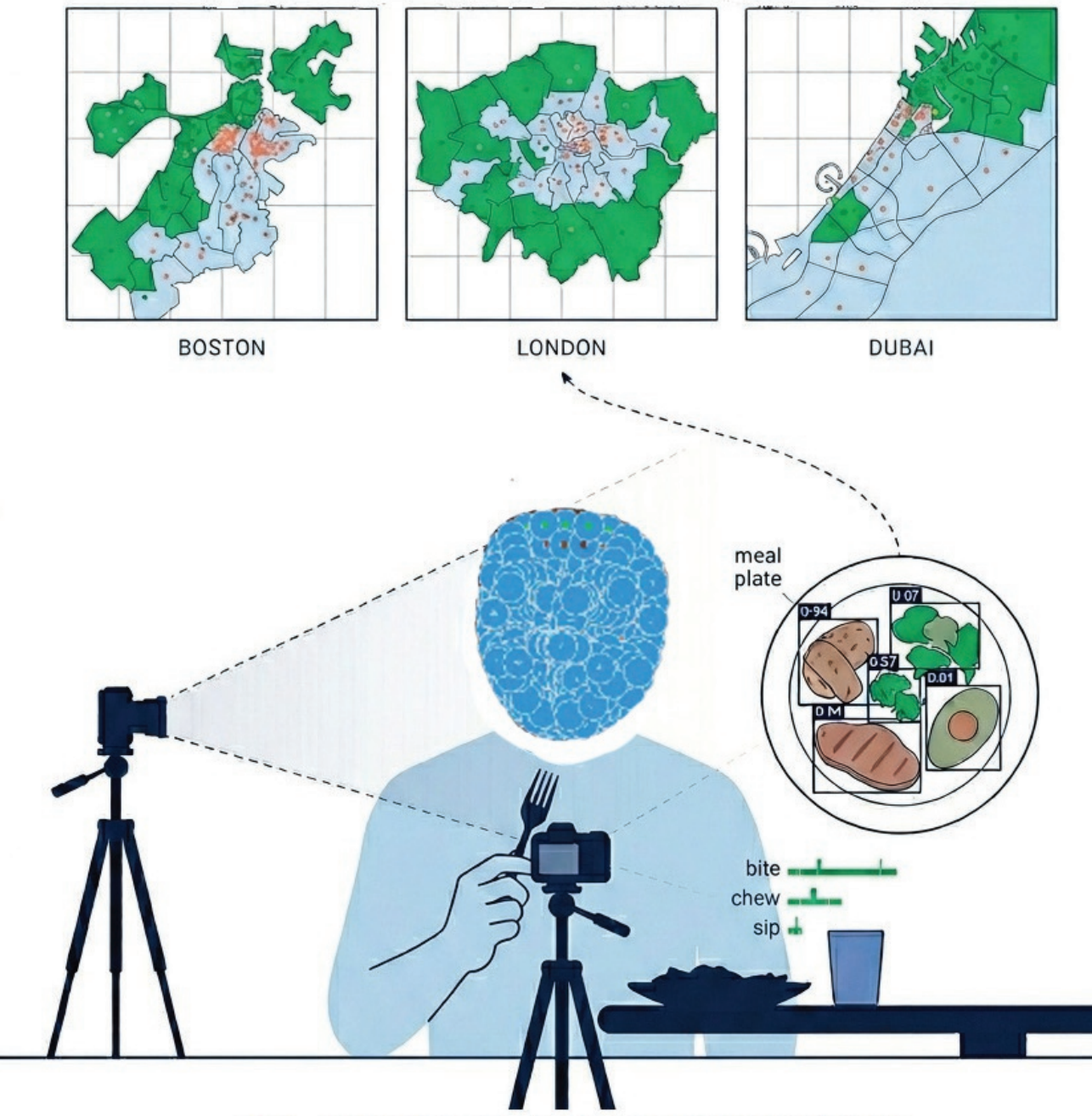


AI AND COMPUTER VISION FOR QUANTITATIVE ASSESSMENT OF EATING BEHAVIOR AND URBAN FOOD LANDSCAPES

MICHELE TUFANO

Propositions

Propositions

1. The burden of filming people while eating limits the effectiveness of real-time eating speed reduction.
(this thesis)

2. It is each citizen's right to have access to nutritious foods.
(this thesis)

3. Interdisciplinary side projects, beyond the PhD thesis, are essential to foster creativity.

4. Assumptions underlying the Impact section of a scientific proposal are best validated in consultation with end users.

5. Consumers' preferences shape the food environment, thereby steering the obesity epidemic.

6. It is the responsibility of every citizen to cultivate critical thinking and the pursuit of knowledge.

Propositions belonging to the thesis, entitled

Artificial intelligence and computer vision for quantitative assessment of eating behavior and urban food landscapes

Michele Tufano
Wageningen, 19/06/2026

Artificial intelligence and computer vision for quantitative assessment of eating behavior and urban food landscapes

Michele Tufano

THESIS COMMITTEE

Promotors
Prof. Dr ir E. J. M. Feskens
Professor in Global Nutrition
Wageningen University & Research

Co-promotor
Dr Guido Camps
Associate professor in Global Nutrition
Wageningen University & Research

Dr Marlou Lasschuijt
Assistant Professor in Sensory Science and Eating Behaviour
Wageningen University & Research

Other members
Prof. Dr A.M. Brouwer, Netherlands Organization for Applied Scientific Research TNO
Dr Philipp Rouast, Rouast Labs, Australia, Newcastle
Dr G.W. Kootstra, Wageningen University & Research
Prof. Dr M.C. Onwezen, Wageningen University & Research

This research was conducted under the auspices of the VLAG Graduate School (Biobased, Biochemical, Chemical, Food, and Nutritional Sciences).

Artificial intelligence and computer vision for quantitative assessment of eating behavior and urban food landscapes

Michele Tufano

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr C. Kroeze,
In the presence of the
Theis Committee appointed by the Academic Board
to be defended in public
on Friday 19 June 2026
at 1 p.m. in the Omnia Auditorium.

Michele Tufano

Artificial intelligence and computer vision for quantitative assessment of eating behavior and urban food landscapes, 166 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2026)

With references, with summary in English

DOI https://doi.org/10.18174/681714

TABLE OF CONTENTS

Chapter 1	Introduction	7
Chapter 2	Data-Driven Nutritional Assessment of Urban Food Landscapes: Insights from Boston, London, and Dubai	25
Chapter 3	Capturing eating behavior from video analysis: a systematic review	51
Chapter 4	Rule-based systems to automatically count bites from meal videos	79
Chapter 5	Automatic count of bites, chews, and sips from meal videos using vision transformers	107
Chapter 6	A Cloud-Based Toolkit for Automated Meal Video Analysis: Enabling Large-Scale Eating Behavior Research	131
Chapter 7	Discussion	147
Summary		159
Acknowledgements		161
About the Author		165
PhD candidate's authorship statement		166

1

Introduction

Artificial Intelligence (AI) is rapidly evolving and driving innovation across many fields, including nutritional science. This thesis explores the application of artificial intelligence as a tool for researchers in nutritional and eating-behavior science. The AI and computer vision developments presented in this dissertation have two main areas of application. The first operates at the population level and focuses on investigating how novel AI technologies can be used to ensure that populations have access to nutritious food. The second area applies AI at the individual level and aims to monitor dietary intake and eating behavior, focusing on how people eat.

URBAN FOOD LANDSCAPES AND NUTRITIONAL ASSESSMENT

Understanding the food environment's impact on individual eating behavior is critical, as consumption patterns vary significantly based on the food environment. The food environment encompasses both processed and whole foods, creating differential effects on individual dietary patterns and eating behavior. Hence, addressing global nutritional challenges requires integrated assessment of both environmental factors and individual behaviors. However, current approaches, such as observational field studies and self-report questionnaires, limit scalability, standardization and temporal resolution, constraining our understanding of nutritional environments in diverse urban settings [1–3]. Meanwhile, recent advances in artificial intelligence and computer vision present opportunities for modernizing food environment assessment. Deep learning architectures now enable automated nutritional classification of digitally available restaurant data. This offers significant potential for large-scale, granular mapping of urban food landscapes and socioeconomic disparities to be used for nutrition research. These computational approaches create the opportunity to monitor the nutrition environment over longer periods of time, supporting demographic research and, potentially, food policies.

Chapter 2 explores these techniques and their application to understanding nutritional accessibility patterns across metropolitan regions. It introduces a scalable, data-driven framework that quantifies urban food landscapes using restaurant menu data from online delivery platforms. By linking menu items to standardized nutritional databases through machine learning, this approach enables neighborhood-level assessment of dietary quality and its relationship with socioeconomic conditions and obesity risk.

EVOLUTION OF EATING BEHAVIOR RESEARCH

The scientific study of eating behavior has expanded from early physiological experiments to modern computational methods, driven by technological innovation. In the early 20th century, Cannon's pioneering Cannon-Washburn experiment established a direct link between gastric contractions and hunger perception, showing for the first time that stomach distension could modulate hunger independent of nutrient intake[4]. This physiological understanding of hunger created the need for methods capable of capturing not only biological signals but also actual food consumption in real-world settings. Consequently, from the mid-20th century onward, dietary assessment methods such as the diet history method were introduced to retrospectively measure long-term habitual dietary intake[5], complemented by controlled laboratory studies that quantified food consumption under standardized conditions[6]. As research increasingly acknowledged the role of cognition and behavior in eating, the 1980s marked a further shift with the emergence of psychometric instruments such as the Three-Factor Eating Questionnaire (TFEQ), which integrated behavioral science into dietary assessment by quantifying cognitive restraint, disinhibition, and hunger[7]. More recently, technological advances have enabled objective and continuous monitoring of eating events in free-living conditions using wearable sensors such as EMG devices and smartwatches[8]. This shift from invasive or subjective questionnaire methods to automated, non-invasive measurement of eating behavior has laid the groundwork for modern computational assessment of eating behaviors.

SATIETY MECHANISMS AND CLINICAL RELEVANCE

Understanding satiety and satiation mechanisms is crucial for developing effective behavioral interventions, as this process shapes portion size selection and meal termination. Individuals rely on expected satiation and expected satiety—anticipating both immediate fullness and prolonged hunger prevention—to decide how much to serve and consume [9,10]. However, this process is shaped by many individual factors and influenced by many environmental factors. For example, packaging and portion size affect food and energy intake[11]. Energy density (kcal/g), food texture, eating rate (g/min), and eating intake rate (kcal/min) influence satiety and are determined by mastication and oral processing duration[12]. For instance, minimal chewing promotes faster consumption of food (g/min) and increased intake, while foods that require extensive mastication—like carrots— slow the eating rate and enhance satiety, which in turn reduces overall intake[13–18]. Given rising rates of metabolic imbalance due to chronic

overconsumption, these insights are increasingly relevant [19,20]. Traditional interventions focused on nutrient composition [21], but recent research findings underscore the importance of oro-sensory processes, such the sensory experience of food's taste, smell, and texture, and behavioral factors, such as bite size and eating rate [16,22]. In this context, bite counting has emerged as a scalable proxy for monitoring oral processing behavior [23]. Metrics like bite rate and bite size can differentiate eating patterns and serve as early indicators of dysregulated food intake. As a quantifiable measure, bite counting provides actionable data for designing interventions to slow eating, strengthen satiety signals, and improve metabolic health.

CURRENT MEASUREMENT METHODS

Accurate measurement of oral processing parameters (bite count, chewing patterns, drinking episodes, and their temporal characteristics) is essential for establishing relationships between food structural properties and satiety responses. This relationship enables evidence-based modeling of how textural food attributes affect satiety and postprandial appetite regulation [24,25]. However, current methodologies for quantifying oral processing behaviors and eating intake predominantly rely on human manual annotation [26]. Despite offering precision and standardization, these approaches present substantial limitations for their objectivity and scalability due to their repetitive nature, time-intensive requirements, and impracticality for large-scale investigations such as prospective studies [27]. Consequently, development of automated and high-throughput methodologies to assess oral processing represents a significant research priority. Innovative methodologies would enable more comprehensive evaluation of eating behavior and would help researchers to address how textural modifications can be used as potential interventions for appetite regulation in clinical populations [28,29].

Recent advances in sensing technologies have enabled novel approaches to measure human eating and drinking behaviors. These measurements can be categorized into several distinct domains based on the specific behavioral aspects being monitored [30]: wearable sensors, environmental sensing, weight-based approaches, acoustic sensing, imaging, and multimodal approaches.

Wearable sensor systems represent a prominent approach for detecting consumption activities. Inertial measurement units (IMUs), which typically incorporate accelerometers, gyroscopes, and magnetometers, can be strategically positioned on the wrist, upper arm, head, and neck to capture movement patterns

associated with eating and drinking [31–34]. These sensors, commonly integrated into commercial smartwatches or purpose-built wearables, can identify hand-to-mouth gestures with minimal user disruption during natural consumption activities [31]. Alternative sensor types, including proximity sensors, radar sensor, and piezoelectric devices, have demonstrated efficacy in detecting specific jaw movements or hand positioning related to food intake [35,36]. Environmental sensing represents another methodological approach. Pressure-sensitive surfaces embedded in tablecloths and serving equipment can quantify actions such as cutting, scooping, or stirring, while simultaneously tracking weight changes in dishware [37]. This approach offers the advantage of minimal user burden but requires instrumentation of the eating environment. Novel developments in this domain include specialized dining trays equipped with multiple weighing stations that continuously monitor food and beverage container weights throughout a meal. Importantly, such instrumented environmental systems overcome limitations of conventional self-report methods by directly measuring consumption parameters without reliance on memory or portion size estimation capabilities [38].

Portion size estimation involves complex methodological challenges requiring diverse measurement approaches. Weight-based approaches calculate consumption by comparing pre- and post-meal measurements of food containers. Image analysis techniques have advanced significantly, with RGB-D (Red, Green, Blue plus Depth) imaging enabling more accurate volume calculations by incorporating height information [39]. Some innovative approaches combine multiple sensing modalities, such as pairing acoustic data with images to enable scaling without reference objects [40,41]. For beverage consumption, various approaches have been explored, including smartwatch-based sip duration analysis [42], capacitive and conductivity sensing for cup fill levels [43], and thermal imaging for volume estimation [44]. Novel integrated systems, such as the Playful Bottle system, combine smartphone accelerometers to detect drinking motions with camera imaging to monitor fluid levels [45].

The measurement of specific consumption parameters (e.g., bites, chews) requires specialized technological approaches. Motion sensing technology has been effectively used to detect bites, which serves as a fundamental metric in understanding eating patterns. instance, bite weight estimation has been accomplished using various sensing modalities, including acoustic monitoring and weight sensors [46,47]. Chewing and swallowing activities represent particularly well-studied aspects of consumption behavior. Multiple sensing modalities have been employed, with gyroscopes (motion sensors that measure rotational movements) often utilized for chewing detection, accelerometers for swallow-

ing events, and proximity sensors for hand-to-mouth gesture identification [48]. Acoustic monitoring via microphones has proven effective for capturing the characteristic sounds of both chewing and swallowing. More specialized approaches include electroglottography (EGG) and electromyography (EMG), which directly measure physiological changes during consumption, though these typically impose greater user burden than other methods [49,50].

Imaging technologies, including standard cameras and depth-sensing devices, provide yet another avenue for monitoring consumption, though these typically necessitate maintaining visibility of the consumer's face throughout the eating episode [51]. Commercial software solutions such as OpenPose have been applied to automate bite count from video recordings [52]. While cameras can provide rich contextual information about eating behaviours, these approaches raise significant privacy concerns and often require active user involvement (e.g., taking photos before and after meals). Video-based approaches provide functionality that extends beyond the capabilities of current wearable sensor systems, enabling not only the quantification of mechanical eating parameters (e.g., consumption rate, bite frequency) but also facilitating the analysis of complex behavioral dimensions including emotional correlates of consumption, social interaction dynamics during meals, and specific interpersonal behaviors such as parent-child feeding interactions [47,53]. Additionally, image analysis algorithms may struggle with lightning conditions, and overall accuracy, but the latest development in computer vision show promising opportunities for this field.

RGB imaging technologies, coupled with advanced computer vision algorithms, represent an optimal approach for quantifying eating behaviours due to their non-invasive nature and comprehensive data capture capabilities [54]. These systems can simultaneously measure multiple oral processing parameters (bite frequency, chewing cycles, drinking, and food manipulations) while providing contextual information about meal environments and social interactions that influence consumption patterns. Unlike wearable sensors or specialized equipment, RGB cameras are non-intrusive during naturalistic eating scenarios, requiring no physical contact with participants that might alter typical consumption behaviours [55]. Privacy concerns can be effectively mitigated through facial masking techniques and anonymization algorithms that protect subject identity while preserving critical motion data necessary for behavioural analysis. Furthermore, the accelerating development of deep learning architectures has addressed previous limitations in variable lighting conditions and subject positioning, enabling robust automated analysis of naturalistic eating scenarios. This computer vision-based methodologies are particularly promising for large-scale, ecologically valid investigations of eating behaviour that can generate the

high-dimensional datasets necessary for modelling relationships between food properties and satiety outcomes [56].

ADVANCEMENTS IN AI AND COMPUTER VISION

Artificial Intelligence (AI) has evolved from a theoretical concept to a powerful and widely adopted tool, offering capabilities from data analysis to personalized recommendations. A key branch of AI—Computer Vision—enables machines to interpret and understand visual information, effectively allowing computers to "see", and quantify the observed behaviors.

Computer vision represents a multidisciplinary field intersecting digital image processing, pattern recognition, machine learning, and artificial intelligence. The technological domain encompasses methodologies for image acquisition, processing, and analytical interpretation. Computer vision architectures leverage convolutional neural networks to extract hierarchical feature representations from visual data, eliminating the necessity for explicit feature engineering that characterized traditional approaches [57]. The predominant paradigm shift toward deep learning methodologies has facilitated significant performance improvements across object detection, segmentation, and recognition tasks, with applications extending to autonomous vehicles, medical diagnostics, agricultural monitoring, biometric authentication, industrial quality control systems, and nutritional science contemporary computer vision frameworks integrate pixel optimization techniques with three-dimensional morphological analysis to enhance image interpretation capabilities[58].

Within the domain of computer vision-based eating behavior analysis, facial and body landmarks (e.g., eyes, nose, arms, and hands) serve as critical anatomical reference points for multimodal behavioral assessment. These morphological features facilitate precise spatiotemporal tracking of orofacial movements, hand-to-mouth trajectories, and body positioning during food consumption episodes [59]. The algorithmic extraction and tracking of these landmarks enable quantitative characterization of mastication patterns, bite mechanics, and gestural components of eating behavior, providing objective metrics for clinical assessment and dietary monitoring applications.

ALGORITHMS

Computational vision systems employ diverse algorithmic approaches that vary in architectural complexity and learning paradigms. **Rule-based algorithms** demonstrate optimal performance in well-constrained problem domains where domain knowledge can be explicitly encoded through formal specifications and heuristics. In contrast, **deep learning** methodologies utilize hierarchical neural network architectures to automatically extract and learn relevant features from extensive training datasets, achieving superior performance in complex perceptual tasks such as object recognition and scene understanding.

In the evolution of computer vision techniques, rule-based approaches represented early methodological frameworks where explicit feature engineering was required. As Chai et al. note, traditional computer vision methods like "K-means, Naive Bayes classifier, Decision Tree, Boosting, Random Forest, Haar Classifier, Expectation–Maximization (EM), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM)" dominated before deep learning's emergence [60]. The Viola-Jones detector exemplifies this approach, utilizing "Haar-like wavelet feature and integral graph method for face detection," with subsequent extensions incorporating "rotated Haar-like features" to form the Haar classifier implemented in OpenCV. Deep learning architectures, particularly Convolutional Neural Networks (CNNs), revolutionized computer vision by enabling automatic feature extraction through hierarchical representations. CNNs consist of convolutional layers, pooling layers, and fully connected layers, where various kernels convolve the whole image and generate feature maps. This approach is more efficient in learning patterns from images because it does not require the manual feature engineering that characterized traditional rule-based systems. The historical progression from AlexNet (2012)[61] through VGGNet[62], GoogLeNet [63], ResNet[64], DenseNet[65], and MobileNets[66] demonstrates the field's rapid advancement in designing increasingly efficient network structures. These networks can be trained more efficiently with more data and can extrapolate more complex patters from images and videos, therefore increasing the level of learning and understanding. Transformer architectures represent the latest paradigm shift in computer vision systems [67]. While initially developed for natural language processing, these models have been adapted for visual tasks due to their "self-attention mechanisms," which efficiently model long-range dependencies in data. Unlike CNNs that process information through local convolutional operations, transformers capture global contextual relationships across image regions [68].

The emergence of transformer architectures represents a significant advancement in vision processing, employing self-attention mechanisms to model long-range dependencies within visual data. These mechanisms establish global contextual relationships more efficiently than previous architectural paradigms such as CNN, particularly when processing sequential or spatially distributed visual information. This architectural innovation aligns with the evolution process from supervised learning to semi-supervised learning, enabling more sophisticated visual understanding with reduced dependency on extensive labeled datasets. [60]

Despite significant advancements in imaging hardware and computational algorithms, a substantial research gap persists in the application of modern computer vision approaches to eating behaviour analysis. Notably, **rule-based algorithms** and **transformer architectures** remain largely unexplored in this domain, despite their demonstrated efficacy in other computer vision applications.

THE PURPOSE OF THIS THESIS

The aim of this thesis is to explore the recent advancements in AI and computer vision to automate the assessment of eating behavior and urban food landscapes, providing scalable tools to support nutrition research at both individual and population levels.

On the individual level, two fundamental research questions require systematic investigation: (1) Which computer vision methodologies should be implemented to optimally automate video-based eating behaviour assessment with high reliability while maintaining minimal invasiveness, and (2) What architectural approaches might optimize the trade-off between system-specific calibration parameters and generalizable computational models across diverse populations?

These challenges necessitate innovative methodological frameworks that address the specificity-generalizability continuum while maintaining robust performance across demographic variations, environmental contexts, and diverse food categories. Current implementations demonstrate adequate performance in controlled experimental settings but exhibit significant performance degradation when confronted with real-world variability in illumination conditions, camera positioning, culturally specific eating practices, and morphological diversity among human subjects. This underscores the necessity for adaptive

computational frameworks that balance individualized measurement precision with broad applicability across population subgroups and consumption patterns.

This thesis presents a computational framework designed for the quantitative assessment of human eating behaviors through video-based analysis of meal consumption episodes and population-level food environment characterization. The research described in this thesis provides three main methodological contributions for the quantitative assessment of eating behaviors at both individual and population scales.

Firstly, it provides an automated, real-time quantification of food intake parameters through computer vision techniques that analyze meal consumption episodes frame-by-frame. This methodological advancement enables researchers to conduct larger-scale, more precise eating behavior studies with significantly reduced manual annotation burden. The automated system captures temporal dynamics of eating behaviors including bites and chewing patterns.

Secondly, it contributes to an algorithmic infrastructure capable of developing feedback systems for behavioral interventions. This technological framework facilitates investigation of behavioral interventions specifically targeting eating rate modification.

Furthermore, the framework extends computational analysis from individual-level assessment to population-level food environment quantification through AI and data integration techniques. This contribution addresses the critical gap between individual eating behaviors and the broader food landscape that influences dietary choices. The methodology provides comprehensive nutritional profiling of available food options across urban environments, enabling quantification of nutrient availability and accessibility patterns at the population scale.

These computational approaches address significant methodological limitations in current eating behavior research, which has historically relied on labor-intensive manual coding procedures or intrusive measurement techniques that potentially alter naturalistic eating patterns. Traditional food environment assessments have been limited by categorical classifications and proximity measures without considering actual nutritional content of available foods.

THESIS STRUCTURE

This thesis consists of five chapters (Figure 1).

Chapter 2 presents methodological approaches for quantifying urban food environments, specifically focusing on computational techniques for assessing food availability across geographical regions. By establishing environmental determinants of food access, this research provides the macro-environmental context that influences individual dietary choices and eating behaviors.

Chapter 3 presents a systematic review of video-based methodologies for eating behavior assessment. This analysis examines approaches for quantifying eating behaviors across diverse recording conditions and evaluates the technical limitations and methodological strengths of existing approaches. Building upon this theoretical framework, **Chapter 4** introduces a rule-based computational system that implements algorithmic bite detection and quantification from video recordings of meal occasions. This approach leverages explicit feature engineering and deterministic decision models to identify and count feeding events.

Chapter 5 presents an advanced methodological framework utilizing deep learning architectures, specifically employing transformer-based models for the prediction and enumeration of bite and mastication events. This approach leverages self-attention mechanisms to capture temporal dependencies in feeding sequences, demonstrating superior performance in variable environmental conditions. Chapters 4 and 5 thus present complementary methodological paradigms—rule-based versus learning-based approaches—providing comparative performance analysis across diverse implementation contexts.

Chapter 6 focuses on the translational aspects of these computational frameworks, detailing the software implementation, validation protocols, and deployment architecture necessary for large-scale adoption by researchers investigating eating behavior patterns. This chapter addresses critical engineering considerations including computational efficiency, interface design, and integration with existing research methodologies.

While Chapter 2 concentrates on population-level environmental assessment, Chapters 3-6 focus on individual-level behavioral analysis. This research integrates individual behavior with environmental determinants, establishing a comprehensive analytical framework that bridges macro-environmental profiling with micro-behavioral measurements.

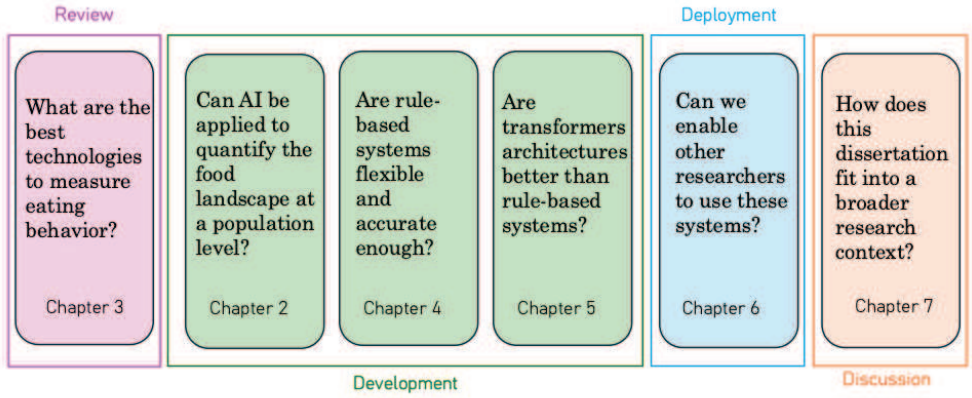


Figure 1 – Overview of the chapters presented in this thesis.

REFERENCES

1. Lytle, L. A. & Sokol, R. L. Measures of the food environment: A systematic review of the field, 2007-2015. Health Place **44**, 18–34 (2017).
2. Gamba, R. J., Schuchter, J., Rutt, C. & Seto, E. Y. W. Measuring the food environment and its effects on obesity in the United States: a systematic review of methods and results. J. Community Health **40**, 464–475 (2015).
3. McKinnon, R. A., Reedy, J., Morrissette, M. A., Lytle, L. A. & Yaroch, A. L. Measures of the food environment: a compilation of the literature, 1990-2007. Am. J. Prev. Med. **36**, S124-133 (2009).
4. Cannon, W. B. & Washburn, A. L. An explanation of hunger. 1911. Obes. Res. **1**, 494–500 (1993).
5. Measurement Toolkit - Diet histories. https://www.measurement-toolkit.org/diet/subjective-methods/diet-history.
6. Robinson, E., Bevelander, K. E., Field, M. & Jones, A. Methodological and reporting quality in laboratory studies of human eating behavior. Appetite **125**, 486–491 (2018).
7. Stunkard, A. J. & Messick, S. The three-factor eating questionnaire to measure dietary restraint, disinhibition and hunger. J. Psychosom. Res. **29**, 71–83 (1985).
8. Nicholls, B. et al. An EMG-based Eating Behaviour Monitoring system with haptic feedback to promote mindful eating. Comput. Biol. Med. **149**, 106068 (2022).
9. Brunstrom, J. M., Brown, S., Hinton, E. C., Rogers, P. J. & Fay, S. H. 'Expected satiety' changes hunger and fullness in the inter-meal interval. Appetite **56**, 310–315 (2011).
10. Brunstrom, J. M. The control of meal size in human subjects: a role for expected satiety, expected satiation and premeal planning. Proc. Nutr. Soc. **70**, 155–161 (2011).
11. Hetherington, M. M. et al. Understanding the science of portion control and the art of downsizing. Proc. Nutr. Soc. **77**, 347–355 (2018).
12. Teo, P. S., van Dam, R. M., Whitton, C., Tan, L. W. L. & Forde, C. G. Consumption of Foods With Higher Energy Intake Rates is Associated With Greater Energy Intake, Adiposity, and Cardiovascular Risk Factors in Adults. J. Nutr. **151**, 370–378 (2021).
13. Chen, J. Food oral processing—A review. Food Hydrocoll. **23**, 1–25 (2009).
14. Lavergne, M. D. de, Velde, F. van de & Stieger, M. Bolus matters: the influence of food oral breakdown on dynamic texture perception. Food Funct. **8**, 464–480 (2017).
15. Lenfant, F., Loret, C., Pineau, N., Hartmann, C. & Martin, N. Perception of oral food breakdown. The concept of sensory trajectory. Appetite **52**, 659–667 (2009).
16. Bolhuis, D. P. & Forde, C. G. Application of food texture to moderate oral processing behaviors and energy intake. Trends Food Sci. Technol. **106**, 445–456 (2020).
17. Bolhuis, D. P. et al. Slow Food: Sustained Impact of Harder Foods on the Reduction in Energy Intake over the Course of the Day. PLOS ONE **9**, e93370 (2014).
18. Forde, C. G., van Kuijk, N., Thaler, T., de Graaf, C. & Martin, N. Oral processing characteristics of solid savoury meal components, and relationship with food composition, sensory attributes and expected satiation. Appetite **60**, 208–219 (2013).
19. Bray, G. A. Medical consequences of obesity. J. Clin. Endocrinol. Metab. **89**, 2583–2589 (2004).

20. Levesque, R. J. R. Obesity and Overweight. in Encyclopedia of Adolescence 1913–1915 (Springer, New York, NY, 2011). doi:10.1007/978-1-4419-1695-2_447.
21. Industry must cut calories in savoury food products by 20%, says Public Health England | The BMJ. https://www.bmj.com/content/360/bmj.k1067.
22. Hogenkamp, P. S. & Schiöth, H. B. Effect of oral processing behaviour on food intake and satiety. Trends Food Sci. Technol. **34**, 67–75 (2013).
23. Cunningham, P. M. & Rolls, B. J. The Satiation Framework: Exploring processes that contribute to satiation. Physiol. Behav. **236**, 113419 (2021).
24. Forde, C. G., Leong, C., Chia-Ming, E. & McCrickerd, K. Fast or slow-foods? Describing natural variations in oral processing characteristics across a wide range of Asian foods. Food Funct. **8**, 595–606 (2017).
25. Miquel-Kergoat, S., Azais-Braesco, V., Burton-Freeman, B. & Hetherington, M. M. Effects of chewing on appetite, food intake and gut hormones: A systematic review and meta-analysis. Physiol. Behav. **151**, 88–96 (2015).
26. Robinson, E. et al. A systematic review and meta-analysis examining the effect of eating rate on energy intake and hunger. Am. J. Clin. Nutr. **100**, 123–151 (2014).
27. Bolhuis, D. P. & Forde, C. G. Application of food texture to moderate oral processing behaviors and energy intake. Trends Food Sci. Technol. **106**, 445–456 (2020).
28. Fogel, A. et al. Faster eating rates are associated with higher energy intakes during an ad libitum meal, higher BMI and greater adiposity among 4·5-year-old children: results from the Growing Up in Singapore Towards Healthy Outcomes (GUSTO) cohort. Br. J. Nutr. **117**, 1042–1051 (2017).
29. McCrickerd, K. & Forde, C. G. Sensory influences on food intake control: moving beyond palatability. Obes. Rev. Off. J. Int. Assoc. Study Obes. **17**, 18–29 (2016).
30. Technology to Automatically Record Eating Behavior in Real Life: A Systematic Review. https://www.mdpi.com/1424-8220/23/18/7757.
31. Diamantidou, E., Giakoumis, D., Votis, K., Tzovaras, D. & Likothanassis, S. Comparing Deep Learning and Human Crafted Features for Recognising Hand Activities of Daily Living from Wearables. in 2022 23rd IEEE International Conference on Mobile Data Management (MDM) 381–384 (2022). doi:10.1109/MDM55031.2022.00085.
32. Kyritsis, K., Diou, C. & Delopoulos, A. A Data Driven End-to-End Approach for In-the-Wild Monitoring of Eating Behavior Using Smartwatches. IEEE J. Biomed. Health Inform. **25**, 22–34 (2021).
33. Kyritsis, K., Tatli, C. L., Diou, C. & Delopoulos, A. Automated analysis of in meal eating behavior using a commercial wristband IMU sensor. in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2843–2846 (2017). doi:10.1109/EMBC.2017.8037449.
34. Mirtchouk, M. et al. Recognizing Eating from Body-Worn Sensors: Combining Free-living and Laboratory Data. Proc ACM Interact Mob Wearable Ubiquitous Technol **1**, 85:1-85:20 (2017).
35. Farooq, M. et al. Validation of Sensor-Based Food Intake Detection by Multicamera Video Observation in an Unconstrained Environment. Nutrients **11**, 609 (2019).
36. Chun, K. S., Bhattacharya, S. & Thomaz, E. Detecting Eating Episodes by Tracking Jawbone Movements with a Non-Contact Wearable Sensor. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**, 4:1-4:21 (2018).
37. Zhou, B. et al. Smart table surface: A novel approach to pervasive dining monitoring. in 2015 IEEE International Conference on Pervasive Computing and Communications (PerCom) 155–162 (2015). doi:10.1109/PERCOM.2015.7146522.

38. Lasschuijt, M. et al. Concept Development and Use of an Automated Food Intake and Eating Behavior Assessment Method. J. Vis. Exp. (2021) doi:10.3791/62144.
39. Zhang, S., Xu, Q., Sen, S. & Alshurafa, N. VibroScale: turning your smartphone into a weighing scale. in Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers 176–179 (Association for Computing Machinery, New York, NY, USA, 2020). doi:10.1145/3410530.3414397.
40. Gao, J., Tan, W., Ma, L., Wang, Y. & Tang, W. MUSEFood: Multi-Sensor-Based Food Volume Estimation on Smartphones. in 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) 899–906 (2019). doi:10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00182.
41. Makhsous, S., Mohammad, H. M., Schenk, J. M., Mamishev, A. V. & Kristal, A. R. A Novel Mobile Structured Light System in Food 3D Reconstruction and Volume Estimation. Sensors **19**, 564 (2019).
42. Hamatani, T., Elhamshary, M., Uchiyama, A. & Higashino, T. FluidMeter: Gauging the Human Daily Fluid Intake Using Smartwatches. Proc ACM Interact Mob Wearable Ubiquitous Technol **2**, 113:1-113:25 (2018).
43. Kreutzer, J. F., Deist, J., Hein, C. M. & Lueth, T. C. Sensor systems for monitoring fluid intake indirectly and directly. in 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN) 1–6 (2016). doi:10.1109/BSN.2016.7516223.
44. SmrtFridge | Proceedings of the 17th Conference on Embedded Networked Sensor Systems. https://dl.acm.org/doi/10.1145/3356250.3360028.
45. Chiu, M.-C. et al. Playful bottle: a mobile social persuasion system to motivate healthy water intake. in Proceedings of the 11th international conference on Ubiquitous computing 185–194 (Association for Computing Machinery, New York, NY, USA, 2009). doi:10.1145/1620545.1620574.
46. Mertes, G. et al. Measuring and Localizing Individual Bites Using a Sensor Augmented Plate During Unrestricted Eating for the Aging Population. IEEE J. Biomed. Health Inform. **24**, 1509–1518 (2020).
47. Papapanagiotou, V., Ganotakis, S. & Delopoulos, A. Bite-Weight Estimation Using Commercial Ear Buds. in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 7182–7185 (2021). doi:10.1109/EMBC46164.2021.9630500.
48. FitByte: Automatic Diet Monitoring in Unconstrained Situations Using Multimodal Sensing on Eyeglasses | Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://dl.acm.org/doi/10.1145/3313831.3376869.
49. Farooq, M., Fontana, J. M. & Sazonov, E. A novel approach for food intake detection using electroglottography. Physiol. Meas. **35**, 739 (2014).
50. Zhang, R. & Amft, O. Monitoring Chewing and Eating in Free-Living Using Smart Eyeglasses. IEEE J. Biomed. Health Inform. **22**, 23–32 (2018).
51. Adachi, K. & Yanai, K. DepthGrillCam: A Mobile Application for Real-time Eating Action Recording Using RGB-D Images. in Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management 55–59 (Association for Computing Machinery, New York, NY, USA, 2022). doi:10.1145/3552484.3555752.

52. Qiu, J., Lo, F. P.-W. & Lo, B. Assessing Individual Dietary Intake in Food Sharing Scenarios with a 360 Camera and Deep Learning. in 2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN) 1–4 (2019). doi:10.1109/BSN.2019.8771095.
53. van Bommel, R., Stieger, M., Visalli, M., de Wijk, R. & Jager, G. Does the face show what the mind tells? A comparison between dynamic emotions obtained from facial expressions and Temporal Dominance of Emotions (TDE). Food Qual. Prefer. **85**, 103976 (2020).
54. Chen, X. & Kamavuako, E. N. Vision-Based Methods for Food and Fluid Intake Monitoring: A Literature Review. Sensors **23**, 6137 (2023).
55. Yarlagadda, S. K. et al. Learning eating environments through scene clustering. Preprint at https://doi.org/10.48550/arXiv.1910.11367 (2019).
56. Wang, W. et al. A review on vision-based analysis for automatic dietary assessment. Trends Food Sci. Technol. **122**, 223–237 (2022).
57. A Review Paper on Computer Vision and Image Processing - R Discovery. https://discovery.researcher.life/article/a-review-paper-on-computer-vision-and-image-processing/0e85f941965b354d9a7743a035685f8d.
58. Shreya M. Shelke et al. A Review Paper on Computer Vision. Int. J. Adv. Res. Sci. Commun. Technol. 673–677 (2023) doi:10.48175/IJARSCT-8901.
59. 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. Pers. Ubiquitous Comput. **16**, 729–739 (2012).
60. Chai, J., Zeng, H., Li, A. & Ngai, E. W. T. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Mach. Learn. Appl. **6**, 100134 (2021).
61. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2017).
62. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1409.1556 (2015).
63. Szegedy, C. et al. Going Deeper with Convolutions. Preprint at https://doi.org/10.48550/arXiv.1409.4842 (2014).
64. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1512.03385 (2015).
65. Huang, G., Liu, Z., Maaten, L. van der & Weinberger, K. Q. Densely Connected Convolutional Networks. Preprint at https://doi.org/10.48550/arXiv.1608.06993 (2018).
66. Howard, A. G. et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Preprint at https://doi.org/10.48550/arXiv.1704.04861 (2017).
67. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
68. Vaswani, A. et al. Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).

2

Data-Driven Nutritional Assessment of Urban Food Landscapes: Insights from Boston, London, and Dubai

Michele Tufano[1], Fábio Duarte[2*], Martina Mazzarello[2], Javad Eshtiyagh[2], Carlo Ratti[2,3], Guido Camps[1]

[1] Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands
[2] Senseable City Lab, Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] ABC Department, Politecnico di Milano, Milan, Italy
Published in Scientific Reports (2025)
https://doi.org/10.1038/s41598-025-08098-9

*

ABSTRACT

Urban food landscapes significantly influence dietary habits and health outcomes, with disparities in food access contributing to obesity, particularly in socioeconomically disadvantaged neighborhoods. This study presents a data-driven approach to assess urban food landscapes using restaurant menu data from online delivery platforms in Boston, London, and Dubai. Machine learning matched menu items to the U.S. FoodData Central database, enabling the calculation of nutritional indices and neighborhood-level nutrient averages.

The analysis revealed significant patterns between urban food landscapes, socioeconomic features, and obesity rates. In London and Boston, higher socioeconomic neighborhoods had better access to nutrient-rich foods, with dietary fibers showing a strong inverse association with obesity ($p = 0.001$, $p = 0.004$, respectively). In Dubai, due to limited health data, the analysis focused on food landscapes and rental prices as a proxy of a neighborhood's socioeconomic profile.

This method offers a scalable alternative to traditional food environment studies and can guide policymakers in identifying neighborhoods at risk for obesity and lack of nutritious foods. Future research should extend this method to diverse regions and advocate for standardized, open-access nutritional data to implement targeted and evidence-based nutritional interventions.

MAIN

Urban food landscapes—defined by the availability and accessibility of food options—play a pivotal role in shaping individuals' dietary habits and consequent health outcomes. Obesogenic environments, characterized by an abundance of fast-food outlets, aggressive marketing strategies, and affordable processed foods, contribute to rising obesity rates, disproportionately affecting low-income and minority communities [1–5]. Neighborhoods with lower median home values and high deprivation are characterized by a higher density of fast-food outlets [6–9]. In contrast, higher-income groups benefit from more healthy food environments with greater access to nutritious foods and more information on healthy eating [10,11]. Notably, the proximity to fast-food outlets is a risk factor for obesity, diabetes, cardiovascular and coronary heart diseases [12–19].

Assessing the nutritional health of food landscapes is a central issue in public health, which requires comprehensive data on food availability and quality. Previous research on food landscapes has primarily relied on observational methods such as diary logs and records to gather data on food availability, quality, and pricing. However, this approach is labor- and time-intensive. The logs and questionnaire require multiple audits of the food landscape to represent a typical condition, and adapting to different food settings is challenging and limited to food regionality [20,21].

Previous studies on food landscapes have primarily relied on Geographic Information Systems (GIS) network analysis, which, while valuable for modeling spatial accessibility, often relies on outdated business data and fails to capture nutritional content [22].

Here, we present data-driven and rapid nutritional quantification of the food landscape. Rather than classifying restaurants in a few typologies, we collect data from all restaurants' menus available on food delivery platforms during the summer of 2023. Our study utilizes machine learning techniques to match menu items with entries in the FoodData Central database, enabling the assignment of their corresponding nutritional values [23]. FoodData Central, the USDA's comprehensive database containing standardized nutritional information for over 375,000 food items, serves as our reference database for matching menu items with their corresponding nutritional values. We then calculate average nutrient levels, the Meal Balance Index (MBI)[24], and the Nutrient-Rich Food Index (NRF) [25,26] per neighborhood. The MBI is a composite score that assesses how well a meal meets daily nutritional recommendations across key nutrients to reflect overall balance, while the NRF score quantifies nutrient density by positively

weighting beneficial nutrients and negatively weighting those to limit, thereby providing a single measure of nutritional quality. A schematic overview of the main workflow is presented in Figure 1.

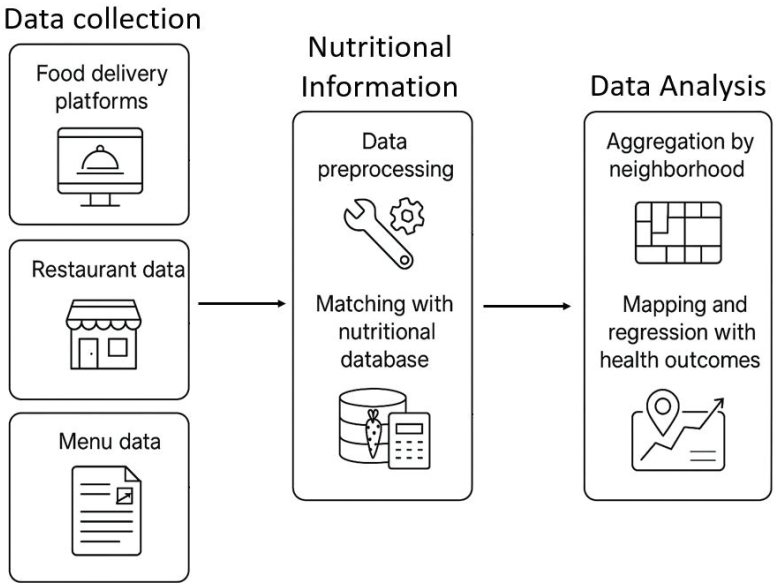


Figure 1. Workflow for the collection, processing, and analysis of food delivery data for nutritional and geographic assessment.
The left column illustrates data collection from food delivery platforms, including restaurant and menu information. The middle column represents data preprocessing and enrichment through matching menu items with a nutritional database to retrieve nutrient values and calculate indices (Meal Balance Index and Nutrient-Rich Foods Index). The right column shows the aggregation of nutritional data by neighborhood, followed by spatial mapping and multivariate regression analyses to investigate associations with obesity and socioeconomic features.

Our method was applied to three cities across different continents—Boston, London, and Dubai—each offering unique characteristics to contribute to a comprehensive analysis of urban food landscapes.

Boston, London, and Dubai represent distinct urban contexts for examining food landscapes: Boston, as an American city, benefits from comprehensive food data within the FoodData Central database, providing a reliable benchmark for nutritional assessment; London, as a major Western city, reflects global dietary trends through its diverse and accessible food environment; and Dubai, with its non-Western cuisine and significant obesity burden[27] , highlights the dietary challenges of a rapidly developing urban center. These cities were selected due to the availability of robust data for analysis. By comparing these diverse cities,

our study provides valuable insights into the applicability of our food landscape assessment method across varying urban environments and socio-economic contexts, while also exploring its relationship with neighborhood characteristics and obesity rates.

In contrast, in Boston and London, the study explores the relationships between the food landscape, obesity rates, and neighborhood socioeconomic features; whereas in Dubai, the analysis focuses on quantifying the nutritional composition of the urban food landscape and its correlation with rental housing prices as a proxy of each neighborhood's socioeconomic profile.

RESULTS

This study analyzed restaurant menu information from three cities to quantify the urban food landscapes and their association with obesity rates and neighborhood socioeconomic features. Data were collected from restaurants with a physical presence in each city through online food delivery platforms, encompassing thousands of restaurants and millions of menu items: London (18,002 restaurants and 3,108,134 menu items), followed by Dubai (9,288; 1,556,243, respectively), and Boston (2,375; 222,964). After filtering the menus to discard drink-, tobacco-, portion-related-, and sauce-related items, we calculated the nutritional indexes of MBI and NRF as indicative of the nutritional content of each food item. The nutrients were then aggregated per restaurant and neighborhood. For each neighborhood, we included socioeconomic features, such as housing prices and income. In London and Boston, we investigated the correlation between the food landscape and health outcomes. In Dubai, we limited the correlation between the food landscape and neighborhood socioeconomic features, due to the lack of health outcomes data. We begin by presenting the food landscapes of each city, followed by an analysis of the relationships between these food landscapes, obesity rates, and socioeconomic characteristics.

FOOD LANDSCAPES

The number of restaurants and foods with nutrients differed per city (Table 1). Foods from restaurant menus achieved the highest coverage in Boston, where 71.1% of the foods were found in the U.S.-based database. London and Dubai obtained lower database coverage with 56.1% and 41.7%, respectively. When a food item from a restaurant menu matched an entry in the FoodData Central database, its nutrient information was retrieved and assigned. This process

successfully assigned nutrients in approximately 98% of cases. To investigate the inequalities of nutrients on a city level, we aggregate nutrients data by restaurant and neighborhood. The food landscapes depicted below for London (Fig. 2), Boston (Fig. 3), and Dubai (Fig. 4) show different concentrations of nutrient-rich foods, dietary fibers, MBI, NRF, and potassium across the cities.

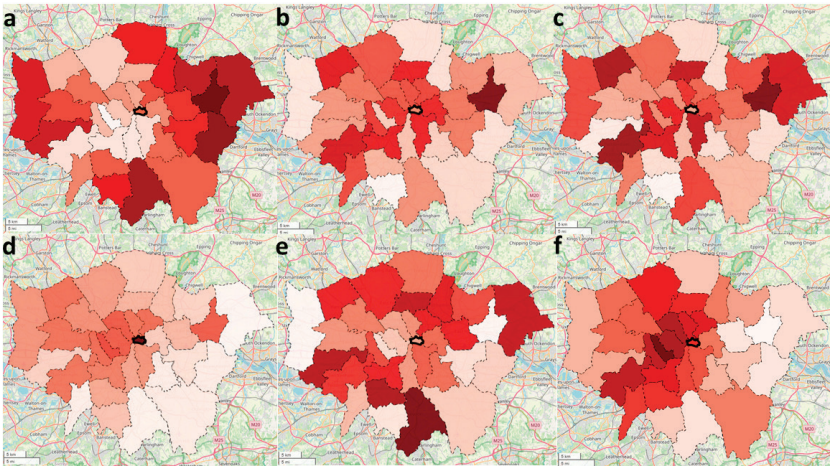


Figure 2 - London food landscape per neighborhoods, represented for a) Obesity prevalence in adults, b) Meal Balance Index, c) Nutrient-rich score, d) Dietary fibers, e) Potassium, and f) Housing prices. Areas with darker red gradients correspond to higher values. The City of London is marked in black at the center of the map. Implemented using Python (v3.10., https://www.python.org/downloads/release/python-3100/), geopandas (v0.8.1, https://geopandas.org/en/v0.8.1/reference.html), and Folium (v0.11.0, https://python-visualization.github.io/folium/version-v0.11.0/index.html).

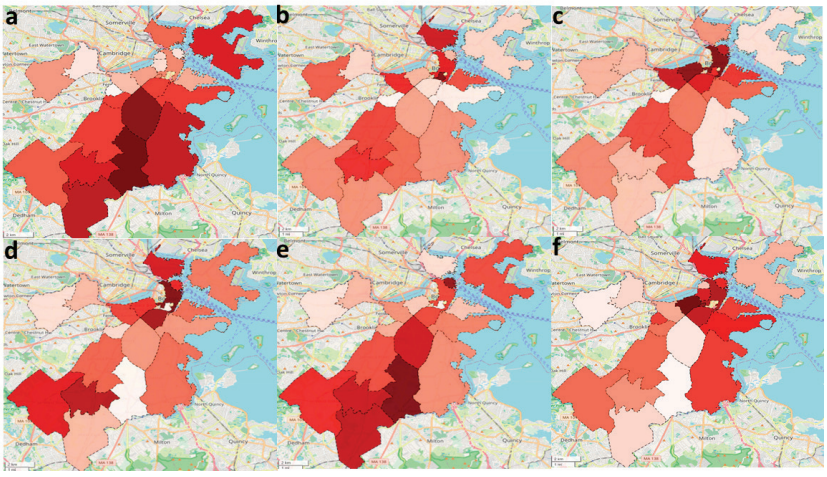


Figure 3 - Boston food landscape per neighborhoods, represented for a) Obesity prevalence in adults, b) Meal Balance Index, c) Nutrient-rich score, d) Dietary fibers, e) Potas-

sium, and f) Housing prices. Areas with darker red gradients correspond to higher values. Implemented using Python (v3.10., https://www.python.org/downloads/release/python-3100/), geopandas (v0.8.1, https://geopandas.org/en/v0.8.1/reference.html), and Folium (v0.11.0, https://python-visualization.github.io/folium/version-v0.11.0/index.html).



Figure 4 - Dubai food landscape per neighborhoods, represented for a) Meal Balance Index, b) Nutrient-rich score, c) Dietary fibers,
d) Potassium, and e) Rental price. Areas with darker red gradients correspond to higher values. Implemented using Python (v3.10., https://www.python.org/downloads/release/python-3100/), geopandas (v0.8.1, https://geopandas.org/en/v0.8.1/reference.html), and Folium (v0.11.0, https://python-visualization.github.io/folium/version-v0.11.0/index.html).

These indices were selected for their relevance in assessing overall meal quality, nutrient density, and potential health impacts at the community level. A detailed description of the nutritional values per neighborhood is provided in the Supplementary section online.

Table 1 – Overview of restaurant and menu data collected from online delivery platforms in London, Dubai, and Boston, including total items, filtered items, and the percentage of foods successfully matched to the FoodData Central database. Nutrient assignments are also reported as a percentage of foods matched to the database.

City	Restaurants, N	Total items, N	Restaurants after filtering, N	Foods after filtering, N (unique), %	Foods found in the database, N (unique), %	Foods with nutrients, N (unique), %
London	20,304	3,359,061	18,002	3,108,134 (485,016), 92.5%	1,746,424 (64,572), 56.1%	1,702,479 (62,686), 97.5%

Table 1 – Overview of restaurant and menu data collected from online delivery platforms in London, Dubai, and Boston, including total items, filtered items, and the percentage of foods successfully matched to the FoodData Central database. Nutrient assignments are also reported as a percentage of foods matched to the database. (*continued*)

City	Restaurants, N	Total items, N	Restaurants after filtering, N	Foods after filtering, N (unique), %	Foods found in the database, N (unique), %	Foods with nutrients, N (unique), %
Dubai	9,529	1,775,769	9,288	1,556,243 (245,879) , 87,6%	649,936 (43,142), 41.7%	640,551 (41,979), 98.5%
Boston	2,857	264,269	2,375	222,964 (100,796), 84.4%	158,295 (26,869) 71.1%	155,813 (26,150), 98.4%

Table 2 – Regression Coefficients, Standard Errors, t-statistics, p-values, and 95% Confidence Intervals for Predictors of Obesity in Adults in London

Variable	Coefficient	Standard Error	t-statistic	p-value	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)	VIF
Intercept	19.7112	0.609	32.345	<0.001	18.461	20.962	1
Meal Balance Index	0.5822	1.118	0.521	0.607	-1.713	2.877	3.368
Nutrient-Rich Score	-0.1999	1.065	-0.188	0.853	-2.385	1.986	3.055
Dietary Fibers	-2.5392	0.707	-3.59	0.001	-3.99	-1.088	1.347
Potassium	-1.9207	0.642	-2.99	0.006	-3.239	-0.603	1.111

Table 3 – Regression Coefficients, Standard Errors, t-statistics, p-values, and 95% Confidence Intervals for Predictors of Obesity in Adults in Boston

Variable	Coefficient	Standard Error	t-statistic	p-value	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)	VIF
Intercept	21.9992	0.741	29.697	<0.001	20.443	23.556	1
Meal Balance Index	0.3811	0.916	0.416	0.682	-1.543	2.305	1.528
Nutrient-Rich Score	-1.3804	0.743	-1.858	0.08	-2.941	0.18	1.006
Dietary Fibers	-3.046	0.908	-3.355	0.004	-4.953	-1.139	1.502
Potassium	2.7189	0.754	3.604	0.002	1.134	4.304	1.037

Food landscapes and Obesity

We investigated the relationship between the food landscape and obesity in cities with available health data, specifically London (Table 2) and Boston (Table 3). Multivariate regression models revealed varying degrees of predictive power for obesity rates in both cities. Both models exhibited statistically significant relationships between predictor variables and obesity. However, the explained variance differed, with R-squared of 0.4 for London and 0.621 for Boston, suggesting a stronger predictive capability in the latter.

Notably, dietary fibers consistently showed a significant negative association with obesity in both cities (p-value: 0.001 for London; 0.004 for Boston). Potassium had differing effects in London and Boston. No significant association was found between obesity and MBI nor NRF in either city.

FOOD LANDSCAPES AND NEIGHBORHOOD SOCIOECONOMIC FEATURES

This section explores the relationship between neighborhood socioeconomic characteristics and the local food landscape.

The analysis considers key socioeconomic metrics, including median housing prices in London, median household income in Boston, and median rental prices in Dubai. Multivariate regression models are used to assess the associations between food-related variables and these socioeconomic factors. In London, the multivariate regression model demonstrated a strong predictive power for housing prices (R-squared = 0.703). Both dietary fiber intake and the daily consumption of five portions of fruits and vegetables were significantly and positively associated with higher housing prices ($p < 0.001$ for both). NRF showed no significant association with housing prices (Table 4).

Table 4 – Regression results for predictors of housing prices in London: coefficients, standard errors, t-statistics, p-values, 95% confidence intervals, and variance inflation factors

Variable	Coefficient	Standard Error	t-statistic	p-value	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)	VIF
Intercept	569.0893	18.141	31.37	<0.001	531.866	606.312	1

Table 4 – Regression results for predictors of housing prices in London: coefficients, standard errors, t-statistics, p-values, 95% confidence intervals, and variance inflation factors (continued)

Variable	Coefficient	Standard Error	t-statistic	p-value	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)	VIF
Eat 5-a-day Fruit/ Vegetable - Adults (%)	84.6118	18.753	4.512	<0.001	46.134	123.089	1.069
Meal Balance Index	-39.3576	32.98	-1.193	0.243	-107.027	28.312	3.305
Nutrient-Rich Score	3.7171	31.541	0.118	0.907	-60.999	68.433	3.023
Dietary Fibers	124.2368	20.692	6.004	<0.001	81.78	166.694	1.301

In Boston, the multivariate regression model revealed robust predictive power for home values (R-squared = 0.725). Both NRF and dietary fiber intake were significantly and positively associated with higher home values ($p<0.001$ for both). In contrast, the MBI was significantly and negatively associated with home values ($p=0.021$) (Table 5).

Table 5 –Regression results for predictors of housing prices in Boston: coefficients, standard errors, t-statistics, p-values, 95% confidence intervals, and variance inflation factors

Variable	Coefficient	Standard Error	t-statistic	p-value	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)	VIF
Intercept	779700	24600	31.695	<0.001	728000	831000	1
Meal Balance Index	-75170	29900	-2.513	0.021	-138000	-12600	1.478
Nutrient-Rich Score	144100	24600	5.849	<0.001	92500	196000	1.003
Dietary Fibers	124000	29900	4.15	0.001	61400	187000	1.475

In Dubai, the regression model accounted for only a small proportion of the variance in rental prices (R-squared = 0.046). Importantly, none of the dietary variables exhibited a statistically significant relationship with rental prices at the 0.05 level (Table 6).

Table 6 – Regression results for predictors of rental prices in Dubai: coefficients, standard errors, t-statistics, p-values, 95% confidence intervals, and variance inflation factors

Variable	Coefficient	Standard Error	t-statistic	p-value	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)	VIF
Intercept	71.3771	2.644	26.993	<0.001	66.145	76.609	1
Meal Balance Index	-4.2895	3.34	-1.284	0.201	-10.899	2.32	1.596
Nutrient-Rich Score	-1.3416	3.4	-0.395	0.694	-8.069	5.386	1.653
Dietary Fibers	-3.4064	2.707	-1.259	0.21	-8.762	1.949	1.048

OBESITY AND NEIGHBORHOOD SOCIOECONOMIC FEATURES

Lastly, we examine the associations between obesity rates and neighborhood socioeconomic features. Regression models for London and Boston were used to explore the correlation between obesity rates and key neighborhood socioeconomic factors, such as housing prices and income levels. In London, the regression model revealed significant negative associations between obesity rates and weekly income (Table 7), indicating that higher economic status is linked to lower obesity prevalence.

Table 7 – Regression results for neighborhood socioeconomic predictors of obesity in London: coefficients, standard errors, t-statistics, p-values, 95% confidence intervals, and variance inflation factors

Variable	Coefficient	Standard Error	t-statistic	p-value	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)	VIF
Intercept	19.7112	0.506	38.96	<0.001	18.676	20.746	1
Housing price	-2.2092	0.593	-3.727	0.001	-3.421	-0.997	1.372
Income (weekly)	-1.2446	0.593	-2.1	0.045	-2.457	-0.032	1.372

Similarly, in Boston, the regression results show limited ability to explain variability in obesity rates (R-squared = 0.223), with neither income (p=0.233) nor home values (p=0.223) showing statistically significant relationships with obesity (Table 8). The adjusted R-squared value of 0.145 suggests that a substantial

portion of the variation in obesity rates is not explained by the included predictors.

Table 8 – Regression results for neighborhood socioeconomic predictors of obesity in Boston: coefficients, standard errors, t-statistics, p-values, 95% confidence intervals, and variance inflation factors

Variable	Coef-ficient	Standard Error	t-statistic	p-value	95% Con-fidence Interval (Lower)	95% Con-fidence Interval (Upper)	VIF
Intercept	21.9992	1.007	21.848	<0.001	19.899	24.1	1
Income	-1.3959	1.134	-1.231	0.233	-3.762	0.97	1.269
Housing price	-1.4267	1.134	-1.258	0.223	-3.793	0.939	1.269

DISCUSSION

This study presents a quantitative nutritional analysis of urban food landscapes using menu data from restaurants available on online delivery platforms. Restaurant menus from Boston, London, and Dubai were matched to the FoodData Central database to retrieve food nutrients for each menu item, which were then aggregated per area and associated with obesity prevalence and neighborhood socioeconomic features.

Our findings demonstrate that leveraging online delivery platforms and AI-based nutritional analysis provides an efficient and scalable method for assessing urban food landscapes with potential global applications. This approach proved particularly effective in Boston and is likely applicable to other American cities, where a high match rate (71%) between restaurant foods and the standardized FoodData Central database was observed. This enabled the estimation of nutrient content and its assignment to online food items. Consequently, this method facilitated the visualization of food landscapes and the identification of neighborhoods or areas where access to healthy food may be limited, such as food deserts.

Nutritional data for each menu item were aggregated at the restaurant and neighborhood (or borough) levels to analyze the food landscape and examine variations across different areas within a city. Interestingly, dietary fibers from the food landscape were significantly associated with obesity rates and neighborhood socioeconomic features. Our findings highlight dietary fiber as a key nutrient consistently associated with lower obesity rates across both cities. In

London, dietary fibers demonstrated a significant negative relationship with obesity ($p = 0.001$), and a similar association was observed in Boston ($p = 0.004$). This suggests that low concentrations of dietary fibers in the food landscape may serve as an indicator of higher obesity prevalence, consistent with epidemiological studies showing an inverse relationship between dietary fibers intake and body weight[28]. Neither the Meal Balance Index (MBI) nor the Nutrient-Rich Food (NRF) score showed significant associations with obesity in either location. This absence of correlation may reflect limitations in aggregate nutritional indices to capture the complex relationship between food landscapes and health outcomes.

Furthermore, the results highlight distinct patterns in the association between food landscape variables and neighborhood socioeconomic features across the three cities. Due to data availability, we employed diverse neighborhood socioeconomic features, such as median housing prices in London, median housing prices and median household income in Boston, and median rental housing prices in Dubai.

In London, the regression model demonstrated strong predictive power for housing prices (R-squared = 0.703), with significant associations observed for dietary fibers and fruit/vegetable consumption ($p < 0.001$ for both). Similarly, the model for Boston demonstrated strong predictive power for home values (R-squared = 0.725). Dietary fibers were significantly and positively associated with higher home values ($p<0.001$). Surprisingly, the MBI was significantly negatively associated with home values ($p=0.021$). This indicates that neighborhoods with higher socioeconomic conditions have better access to nutritious foods. These findings align with prior research indicating that higher-income neighborhoods generally have greater and more diverse access to nutritious foods [10,11], reflecting socioeconomic disparities in food availability. In contrast, the model for Dubai revealed a very low explanatory power for rental prices (R-squared = 0.046), with no significant associations observed for any of the dietary variables.

These findings suggest that data interpretability and local context will influence the ability to quantify the relationship between food landscapes and socioeconomic factors across cities. Strong associations in London and Boston indicate that dietary fiber and fruit/vegetable intake reflect socioeconomic disparities in food access. However, the weaker results observed in Dubai underscore the necessity for further research in non-Western urban contexts, particularly for cuisines or dishes that are not represented in the FoodData Central database. Furthermore, alternative sources of healthy food access, such as supermarket

distribution patterns, may also contribute to the observed differences and require additional investigation.

The neighborhood socioeconomic features also had a relationship with obesity rates, regardless of the food landscape. In London, the regression model demonstrated significant negative relationships between obesity and both housing prices and weekly income. In Boston, the regression model showed low predicting power (R-squared = 0.223) and two non-significant negative relationships between obesity and both housing prices (p=0.223) and income (p=0.233). These results suggest that neighborhoods with higher socioeconomic status, as reflected by higher housing prices and income levels, are associated with lower obesity rates, and vice versa. Although the strong statistical significance of these associations in London (p = 0.001 for housing prices and p = 0.045 for income) underscores the link between neighborhood socioeconomic features and obesity, the Boston results suggest that these factors alone may not strongly predict obesity in this context, highlighting the need for further research.

These findings are consistent with previous studies indicating that higher income is associated with a reduced risk of obesity [29–31]. In contrast, residing in economically disadvantaged neighborhoods has been linked to a higher risk of obesity in both adults and adolescents [32,33]. In Dubai, the absence of obesity data prevented the analysis from being conducted.

Several limitations should be noted. A major limitation of this study is the varying coverage of menu items matched to the U.S. Department of Agriculture's FoodData Central across the three cities. The limited database coverage in Dubai highlights the difficulty of applying the U.S.-based food composition data to culturally different cuisines, despite the availability of Gulf-specific dietary guidelines and food composition tables [34,35]. This limitation affects the generalizability of our approach, which depends on foods (1) listed on delivery platforms and (2) their compatibility with the U.S. nutrient database. The lower coverage in non-U.S. cities underscores the need for open-access, region-specific food composition databases to ensure comprehensive analysis—a limitation that contributes to the literature, highlighting bias in data availability. The lower database coverage in London (56%) and Dubai (42%) can be attributed to two main factors: differences in food terminology (e.g., "French fries" in the American database versus "chips" in British menus) and the absence of certain foods (e.g., common Arabic dishes are not included in the FoodUSDA database). This presents a limitation of the study, as not all cities and food items had nutritional values assigned consistently. While Natural Language Processing, Generative AI, or using a dedicated food ontology (e.g. FoodOn) could improve the seman-

tic matching between menu items and food database entries, the lack of nutritional information for underrepresented foods remains a challenge. Therefore, region-specific nutritional datasets are crucial for accurate food landscape assessment. Despite this limitation, the data-driven approach reaches a menu-item level of ingredient identification that far exceeds traditional studies that use coarser categorizations (fast vs traditional restaurants, e.g.). In the future, other nutritional databases can be explored to include better descriptions of cuisines in all countries and foods from all ethnicities. This would enable easier access to nutritional information and more comprehensive assessments of food environments using these and other methods.

Another limitation of this study is data mining on food delivery platforms. Our datasets included all the foods served by restaurants with a physical presence that could be found online. Notwithstanding our data mining efforts, access to data was limited to a subset of food delivery platforms. As previously advocated by other researchers, regulations on food delivery platforms are needed to ensure transparent nutrition information and accessibility [36,37].

The ecological, cross-sectional design of this study allows for the analysis of area-level patterns but does not account for individual-level variability. This limitation restricts the ability to infer causation or fully unravel complex associations, as city areas were used as the units of observation.

However, our approach enables rapid and automatic data collection of food availability, quality, and pricing in city restaurants. Unlike traditional food environment research, which is labor-intensive because it relies on observational scans and multiple audits [20–22], our method offers a more efficient alternative, although comparative validation remains limited. While traditional approaches such as in-person audits and food logs offer detailed individual dietary information, they are resource-intensive and geographically constrained. Our methodology enables comprehensive city-scale analysis with minimal manual intervention, though it primarily utilizes online menu data, it does not represent food consumption, and it might exclude individuals who cannot afford to purchase food regularly or who do not use online platform regularly. Future research should prioritize methodological validation through systematic comparisons with established techniques in selected urban regions, such as stratified random sampling, to evaluate consistency, identify discrepancies, and enhance accuracy. Our approach has potential applications in food policy by providing a quick and inexpensive way to identify areas for further study in urban health policy. The integration of ground-truthing protocols or hybrid methodologies

combining digital and traditional data sources represents a promising approach to optimize reliability in urban nutritional landscape evaluation.

This study raises many opportunities for future research on the food environment using data-driven approaches, and for monitoring policy of obesity in the urban context. For example, the Al Ttay area, Dubai, showed the lowest average dietary fiber. Based on the correlation between nutrition data and obesity, our results indicate that this area might be at risk for obesity and lack of nutritious food—which could be used by Dubai to promote healthy food options to this community. Similarly, other cities may employ our data-driven approach to identify areas that may lack nutritious foods or that may be at increased risk for obesity and then intervene to counteract the obesogenic environment.

Restaurants were prioritized in our analysis, as they represent the most prevalent category on online delivery platforms, while grocery retail availability exhibits significant regional and platform-specific variation. Yet, nutritional information gathered with big data has already shown its potential for supermarket websites[38] and if coupled with the food landscape could represent a crucial innovation for monitoring nutrition and public health. To achieve this, we advocate for an increased number of open-source food databases per country, as well as more standardized food names on delivery platforms and across countries, ideally including nutritional information. This would enable our method to be applied in non-English-speaking cities and support global scalability. Menus collected from a new city could first be matched with the local food database, and any unmatched items could then be translated and queried against FoodData Central or other international databases.

In conclusion, our data-driven algorithm is specifically useful for assessing food landscapes based on restaurant data from online delivery platforms and food items identifiable in the US food system database (FoodData Central). Future research is needed to evaluate the algorithm's applicability in other contexts and with alternative food databases to extend its utility beyond the USA. Attention should be paid to neighborhoods that report a low average of fibers. This data-driven approach can serve as a valuable tool for policy makers to rapidly and automatically assess the urban food landscapes in a city and implement targeted interventions in neighborhoods at risk for lack of inadequate access to nutritious food and obesogenic environments.

ONLINE METHODS

Data collection, processing and storage

Starting at the food delivery platform's website, the postcodes and addresses were used as delivery proxies including all neighborhoods for the analyzed cities. Boston, London, and Dubai were selected from a list of cities with attainable datasets on neighborhood socioeconomic features, health, and food menu items as representatives of three geographic regions with different dietary habits. The search was restricted to restaurant categories, therefore excluding grocery stores and other store types (e.g., liquor, pet, or flower stores). This focus was chosen because restaurants are the most prevalent category on the online delivery platforms used as our data source, whereas grocery stores are not consistently available across all cities and platforms. The time of delivery was changed to 'anytime' to include restaurants that did not deliver at the time of the analysis. The URL of each restaurant was retrieved and subsequently opened to fetch (where available) the name, address, latitude, longitude, postcode, city, geohash, cuisine tags, rating, and number of reviews. We assigned an identification number to each restaurant to link it with its menu information, which was stored separately. During the summer of 2023, we retrieved online menu information including (where available) the name of the menu item, description, price, image, and currency. Unique internal identifiers were used to identify individual instances of each menu item [39]. After removing duplicate restaurants, the data was aggregated per city and stored in a database using the open-source software for relational database management SQLite V.3.40 [40]. All collection and processing code was written in Python V.3.10, using open-source Python Libraries request, BeautifulSoup, and selenium. These libraries were employed to send requests to website servers and manage HTML files or dynamically generated web pages of restaurants. The data collection code included built-in pauses to prevent overloading the website's servers. Multithreading was implemented to enhance efficiency.

Nutritional Information

Data preprocessing was implemented on the dataset containing the dishes. Firstly, all food names and descriptions were standardized to lowercase, removing special characters, redundant single characters, and extra spaces. Drink-, size-, tobacco-, and sauce-related entries were removed using extensive lists of keywords to ensure a focus on solid food items. To reduce the frequency of database queries, a dictionary was created to map the food items to restaurant ID numbers. The dictionary was sorted by the number of unique restaurant IDs. Further word replacements were used to standardize variations in food item naming. Certain words (e.g., specialty, combo) and characters (i.e., non-ASCII)

were excluded to enhance data quality. Regular expressions were used for text replacement. To match the dishes from restaurant menus and the database entries, we used sentence transformers and FAISS index. The dishes and database entries were encoded using sentence transformers (model BAAI/bge-base-en-v1.5) using NVIDIA RTX A5500 Laptop GPU on Windows 10 Pro (12th Gen Intel(R) Core (TM) i9-12950HX, 2.30 GHz). The embeddings were moved from GPU to CPU, converted to an array, normalized, and an index was generated using the FAISS library [41]. After the semantic search, we queried a SQL database to retrieve the nutritional values of the best matches. SQL queries were created to extract detailed nutrient information for each menu item about calcium, carbohydrate, energy, fiber, iron, magnesium, potassium, protein, saturated fatty acids, sodium, sugars, total fat, vitamins, and zinc.

The nutrients were retrieved from a database created locally from Food Data-Central tables[23] (Supplementary Fig. 1). The Meal Balance Index (MBI) and the Nutrient-Rich Foods (NRF) index were calculated based on previously established methods [24–26,] using functions that process nutritional values derived from databases.

The MBI and NRF scores quantify food nutritional balance and quality. These composite indices enhance our analysis by evaluating multidimensional nutritional properties across urban food landscapes, providing a more comprehensive assessment than single-nutrient measurements alone. The MBI was calculated to assess the nutritional balance of meals using a Python-based function. The MBI evaluates the proportion of key nutrients relative to the meal's total energy content and compares these proportions to established dietary recommendations. Each nutrient was normalized to a standard daily intake or energy content, scored against lower and upper intake limits, and assigned a weight based on its dietary significance. Nutrient scores were aggregated into an overall MBI by computing a weighted average. Higher MBI values indicate a more balanced meal composition, whereas lower scores suggest nutritional inadequacy or excesses.

The NRF was calculated to quantify the nutrient density of foods using a Python-based function. The NRF score incorporates nine encouraged nutrients (e.g., fiber, protein, calcium) and three limited nutrients (e.g., saturated fat, added sugars, sodium). Nutritional information for each food was normalized per 100 kcal to standardize comparisons. Each nutrient contribution was scaled relative to its daily value or maximum recommended value, with contributions capped at 100% to prevent overrepresentation. Positive contributions from encouraged nutrients were summed up, while negative contributions from limited nutrients

were subtracted. The resulting NRF score reflects the overall balance of nutrient density, with higher scores indicating healthier food choices. Finally, the dataset was enriched with the calculated MBIs and NRFs scores.

Data analysis

The exploratory analysis was conducted using pandas [42], by extracting nutritional information from the dataset. The outliers were removed using interquartile ranges. For each numeric column, a statistical summary was generated including mean, median, standard deviation, and skewness. Histograms, box plots, density plots, and Q-Q plots were generated to inspect the characteristics and distribution of the dataset. The most common foods per city were plotted using plotly[43]. ChatGPT model 4 (OpenAI) was used for refactoring code and rephrasing text.

The food landscape analysis was performed by aggregating the nutritional and index values for all dishes associated with each restaurant and computing the averages per restaurant and per neighborhood (or borough, when applicable).

To map the nutrient distributions across the cities, geographical data were used to create polygons representing neighborhoods or boroughs using GeoPandas[44] and Folium[45]. Missing restaurants' latitude and longitude values were retrieved from the address using the Google Maps Geocoding API. Latitude and longitude were matched against the polygon list to assign restaurants to a specific area. The nutritional content was aggregated per area to visually represent the spatial data. Each nutritional metric was treated as a separate layer in the map, with custom style functions applied to each layer to assign specific colors and labels using branca and geopandas. After including layer controls, the maps were saved to HTML format and subsequently converted to PNG format for final presentation. Multivariate regression analysis with obesity rates was performed in London and Boston. The London latest data release from 2022 was retrieved from the London datastore[46], and from the Office for Health Improvement and Disparities (https://fingertips.phe.org.uk/). The Boston data were retrieved from the PLACES project from the Center for Disease Control and Prevention (CDC), using the 2023 PLACES: Local Data for Better Health, Census Tract Data 2023 release [47]. The nutritional data aggregated per neighborhood, or borough, was normalized and analyzed for multivariate regression with prevalence of obesity or neighborhood socioeconomic features. The Boston data were retrieved from https://data.boston.gov/. The Boston housing values corresponding to 31st December 2023 were obtained from https://www.zillow.com/research/data/ . Due to absence of home price data, we collected rental prices in Dubai from www.bayut.com/ .

Data availability

The dataset associated with this manuscript, entitled "Restaurants with menus collected from online delivery Platforms in Boston, London, and Dubai", is available at https://doi.org/10.6084/m9.figshare.28302494.v1

Code availability

The code can be found at https://github.com/mtufano/foodpedia-public

REFERENCES

1. Egger, G. & Swinburn, B. An 'ecological' approach to the obesity pandemic. *BMJ* **315**, 477–480 (1997).
2. Cummins, S. & Macintyre, S. Food environments and obesity—neighbourhood or nation? *Int. J. Epidemiol.* **35**, 100–104 (2006).
3. Story, M., Kaphingst, K. M., Robinson-O'Brien, R. & Glanz, K. Creating Healthy Food and Eating Environments: Policy and Environmental Approaches. *Annu. Rev. Public Health* **29**, 253–272 (2008).
4. Pineda, E. & Mindell, J. Association of the food environment with obesity: a systematic review of geographical and statistical methods. *The Lancet* **388**, S89 (2016).
5. Jones-Smith, J. C. *et al.* Obesity and the Food Environment: Income and Ethnicity Differences Among People With Diabetes. *Diabetes Care* **36**, 2697–2705 (2013).
6. Powell, L. M., Chaloupka, F. J. & Bao, Y. The availability of fast-food and full-service restaurants in the United States: associations with neighborhood characteristics. *Am. J. Prev. Med.* **33**, S240-245 (2007).
7. Block, J. P., Scribner, R. A. & DeSalvo, K. B. Fast food, race/ethnicity, and income: a geographic analysis. *Am. J. Prev. Med.* **27**, 211–217 (2004).
8. Kwate, N. O. A., Yau, C.-Y., Loh, J.-M. & Williams, D. Inequality in obesigenic environments: fast food density in New York City. *Health Place* **15**, 364–373 (2009).
9. Pearce, J., Blakely, T., Witten, K. & Bartie, P. Neighborhood deprivation and access to fast-food retailing: a national study. *Am. J. Prev. Med.* **32**, 375–382 (2007).
10. Gordon, C. *et al.* Measuring food deserts in New York City's low-income neighborhoods. *Health Place* **17**, 696–700 (2011).
11. Storr, R., Carins, J. & Rundle-Thiele, S. Assessing Support for Advantaged and Disadvantaged Groups: A Comparison of Urban Food Environments. *Int. J. Environ. Res. Public. Health* **16**, 1135 (2019).
12. Kusuma, D. *et al.* Food environment and diabetes mellitus in South Asia: A geospatial analysis of health outcome data. *PLOS Med.* **19**, e1003970 (2022).
13. Kanchi, R. *et al.* Longitudinal Analysis of Neighborhood Food Environment and Diabetes Risk in the Veterans Administration Diabetes Risk Cohort. *JAMA Netw. Open* **4**, e2130789 (2021).
14. Davis, B. & Carpenter, C. Proximity of Fast-Food Restaurants to Schools and Adolescent Obesity. *Am. J. Public Health* **99**, 505–510 (2009).
15. Han, J., Schwartz, A. E. & Elbel, B. Does proximity to fast food cause childhood obesity? Evidence from public housing. *Reg. Sci. Urban Econ.* **84**, 103565 (2020).
16. Bodicoat, D. H. *et al.* Is the number of fast-food outlets in the neighbourhood related to screen-detected type 2 diabetes mellitus and associated risk factors? *Public Health Nutr.* **18**, 1698–1705 (2015).
17. Walker, B. B. *et al.* The Local Food Environment and Obesity: Evidence from Three Cities. *Obes. Silver Spring Md* **28**, 40–45 (2020).
18. Poelman, M. *et al.* Relations between the residential fast-food environment and the individual risk of cardiovascular diseases in The Netherlands: A nationwide follow-up study. *Eur. J. Prev. Cardiol.* **25**, 1397–1405 (2018).
19. Ntarladima, A.-M. *et al.* Associations between the fast-food environment and diabetes prevalence in the Netherlands: a cross-sectional study. *Lancet Planet. Health* **6**, e29–e39 (2022).

20. Lytle, L. A. & Sokol, R. L. Measures of the food environment: A systematic review of the field, 2007–2015. *Health Place* **44**, 18–34 (2017).
21. McKinnon, R. A., Reedy, J., Morrissette, M. A., Lytle, L. A. & Yaroch, A. L. Measures of the food environment: a compilation of the literature, 1990-2007. *Am. J. Prev. Med.* **36**, S124-133 (2009).
22. Gamba, R. J., Schuchter, J., Rutt, C. & Seto, E. Y. W. Measuring the food environment and its effects on obesity in the United States: a systematic review of methods and results. *J. Community Health* **40**, 464–475 (2015).
23. FoodData Central. https://fdc.nal.usda.gov/.
24. Mainardi, F., Prozorovscaia, D., Sweeney, G. & Green, H. Development and validation of a meal quality index with applications to NHANES 2005-2014. *PLoS ONE* **15**, e0244391 (2020).
25. Fulgoni, V. L., Keast, D. R. & Drewnowski, A. Development and Validation of the Nutrient-Rich Foods Index: A Tool to Measure Nutritional Quality of Foods, ,. *J. Nutr.* **139**, 1549–1554 (2009).
26. Drewnowski, A. Defining Nutrient Density: Development and Validation of the Nutrient Rich Foods Index. *J. Am. Coll. Nutr.* **28**, 421S-426S (2009).
27. Mamdouh, H. *et al.* Prevalence and associated risk factors of overweight and obesity among adult population in Dubai: a population-based cross-sectional survey in Dubai, the United Arab Emirates. *BMJ Open* **13**, e062053 (2023).
28. Slavin, J. L. Dietary fiber and body weight. *Nutrition* **21**, 411–418 (2005).
29. Grafova, I. B., Freedman, V. A., Kumar, R. & Rogowski, J. Neighborhoods and Obesity in Later Life. *Am. J. Public Health* **98**, 2065–2071 (2008).
30. Sheehan, C. M., Cantu, P. A., Powers, D. A., Margerison-Zilko, C. E. & Cubbin, C. Long-term neighborhood poverty trajectories and obesity in a sample of california mothers. *Health Place* **46**, 49–57 (2017).
31. Zick, C. D. *et al.* Running to the store? The relationship between neighborhood environments and the risk of obesity. *Soc. Sci. Med. 1982* **69**, 1493–1500 (2009).
32. Stuck in Unhealthy Places: How Entering, Exiting, and Remaining in Poor and Nonpoor Neighborhoods Is Associated with Obesity during the Transition to Adulthood - Adam M. Lippert, 2016. https://journals.sagepub.com/doi/full/10.1177/0022146515627682.
33. Alvarado, S. E. Neighborhood disadvantage and obesity across childhood and adolescence: Evidence from the NLSY children and young adults cohort (1986–2010). *Soc. Sci. Res.* **57**, 80–98 (2016).
34. Musaiger, A. O., Takruri, H. R., Hassan, A. S. & Abu-Tarboush, H. Food-based dietary guidelines for the arab gulf countries. *J. Nutr. Metab.* **2012**, 905303 (2012).
35. Bawajeeh, A. *et al.* Development of an Arabic food composition database for use in an Arabic online dietary assessment tool (myfood24). *J. Food Compos. Anal.* **102**, 104047 (2021).
36. Hetz, K. *et al.* The Development and Potential of a Digital out of Home Food Environment Monitoring Platform. *Nutrients* **15**, 3887 (2023).
37. Bates, S., Reeve, B. & Trevena, H. A narrative review of online food delivery in Australia: challenges and opportunities for public health nutrition policy. *Public Health Nutr.* **26**, 262–272 (2023).
38. Harrington, R. A., Adhikari, V., Rayner, M. & Scarborough, P. Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure. *BMJ Open* **9**, e026652 (2019).

39. Tufano, Michele (2024). Restaurants with menus collected from online delivery Platforms in Boston, London, and Dubai. figshare. Dataset. https://doi.org/10.6084/m9.figshare.25976056.
40. Hipp, R.D., 2020. SQLite, Available at: https://www.sqlite.org/index.html.
41. Douze, M. *et al.* The Faiss library. Preprint at http://arxiv.org/abs/2401.08281 (2024).
42. McKinney, W. & others, 2010. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference. pp. 51–56.
43. Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. https://plot.ly.
44. Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, Matthew Perry, Adrian Garcia Badaracco, Carson Farmer, Geir Arne Hjelle, Alan D. Snow, Micah Cochran, Sean Gillies, Lucas Culbertson, Matt Bartos, Nick Eubank, maxalbert, Aleksey Bilogur, ... François Leblanc. (2020). geopandas/geopandas: v0.8.1 (v0.8.1). Zenodo. https://doi.org/10.5281/zenodo.3946761
45. Folium python-visualization. Version 0.11.0 (2020).
46. London Datastore – Greater London Authority. https://data.london.gov.uk/.
47. CDC. PLACES: Local Data for Better Health. *Centers for Disease Control and Prevention* https://www.cdc.gov/places/index.html (2023).

ACKNOWLEDGEMENTS

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the grant PROJ-01041/Province of Gelderland -Op Oost -EFRO InToEat. The authors wish to thank VLAG Graduate School from Wageningen University and Research for their support.

J.E., M.M., F.D, and C.R. thank Dubai Future Foundation, UnipolTech, Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria, Volkswagen Group America, FAE Technology, Samoo Architects & Engineers, Shell, GoAigua, ENEL Foundation, Kyoto University, Weizmann Institute of Science, Universidad Autónoma de Occidente, Instituto Politecnico Nacional, Imperial College London, Universitá di Pisa, KTH Royal Institute of Technology, AMS Institute and all the members of the MIT Senseable City Lab Consortium for supporting this research.

Grubhub data were sourced from the Grubhub website at https://www.grubhub.com/. JustEat's data were sourced from the JustEat website at https://www.just-eat.co.uk/. UberEats's data were sourced from the UberEats website at https://www.ubereats.com/. Talabat's data were sourced from the Talabat website at https://www.talabat.com. Deliveroo's data were sourced from the Deliveroo website at https://www.deliveroo.co.uk and https://www.deliveroo.ae. The authors wish to thank Lucrezia Bertoni and Karin Borgonjen for their nutritional advice.

AUTHOR CONTRIBUTIONS

MT, FD, MM, JE, and GC conceived the experiments. MT, and JE created the primary data processing pipelines and performed the first data exploration. MT, FD, MM, JE, and GC created the final data presentation and interpretation. All authors reviewed and approved the final manuscript.

ETHICS DECLARATIONS

COMPETING INTERESTS

The authors declare no competing interests.

3

Capturing eating behavior from video analysis: a systematic review

Michele Tufano[1], Marlou Lasschuijt[1], Aneesh Chauhan[2], Edith J.M. Feskens[1] and Guido Camps [1,3]

[1] Division of Human Nutrition and Health, Wageningen University & Research, the Netherlands
[2] Wageningen Food and Biobased Research, Wageningen University & Research, the Netherlands
[3] OnePlanet Research Center, Wageningen, the Netherlands

Published in Nutrients (2022)
https://doi.org/10.3390/nu14224847

ABSTRACT

Current methods to detect eating behavior events (i.e., bites, chews, and swallows) lack objective measurements, standard procedures, and automation. The video recordings of eating episodes provide a non-invasive and scalable source for automation. Here, we reviewed the current methods to automatically detect eating behavior events from video recordings. According to PRISMA guidelines, publications from 2010-2021 in PubMed, Scopus, ScienceDirect, and Google Scholar were screened through title and abstract, leading to the identification of 277 publications. We screened the full text of 52 publications and included 13 for analysis. We classified the methods in five distinct categories based on their similarities and analyzed their accuracy. Facial landmarks can count bites, chews, and food liking automatically (accuracy: 90%, 60%, 25%). Deep neural networks can detect bites and gesture intake (accuracy: 91%, 86%). The active appearance model can detect chewing (accuracy: 93%), and optical flow can count chews (accuracy: 88%). Video fluoroscopy can track swallows but is currently not suitable beyond clinical settings. The optimal method for automated counts of bites and chews is facial landmarks, although further improvements are required. Future methods should accurately predict bites, chews, and swallows using inexpensive hardware and limited computational capacity. Automatic eating behavior analysis will allow the study of eating behavior and real-time interventions to promote healthy eating behaviors.

INTRODUCTION

Eating behavior determines the nutritional intake and the health status of adults and children. Eating behavior is defined as the ensemble of food choices, eating habits, and eating events (bites, chews, and swallows)[1]. Eating rate, which is the amount of food consumed per unit of time (g/min), can affect food intake[2], energy intake[3], and weight gain[4,5], as well as the risk of obesity[6,7], and metabolic diseases[8,9]. Eating behavior can be influenced by the food consumed, although it develops through parent-child interactions, individual child growth, neural mechanisms, and social influences[10]. For example, eating rate is an individual trait but it strongly depends on food properties, such as food texture and matrix[11,12]. Solid foods with hard textures (difficult to bite and chew) decrease eating rate, food intake, and energy intake, whereas semi-solid or liquid foods increase eating rate, food intake, and energy intake[13]. To prevent food overconsumption and obesity, interventions in food texture and eating rate can manipulate individual eating behavior and lower food and energy intake [14,15].

Tracking each eating episode (i.e., a meal) is crucial for a valid comprehension of individual eating behavior. The golden standard for this process consists of two or three independent researchers that watch the videos of each eating episode and record the eating behavior events [16]. For the annotation, the number of eating events, bite-size, chewing frequency, eating rates, meal duration, and rate of ingestion [17] must be recorded. Measuring eating behavior events requires the training of human annotators and often the purchase of expensive software licenses. The most used software packages for eating behavior annotation are Noldus Observer XT[18], ELAN 4.9.1 Max Planck Institute for Psycholinguistics[19], and ChronoViz[20]. Although human annotation can be accurate, often this task is prone to subjectivity and attentional lapses, due to its repetitive and time-consuming nature. Furthermore, large prospective studies are unfeasible due to the large number of videos to annotate. Because of this, the evidence in the eating behavior field is confined to cross-sectional and short-term experimental studies[3,21]. Therefore, according to the experts in the field, the human annotation process should be automated[22,23].

Despite the recent advancements in smart devices for tracking eating behavior, including the wristband[24], ear sensors[25], smart fork[26], smart utensils[27], smart plate[28], smart tray[29], and wearable cameras[30], the video recordings of eating episodes remain the least intrusive and most scalable approach. Video recordings are able not only to reproduce wearables functionalities (e.g., eating rate, number of bites) but also to expand them towards more complex

eating behavior events (e.g., emotion detection for eating behavior, social interactions at the table, or parent-child interaction[31]).

Such fortes make video recordings of eating episodes a strong candidate for tracking eating behavior. The automatic analysis of video recordings can replace the expensive and time-consuming manual annotation and lead to better interventions to manipulate eating behavior. However, it remains unclear what methods are applicable to analyze meal videos automatically.

Therefore, the aim is to determine accuracy, advantages and disadvantages of the current video-based automated measures of eating behavior. This review focuses on video-based methods that aim to predict bites, chews, swallows of consuming foods and food liking.

MATERIALS AND METHODS

Search strategy

This systematic review was performed to assess the available methods to automatically detect eating behavior events. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed for the literature search[32]. The databases PubMed, Scopus (Elsevier), ScienceDirect, and Google Scholar were screened. The terms included in the search strategy were eating, behavior (or behaviour), video, methodology and further analysis terms (see Supplementary Methods section 8.1 for the comprehensive list of search queries). Additional author search was performed for the most common authors found, using the snowballing search strategy. All the citations were exported to the reference manager software Zotero (version 5.0.96.3), where the first author (MT) screened all the titles and abstracts to select the scientific publications that met the criteria as outlined below.

Inclusion and exclusion criteria

Original research articles were considered as valid exclusively if published in the English language and containing findings on video analysis for human eating behavior from January 2010 to December 2021. This temporal cut-off was chosen to ensure that outdated (computer vision and machine learning) technologies would be excluded. Conference papers were included. These studies might contain preliminary results therefore validity and precision in Table 1 were considered while writing this review. Articles concerning non-human studies were excluded. We excluded research articles on eating behavior with video electroencephalogram monitoring, verbal interaction analysis, or sensors, as

well as research studies not focusing on automated measures as they are beyond the scope of video analysis.

Article selection

A complete overview of the selection process is depicted in the PRISMA flow diagram[32] (See Appendix A). 239 records were found through database searching and 38 through author search, for a total of 277 records identified from all sources. After removing 20 internal or external duplicates, 257 unique research articles were assessed. To determine their eligibility, the research articles were screened based on their titles and abstracts. The main exclusion criterium was the absence of video analysis from the title and abstract (n = 92). We discarded the publications not concerning eating behavior research (n=57). These publications reported about types of behavior not related to eating behavior and were therefore unrelated. Non-human studies (n=23), non-original research studies (n=26), and publications written in a language other than English (n=7) were considered non-eligible. The full text of each of the remaining scientific publications (n=52) was screened rigorously. The articles were excluded due to the following criteria: videos on EEG monitoring (n=2), eating behavior not including automated measures(n=15), using verbal interactions between patients and caregivers for eating behavior (n=6), published before 2010 (n=6), and using sensors or wearables for tracking eating behavior (n=10). Finally, the remaining 13 publications were included in this study for review and data extraction.

Summary characteristics and data extraction

For all the eligible studies, study characteristics such as the methodology used, the research outcomes, and the validation procedure were retrieved. Additionally, the precision of the outcomes was reported using the metrics from the scientific publications: F1-score, accuracy score, precision score, or average error with standard deviation. We tabulated the results from the original publications and summarized the data narratively.The methods were classified by the authors based on their similarities and dissimilarities. This resulted in 5 categories: facial landmarks, deep learning, optical flow, active appearance model, and video fluoroscopy. The extraction was performed by one author (MT). Since this review is centered on the methodology rather than on the outcomes of the studies, the risk of bias was not assessed.

RESULTS

The literature was reviewed systematically according to the PRISMA guidelines to assess the methods to automatically detect eating behavior from video recordings. Overall, the main methods found were facial landmarks, deep learning, optical flow, and active appearance model which can be combined to detect bites, chews, food intake, and food liking. Video fluoroscopy was the only method applied to detect swallows. Facial landmarks are the most used method for detecting bites and chews automatically from video recordings of eating episodes. A summary of the main methods is provided in the next section with their application for eating behavior events (Table 1).

Table 1. – The results table presents the publications included in this review study. The following information is summarized from left to right: first author, year of publication, journal, methods, outcomes of the study, validation method, and precision (as reported in the paper).

1. First author	3. Year	4. Journal	5. Methods	6. Study Setting	7. Device	8. Outcomes	9. Validity	10. Precision
11. Cadavid 12. [33]	14. 2012	15. Pers. Ubiquit. Comput.	16. Active Appearance Model (AAM) for face tracking, and spectral analysis on the temporal window of the model parameter values; binary support vector machine classifier for chewing events	17. Laboratory	18. Not reported; 37 videos at 24 fps; frame resolution: 640x480	19. Chewing detection	20. Manual annotation for chewing events	21. 93% after cross-validation
22. Okamoto [34]	24. 2014	25. IEEE International Conference on Multimedia and Expo Workshops	26. Mouth detector limited to the lower part of detected face; Chopstick detection using OpenCV Hough transform for straight lines	27. Laboratory	28. Smartphone Google Nexus 5 (2.3GHz Quad Core, Android 4.4), inner camera; frontal view	29. Food intake estimation	30. N.A.	31. N.A.
32. Hantke 33. [35]	35. 2018	36. Proceedings of the 20th ACM International Conference on Multimodal Interaction	37. OpenFace facial landmarks extraction for tracking the mouth	38. Office room	39. Logitech HD Pro Webcam C920; 30 fps; resolution: 1280 x 720; frontal view	40. Food liking	41. Leave-One-Out Cross-Validation and SVM	42. Likability 0.583

Table 1. – The results table presents the publications included in this review study. The following information is summarized from left to right: first author, year of publication, journal, methods, outcomes of the study, validation method, and precision (as reported in the paper). (*continued*)

1. First author	3. Year	4. Journal	5. Methods	6. Study Setting	7. Device	8. Outcomes	9. Validity	10. Precision
43. Haider 44. [36]	46. 2018	47. Proceedings of the 20th ACM International Conference on Multimodal Interaction	48. OpenSMILE for facial landmarks extraction, coupled with OpenSMILE audio-feature extraction	49. Office room	50. Logitech HD Pro Webcam C920; 30 fps; resolution: 1280 x 720; frontal view	51. Food liking	52. Leave-One-Out Cross-Validation and active feature transformation	53. 0.61
54. Konstantinidis 55. [37]	57. 2019	58. Computer Vision Systems	59. OpenPose for mouth and hands tracking; Deep Network (3 Conv + shortcut, 3 Conv + shortcut, 3 LSTM)	60. Laboratory	61. 85 videos; Samsung digital camcorder; 1.5 m away from the subject; side view	62. Automatic bite detection	63. F-Score: 0.9173	64. 0.9175
65. Qiu 66. [38]	68. 2019	69. IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks	70. Mask R-CNN for 360-degree camera meal videos; Thresholds for assessing pixel intersection between hand-face and hand-food to infer eating events	71. Free-living (Indoor food sharing scenarios)	72. Samsung's gear 360 camera; 1024 x 1024 pixels	73. Food intake estimation	74. N.A.	75. N.A.

Table 1. – The results table presents the publications included in this review study. The following information is summarized from left to right: first author, year of publication, journal, methods, outcomes of the study, validation method, and precision (as reported in the paper). (*continued*)

1. First author	3. Year	4. Journal	5. Methods	6. Study Setting	7. Device	8. Outcomes	9. Validity	10. Precision
76. Hossain 77. [39]	79. 2020	80. IEEE Access	81. Face detection with manual selection of the region of interest; CNN for bite/non bite classification; Optical flow for spatial chewing motion at every pixel	82. Laboratory	83. 84 videos;SJCAM SJ4000 Action Camera; 84. 1080p video at 30fps; side view	85. Automatic count of bites and chews	86. Manual annotation with 3-button system and LabView software (custom-made)	87. Bites: 88.9% ±7.4%; Chews: 88.64% ± 5.29%
88. Rouast and Adam 89. [40]	91. 2020	92. IEEE J. Biomed. Health Inform	93. CNN for hand-to-mouth movement in 360-degree meal videos	94. Free-living (Indoor group meal)	95. 102 videos; 360fly-4 K camera; 24fps	96. Intake gesture detection	97. N.A.	98. F1-score: 0.858
99. Konstantinidis 100. [41]	102. 2020	103. Nutrients	104. OpenPose skeletal and mouth features extracted for training the RABiD algorithm. Two stream data: 2D coordinates and distances from mouth corners, and from upper body	105. Laboratory	106. Samsung digital camcorder; 1.5 m away from the subject; side view; resolution: 576p (720 × 576 pixels) at 25fps	107. Meal duration and bite counts	108. Manual annotation (Noldus Observer XT)	109. F1-score: 0.948

Table 1. – The results table presents the publications included in this review study. The following information is summarized from left to right: first author, year of publication, journal, methods, outcomes of the study, validation method, and precision (as reported in the paper). (*continued*)

1. First author	3. Year	4. Journal	5. Methods	6. Study Setting	7. Device	8. Outcomes	9. Validity	10. Precision
110. Nour 111. [42]	113. 2021	114. Advances in Social Sciences Research Journal	115. Facial landmarks (dlib) for tracking jawline movement; OpenPose for 2D pose estimation	116. N.A.	117. N.A.	118. Real-time eating activity tracking	119. Manual annotation	120. N.A.
121. Park 122. [43]	124. 2020	125. Robotics and Autonomous System	126. Facial landmarks (dlib) for mouth-pose estimator; Algorithmic model for improving 3D estimation, location, and orientation of the mouth	127. Laboratory	128. Intel SR300 RGB-D camera	129. Robot active feeding assistance	130. Wrist-mounted camera	131. N.A.
132. Alshboul 133. [44]	135. 2021	136. Sensors	137. Time series data consisting of Euclidean distance between jaw/mouth landmarks and a reference facial landmark	138. Free-living (outdoors, indoors, and public spaces)	139. 300 videos; Huawei Y7 Prime 2018 smartphone; 13 MP camera; resolution: 1080p at 30fps; frontal view	140. Number of chews	141. Manual annotation (Intra-class correlation coefficient = slow: 0.96, normal: 0.94, fast: 0.91)	142. Avg Error ± SD: 5.42% ± 4.61 (slow chewing) 7.47% ± 6.85 (normal chewing) 9.84% ± 9.55 (fast chewing)

Table 1. – The results table presents the publications included in this review study. The following information is summarized from left to right: first author, year of publication, journal, methods, outcomes of the study, validation method, and precision (as reported in the paper). (*continued*)

1. First author	3. Year	4. Journal	5. Methods	6. Study Setting	7. Device	8. Outcomes	9. Validity	10. Precision
143. 144. Kato [45]	146. 2021	147. Gerodontology	148. Video fluoroscopy of swallowing for determining which foods are more appropriate for elderly people	149. Laboratory	150. N.A.	151. Association between masticatory movements and food texture in older adults	152. N.A.	153. N.A.

Facial Landmarks

Face detection is a computer technology able to recognize human faces in an image or video. Facial landmarks (or key points) detection is a computer vision task to localize and track key points on a human face.

Several open-source computer vision packages have been developed for facial recognition and landmarks. OpenSMILE is a toolkit for facial detection, which can also extract and analyze audio features for sound and speech interpretation[46]. OpenFace is an open-source package for facial landmark detection, eye-gaze and head pose estimation with real-time performance (using a common webcam)[47]. OpenPose is an open-source package for multi-person, real-time, 2D pose detection with facial, body, hand, and foot landmarks[48]. The Viola-Jones face detector is a machine learning framework for face detection. It combines Haar-like Features, the AdaBoost Algorithm, and the Cascade Classifier, and it can also be adapted to object detection[49]. The Kazemi and Sullivan landmark detector is a machine learning framework for face alignment. It uses an ensemble of regression trees to estimate the facial landmarks from the pixels of an image or video[50].

Facial landmarks (or detection) were used in 7 out of 13 (54%) of the studies included. To predict food liking, Hantke et al. (2018) applied OpenFace for facial landmark extraction during the 2018 EAT challenge. The participants scored their food preferences using a continuous slider with values between 0 (extremely dislike) and 1 (extremely like), which were later mapped into two variables ('Neutral' or 'Like') using a threshold. Facial landmarks were used to predict food liking. The mouth-related subset of landmarks yielded a better performance than landmarks from the entire face, which showed overfitting. A support vector machine (SVM) was optimized using Leave-One-Out cross validation to recognize food liking automatically (accuracy: 0.583)[35].

To classify food liking, Haider et al., (2018) used the OpenSMILE[46] package for facial landmarks coupled with audio feature extraction. In a Leave-One-Out cross validation setting, they employed the active feature transformation method to find a subset of 104 features that provided better results (Unweighted Average Recall = 0.61) for food liking (out of the 988 features from the whole dataset)[36].

To detect eating activity, Nour et al. (2021)[42] used 68 facial landmarks to locate the mouth and OpenPose to detect hands during eating episodes[48,51]. Eating activity was detected when the hands were in proximity of the mouth (accuracy: not reported)[42].

To detect chewing, Alshboul et al. (2021) used the Viola-Jones face detector[44] to detect faces and the Kazemi and Sullivan landmark detector[50] to apply facial landmarks. The videos were recorded outdoors, indoors, and in public spaces, with different light intensities. The video frame rate was 30 fps. The Euclidean distances were calculated between the jaw and mouth landmarks and a reference point (upper left corner of the face rectangle)[44].

To automatically detect bites, Kostantinidis et al. (2019) extracted x- and y-mouth features coordinates applying OpenPose in videos of eating episodes. Subsequently, the coordinates were used to model bites, by calculating the mouth corner between mid-upper and mid-lower lips (the classification task was performed with deep neural networks, see paragraph 3.2)[37].

To estimate food intake, Okamoto et al. (2014) restricted the detection area of the Viola-Jones face detector to the lower half of the face to minimize wrong predictions[34]. Interestingly, Okamoto et al. (2014) developed a chopstick detector by separating the front portion of the image from the background and applying the OpenCV Hough transform[52] for straight-line detection to enhance linear object detection (i.e., chopsticks)[34].

To feed people with impaired mobility, Park et al., (2020) developed a robotic system with facial landmarks. The system localizes the user's face and detects 68 facial landmarks using dlib[51] with the histogram of oriented gradient feature, a sliding window detector, and a linear classifier. They improved the model for light variations, 3D facial estimation, and facial orientation to detect when the mouth is opened. The facial recognition system is combined with a mobile manipulator to automatically deliver the food to the user's mouth [43]. In conclusion, facial landmarks can predict bites and chews, although the camera angle can impact their performance and only 2D facial landmarks have been tested so far.

Deep Neural Networks

Deep learning approaches are a subset of machine learning methods, in particular artificial neural networks, designed to automatically extract representations (also referred to as features) directly from raw input [53]. In deep neural networks, as the network gets deeper, several levels of features (from raw input to more abstract representations) are extracted by composing simple non-linear modules (artificial neurons distributed over hidden layers). Unlike conventional machine learning methods, these features are not usually designed by humans and learned directly from raw data using a general-purpose learning procedure.

A multitude of deep neural networks have been proposed in literature which are distinct in design and architecture [53–55].

Numerous deep neural networks are commonly used for image and video classification tasks. Convolutional Neural Networks (CNNs) are an example of deep neural networks. CNN is an architecture inspired by biological neuron connections. It consists of an input layer, hidden layer(s) and an output layer which are connected by activation functions. Some CNNs can be differentiated based on the input file. 2D-CNNs are commonly used to process RGB images (or video frame by frame). In contrast, 3D-CNNs use a tridimensional input file, such as a video file or a sequence of 2D frames. Other CNNs can combine different input files. For example, Two-Stream CNNs combine data from the RGB-images with optical flow for action recognition.

CNNs can analyze input with a temporal component. CNN-LSTM (Long Short-Term Memory) with feedback connections are well suited for time-series data. SlowFast combines a slow and fast pathway for analyzing the dynamic and static content of a video.

Some CNNs are specialized in object detection (e.g., Faster R-CNN [56]) and instance segmentation (e.g., Mask R-CNN [57]). To detect an object, Faster R-CNN replaced the selective region search with a region proposal network, which boosts the detection task. Mask R-CNN are an extension of Faster R-CNN. After the region proposal network, Mask R-CNN classifies the region and then it classifies the pixels within the region to generate an object mask. CNNs were used in 4 out of the 13 (30%) of the included studies, which assessed CNNs performance for a given classification task (e.g., bite or no-bite).

To assess food intake in shared eating settings, Qui et al. (2019) rescaled the videos from a 360-degree camera to use it as the input for Mask R-CNN [38]. In this free-living setup, a 360 camera recorded a video from the center of a table where three subjects shared a meal. For each subject, a box was applied on the food, person, face, and hands. When the distance among the pixels in the face-hand-food boxes was relatively short, the system predicted a dietary intake event (accuracy not reported)[38]**.** To detect a bite, Rouast et al. (2020) investigated 2D-CNN (F1-score: 0.795), 3D-CNN (F1-score: 0.840), CNN-LSTM (F1-score: 0.856), Two-Stream CNN (F1-score: 0.836), SlowFast (F1-score: 0.858). The SlowFast with ResNet-50 architecture is the best model to predict a bite through intake gesture detection. The video sessions were recorded using a 360 camera on the table where four subjects shared a meal[40].

Kostantinidis et al. (2020) combined facial landmarks (OpenFace) with CNNs[41] to develop RABiD, a deep learning-based algorithm, for bite classification. RABiD combines temporal and spatial interactions (convolutional layers, max-pooling steps, LSTM, and fully connected layers) in a two-data stream deep learning-based algorithm[41]. In RABiD, the first data stream uses 2D features from mouth corners, while the second data stream uses 2D features from the upper body. The mouth, head, and hands predicted bites with F1-score of 0.948 [41].

To count bites automatically, Hossain et al. (2020) used a Faster R-CNN. Initially, human raters manually marked the participants' faces as the region of interest to train the Faster R-CNN for face detection. The bite images consisted of image frames including the face together with straw/glass/bottle/hand/spoon/fork and food in the field of view. A binary image classifier (with AlexNet architecture) was trained to distinguish between 'bite' images and 'non-bite' images. This method achieved an accuracy of 85.4% ± 6.2% in counting bites automatically in 84 videos of eating episodes[39]. In summary, deep neural networks can detect human body and predict bites. However, deep learning is not efficient in predicting chewing and it requires expensive hardware and software requirements.

Optical Flow

Optical flow is a computer vision approach that tracks motion of surfaces, objects, and edges between consecutive image frames. Each image frame is converted to a 3D vector field to describe space and time. The spatial motion is calculated on the 3D vector fields at every pixel[58]. The resulting values (or parameters) can be used to assess the movement of any object using videos as input.

Optical flow was used in 2 of the 13 (15%) included studies. To estimate chewing activity, Hossain et al. (2020) used optical flow to extract spatial motion parameters from the jaw (accuracy: 88.64% ± 5.29% in 84 meal videos)[39]. To detect a bite, Rouast et al. (2020) used a motion stream with the horizontal and vertical components of the optical flow. The motion stream was integrated into a 2D-CNN. The models using optical flows (Small 2D-CNN, F1-score: 0.487; ResNet50 2D-CNN, F1-score: 0.461) performed worse than models using image frames (Small 2D-CNN, F1-score: 0.674; ResNet50 2D-CNN, F1-score: 0.795) [40]. Optical flow presents the advantage of not being restricted to a certain camera angle. It can predict chews but not detect bites.

Active Appearance Model

AAM is a computer vision algorithm that uses the statistical model of an object's shape and appearance. The model is optimized to detect differences between objects in consecutive video frames. The model parameter values are used for least square techniques (or spectral regression) to match the object's appearance to a new image. The resulting data can be used for training a classifier.

Cadavid et al., (2012) was the only publication (one out of 13, 8%) that used AAM to distinguish between chewing and non-chewing facial actions, for which they achieved a precision of 93%, after cross-validation[33]. AAM can detect chews but not bites and generally it is not widely used.

Video fluoroscopy

Video fluoroscopy is a moving X-ray examination of swallowing, which displays the bolus movement through the oropharyngeal anatomical structures. Physicians use video fluoroscopy to gain insights into the eating mechanisms and the problems concerning mastication (e.g., dysphagia, or choking)[59].

The only study (1 out of 13, 8%) to use video fluoroscopy was Kato et al., (2021) detected swallows in older adults in order to determine which foods are more appropriate for elderly people (accuracy not reported)[45]. Overall, video fluoroscopy can track swallow. However, the disadvantages of video fluoroscopy are its elevated costs and dimensions, which limit the use of this technology to the clinical setting. A summary of the advantages and disadvantages of all the methods reviewed is presented (Table 2).

Table 2. – Advantages and disadvantages of each method reviewed in this study.

Method	Advantages	Disadvantages	Open-source
Facial Landmarks	Bite detection Chew detection	Camera angle and distance can impact performance Only tested in 2D No swallow detection	Yes
Deep neural networks	Bite prediction Human body detection	Not efficient for chewing prediction Hardware and software requirements No swallow detection	Yes
Otical flow	Chewing prediction No camera angle restrictions	No bite detection Motion analysis outperformed by appearance analysis No swallow detection	Yes

Table 2. – Advantages and disadvantages of each method reviewed in this study. (*continued*)

Method	Advantages	Disadvantages	Open-source
Active appearance model	Chewing detection	No bite detection Not widely used No swallow detection	Yes
Video fluoroscopy	Swallow tracking	Cost and dimensions	No

DISCUSSION

In this systematic review we determined the accuracy, advantages and disadvantages of the current video-based automated methods for eating behavior. The main methods found were facial landmarks, deep learning, optical flow, and active appearance model. These methods can detect bites, chews, food intake, and food liking. Facial landmarks can be used to count bites, chews, and food liking automatically (accuracy: 90%, 60%, and 25%, respectively). CNNs can detect bites and gesture intake detection (accuracy: 91%, 86% respectively). AAM can be used to detect chewing (accuracy: 93%), and optical flow can be used to count chews (accuracy: 88%). To detect swallows, video fluoroscopy was the only method found; however, video fluoroscopy is not suited beyond a clinical setting. Facial landmarks are the most used method for detecting bites and chews automatically from video recordings of eating episodes.

To our knowledge, this is the first study that describes and gives an overview of video-based automated measures of eating behavior. Our study provides a comprehensive overview of the available methods for detecting eating behavior events automatically from video recordings. Currently, the manual annotation of eating episodes is a time-consuming and expensive task, which is prone to subjectivity and attentional lapses. Large prospective eating behavior studies are unfeasible using the manual annotation. Thus, there is a demand for the annotation process to be automated. To aid in realizing this, we provided a systematic overview of the methods to automate the annotation process. Facial landmarks can be used to count bites and chews by tracking facial and body motion during videos of eating episodes [35–37,42]. However, the camera's distance, angle, occlusion, darkness and camera-lens focal length can limit their efficiency [60]. The reviewed publications used 2D facial landmarks methods only; however, such methods appear too stringent for the 3D real-world application due to their low performance in tracking facial motion from a side view. Facial landmarks predicted food liking [35,36]. However, these studies did not include emotion detection methods that can predict food liking [61,62]. A synergistic interaction

between facial landmarks and emotion detection can enhance food liking predictions and consumer's acceptance of new food products [63].

CNNs can be used to count bites [50–53]**.** CNNs can be used to model eating behavior gestures that include human hands and body (e.g., intake gestures consisting of fine cutting, loading the food, and leading the food to the mouth). CNNs are effective at recognizing differences between consecutive video frames (e.g., presence/absence of hand in proximity of the mouth). However, CNNs are not effective for tracking movement between consecutive video frames (e.g., jaw movement during chewing).

Optical flow can detect chewing [39] . When coupled with a facial detector, optical flow can track jaw movement during chewing. Optical flow is not restricted to the camera angle: it tracks motion regardless of the user's position. To detect intake gestures, optical flow is not indicated. Rouast et al. (2020) showed that using frames as input (appearance analysis) outperformed motion streams as input (optical flow analysis) [40].

AAM can be used to distinguish between chewing and talking[33]. However, more recent methods for facial and object recognition are commonly preferred to automatically detect eating behavior events.

Video fluoroscopy is an accurate and non-invasive technology that can track swallows[45], however it is inappropriate and inaccessible for automating eating behavior analysis due to its costs and dimensions.

Several limitations may be recognized. First, there is a discrepancy in comparing accuracy and performance metrics, due to different study designs and data magnitude. Second, three conference papers were included, although their future full version may include more details or updates. Third, this study includes only methods that analyze video recordings, without considering the accuracy of other methods for automated eating behavior analysis (e.g., bone conduction microphone, algorithmic modeling from scales data, magnetic jaw displacement).

Future work should focus on addressing current issues to provide updated methods for the eating behavior field. The 3D facial landmarks should be applied to improve accuracy from a lateral camera angle. Furthermore, standard procedures should be established for camera angle, data extraction from video, reference facial landmarks, and algorithms to detect eating behavior events. Importantly, privacy concerns regarding face recognition should be addressed.

To achieve automatic swallow detection, optical flow should be tested for tracking throat movement.

Only three publications [38,40,44] found in this study used free-living conditions: two with a 360 camera in the middle of the table, and one using a combination of video recordings from indoor, outdoor, and public spaces. For sensor-based detection, free-living is a valid setup to detect food intake [64]. In the future, video-based detection of eating events should extend to free-living conditions, possibly placing 360 cameras in different positions.

Video-based methods should consider how the awareness of being monitored affects social modeling of eating and eating intake (particularly for energy-dense foods) [65,66].

Future personalized nutrition recommender systems could be implemented by combining automatic eating behavior analysis with food and calories intake estimation [67]. Advancements should be implemented through open-source software, which can boost collaboration in the field. It is hoped that the developments of future methods will provide objective measures to conduct prospective studies and allow intervention strategies to decrease eating rate and, subsequently, overeating.

CONCLUSIONS

Based on this systematic review, the use of facial landmarks is the most promising method to detect eating behavior events automatically from video recordings because it is the only method that can detect both bites and chews. Improvements of this technology are needed to standardize procedures. CNNs can detect bites automatically and optical flow can detect chews automatically, but feasible method to detect swallows are currently lacking.

Ideally, future methods should detect bites, chews, and swallows from video recordings using inexpensive hardware with low computational requirements. Future methods should be implemented with open-source software to boost collaboration and development. The automated video analysis of eating episodes would improve eating behavior research and provide real-time feedback to the consumers to improve their weight status and health.

Author Contributions: CRediT author statement - Michele Tufano: study design, methodology, writing, original draft preparation, reviewing; Marlou Lasschuijt.:

study design, methodology, reviewing and editing; Aneesh Chauhan: reviewing and editing; Edith J.M. Feskens: reviewing and editing; Guido Camps: study design, reviewing and editing, funding acquisition. All authors have approved the final article.

Funding: This work was supported by the Province of Gelderland - Op Oost - EFRO InToEat [Grant number PROJ-01041].

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to ethical reasons.

Conflicts of Interest: The authors declare no conflict of interest.

APPENDIX A

Search Queries

Scopus
(Only articles in English)

TITLE-ABS-KEY ((video AND eating AND behaviour OR behavior AND methodology OR chews OR bites OR meal OR machine AND learning OR computer AND vision OR automated OR analysis)) AND PUBYEAR > 2009

ScienceDirect
(With date range 2010-2022, all publication types)

video AND eating AND behaviour AND chew AND chewing AND bites AND meal AND machine learning AND computer vision AND human

PubMed
(With date range 2010-2022, all publication types, only human, only articles in English)

(((((video[Title/Abstract]) AND ((eating)[Title/Abstract])) AND ((behaviour[Title/Abstract] OR behavior)[Title/Abstract])) AND ((methodology)[Title/Abstract])) AND ((meal)[Title/Abstract] OR (chews[Title/Abstract]) AND (bites[Title/Abstract]) OR (meal[Title/Abstract]) OR (machine[Title/Abstract]) AND (learning[Title/Abstract]) OR (computer[Title/Abstract]) AND (vision[Title/Abstract]) OR (automated[Title/Abstract]) OR (analysis[Title/Abstract])) AND (computer[Title/Abstract] AND vision[Title/Abstract] OR automated[Title/Abstract] OR analysis)[Title/Abstract]))
Google Scholar

(With date range 2010-2022, all publication types, only articles in English)

human video eating behavior behaviour chew bites chewing meal machine learning computer vision automated analysis automatic methods methodology

PRISMA Flow Diagram

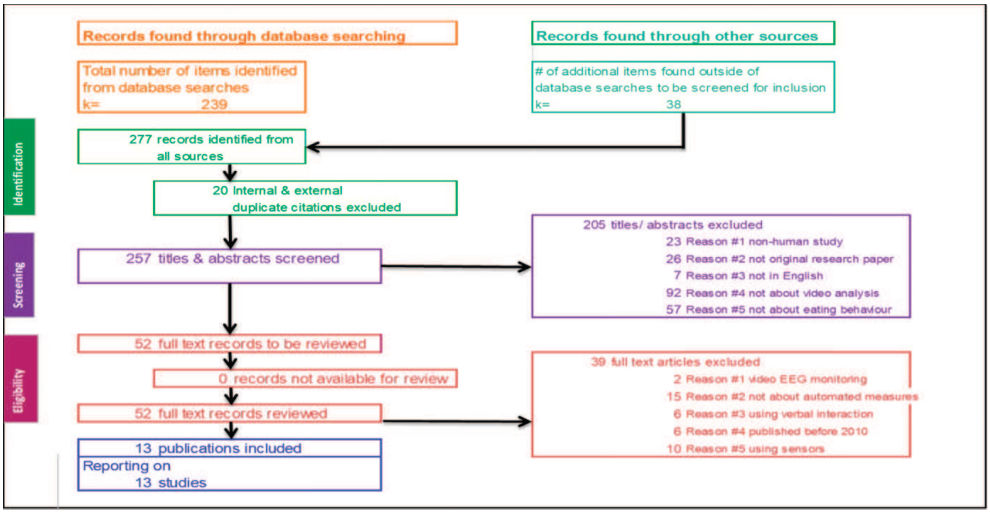


Figure 1. - PRISMA flow diagram that depicts the identification, screening, and eligibility workflow for including the scientific records in this study

REFERENCES

1. LaCaille, L. Eating Behavior. In *Encyclopedia of Behavioral Medicine*; Gellman, M.D., Turner, J.R., Eds.; Springer: New York, NY, 2013; pp. 641–642 ISBN 978-1-4419-1005-9.
2. Viskaal-van Dongen, M.; Kok, F.J.; de Graaf, C. Eating Rate of Commonly Consumed Foods Promotes Food and Energy Intake. *Appetite* **2011**, *56*, 25–31, doi:10.1016/j.appet.2010.11.141.
3. Robinson, E.; Almiron-Roig, E.; Rutters, F.; de Graaf, C.; Forde, C.G.; Tudur Smith, C.; Nolan, S.J.; Jebb, S.A. A Systematic Review and Meta-Analysis Examining the Effect of Eating Rate on Energy Intake and Hunger. *The American Journal of Clinical Nutrition* **2014**, *100*, 123–151, doi:10.3945/ajcn.113.081745.
4. Fogel, A.; McCrickerd, K.; Aris, I.M.; Goh, A.T.; Chong, Y.-S.; Tan, K.H.; Yap, F.; Shek, L.P.; Meaney, M.J.; Broekman, B.F.P.; et al. Eating Behaviors Moderate the Associations between Risk Factors in the First 1000 Days and Adiposity Outcomes at 6 Years of Age. *Am J Clin Nutr* **2020**, *111*, 997–1006, doi:10.1093/ajcn/nqaa052.
5. van den Boer, J.H.W.; Kranendonk, J.; van de Wiel, A.; Feskens, E.J.M.; Geelen, A.; Mars, M. Self-Reported Eating Rate Is Associated with Weight Status in a Dutch Population: A Validation Study and a Cross-Sectional Study. *International Journal of Behavioral Nutrition and Physical Activity* **2017**, *14*, 121, doi:10.1186/s12966-017-0580-1.
6. Otsuka, R.; Tamakoshi, K.; Yatsuya, H.; Murata, C.; Sekiya, A.; Wada, K.; Zhang, H.M.; Matsushita, K.; Sugiura, K.; Takefuji, S.; et al. Eating Fast Leads to Obesity: Findings Based on Self-Administered Questionnaires among Middle-Aged Japanese Men and Women. *Journal of Epidemiology* **2006**, *16*, 117–124, doi:10.2188/jea.16.117.
7. Ohkuma, T.; Hirakawa, Y.; Nakamura, U.; Kiyohara, Y.; Kitazono, T.; Ninomiya, T. Association between Eating Rate and Obesity: A Systematic Review and Meta-Analysis. *Int J Obes* **2015**, *39*, 1589–1596, doi:10.1038/ijo.2015.96.
8. Sakurai, M.; Nakamura, K.; Miura, K.; Takamura, T.; Yoshita, K.; Nagasawa, S.; Morikawa, Y.; Ishizaki, M.; Kido, T.; Naruse, Y.; et al. Self-Reported Speed of Eating and 7-Year Risk of Type 2 Diabetes Mellitus in Middle-Aged Japanese Men. *Metabolism* **2012**, *61*, 1566–1571, doi:10.1016/j.metabol.2012.04.005.
9. Zhu, B.; Haruyama, Y.; Muto, T.; Yamazaki, T. Association Between Eating Speed and Metabolic Syndrome in a Three-Year Population-Based Cohort Study. *Journal of Epidemiology* **2015**, *25*, 332–336, doi:10.2188/jea.JE20140131.
10. Gahagan, S. The Development of Eating Behavior - Biology and Context. *J Dev Behav Pediatr* **2012**, *33*, 261–271, doi:10.1097/DBP.0b013e31824a7baa.
11. Forde, C.G.; de Graaf, K. Influence of Sensory Properties in Moderating Eating Behaviors and Food Intake. *Frontiers in Nutrition* **2022**, *9*.
12. Forde, C.G.; Bolhuis, D. Interrelations Between Food Form, Texture, and Matrix Influence Energy Intake and Metabolic Responses. *Curr Nutr Rep* **2022**, doi:10.1007/s13668-022-00413-4.
13. Bolhuis, D.P.; Forde, C.G. Application of Food Texture to Moderate Oral Processing Behaviors and Energy Intake. *Trends in Food Science & Technology* **2020**, *106*, 445–456, doi:10.1016/j.tifs.2020.10.021.
14. Abidin, N.Z.; Mamat, M.; Dangerfield, B.; Zulkepli, J.H.; Baten, M.A.; Wibowo, A. Combating Obesity through Healthy Eating Behavior: A Call for System Dynamics Optimization. *PLoS One* **2014**, *9*, e114135, doi:10.1371/journal.pone.0114135.

15. Shavit, Y.; Roth, Y.; Teodorescu, K. Promoting Healthy Eating Behaviors by Incentivizing Exploration of Healthy Alternatives. *Front Nutr* **2021**, *8*, 658793, doi:10.3389/fnut.2021.658793.
16. Pesch, M.H.; Lumeng, J.C. Methodological Considerations for Observational Coding of Eating and Feeding Behaviors in Children and Their Families. *International Journal of Behavioral Nutrition and Physical Activity* **2017**, *14*, 170, doi:10.1186/s12966-017-0619-3.
17. Doulah, A.; Farooq, M.; Yang, X.; Parton, J.; McCrory, M.A.; Higgins, J.A.; Sazonov, E. Meal Microstructure Characterization from Sensor-Based Food Intake Detection. *Frontiers in Nutrition* **2017**, *4*, 31, doi:10.3389/fnut.2017.00031.
18. Resources - The Observer XT | Noldus Available online: https://www.noldus.com/observer-xt/resources (accessed on 11 January 2022).
19. "ELAN (Version 6.2) [Computer Software]. (2021). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive.
20. Fouse, A.; Weibel, N.; Hutchins, E.; Hollan, J. *ChronoViz: A System for Supporting Navigation of Time-Coded Data*; 2011; p. 304;.
21. Krop, E.M.; Hetherington, M.M.; Nekitsing, C.; Miquel, S.; Postelnicu, L.; Sarkar, A. Influence of Oral Processing on Appetite and Food Intake – A Systematic Review and Meta-Analysis. *Appetite* **2018**, *125*, 253–269, doi:10.1016/j.appet.2018.01.018.
22. Ioakimidis, I.; Zandian, M.; Eriksson-Marklund, L.; Bergh, C.; Grigoriadis, A.; Sodersten, P. Description of Chewing and Food Intake over the Course of a Meal. *Physiology & behavior* **2011**, *104*, 761–769, doi:10.1016/j.physbeh.2011.07.021.
23. Hermsen, S.; Frost, J.H.; Robinson, E.; Higgs, S.; Mars, M.; Hermans, R.C.J. Evaluation of a Smart Fork to Decelerate Eating Rate. *Journal of the Academy of Nutrition and Dietetics* **2016**, *116*, 1066–1068, doi:10.1016/j.jand.2015.11.004.
24. Kyritsis, K.; Tatli, C.L.; Diou, C.; Delopoulos, A. Automated Analysis of in Meal Eating Behavior Using a Commercial Wristband IMU Sensor. *Annu Int Conf IEEE Eng Med Biol Soc* **2017**, *2017*, 2843–2846, doi:10.1109/EMBC.2017.8037449.
25. Bi, S.; Caine, K.; Halter, R.; Sorber, J.; Kotz, D.; Wang, T.; Tobias, N.; Nordrum, J.; Wang, S.; Halvorsen, G.; et al. Auracle: Detecting Eating Episodes with an Ear-Mounted Sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **2018**, *2*, 1–27, doi:10.1145/3264902.
26. Hermsen, S.; Mars, M.; Higgs, S.; Frost, J.H.; Hermans, R.C.J. Effects of Eating with an Augmented Fork with Vibrotactile Feedback on Eating Rate and Body Weight: A Randomized Controlled Trial. *International Journal of Behavioral Nutrition and Physical Activity* **2019**, *16*, 90, doi:10.1186/s12966-019-0857-7.
27. Smart-U: Smart Utensils Know What You Eat | IEEE Conference Publication | IEEE Xplore Available online: https://ieeexplore.ieee.org/document/8486266 (accessed on 11 January 2022).
28. Mertes, G.; Ding, L.; Chen, W.; Hallez, H.; Jia, J.; Vanrumste, B. *Quantifying Eating Behavior With a Smart Plate in Patients With Arm Impairment After Stroke*; 2019;
29. Lasschuijt, M.P.; Brouwer-Brolsma, E.; Mars, M.; Siebelink, E.; Feskens, E.; Graaf, K. de; Camps, G. Concept Development and Use of an Automated Food Intake and Eating Behavior Assessment Method. *JoVE (Journal of Visualized Experiments)* **2021**, e62144, doi:10.3791/62144.
30. Gemming, L.; Doherty, A.; Utter, J.; Shields, E.; Ni Mhurchu, C. The Use of a Wearable Camera to Capture and Categorise the Environmental and Social Context

of Self-Identified Eating Episodes. *Appetite* **2015**, *92*, 118–125, doi:10.1016/j.appet.2015.05.019.
31. van Bommel, R.; Stieger, M.; Visalli, M.; de Wijk, R.; Jager, G. Does the Face Show What the Mind Tells? A Comparison between Dynamic Emotions Obtained from Facial Expressions and Temporal Dominance of Emotions (TDE). *Food Quality and Preference* **2020**, *85*, 103976, doi:10.1016/j.foodqual.2020.103976.
32. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Ann Intern Med* **2009**, *151*, 264–269, W64, doi:10.7326/0003-4819-151-4-200908180-00135.
33. 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. *Pers Ubiquit Comput* **2012**, *16*, 729–739, doi:10.1007/s00779-011-0425-x.
34. Okamoto, K.; Yanai, K. Real-Time Eating Action Recognition System on a Smartphone. In Proceedings of the 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW); July 2014; pp. 1–6.
35. Hantke, S.; Schmitt, M.; Tzirakis, P.; Schuller, B. EAT -: The ICMI 2018 Eating Analysis and Tracking Challenge. In Proceedings of the Proceedings of the 20th ACM International Conference on Multimodal Interaction; ACM: Boulder CO USA, October 2 2018; pp. 559–563.
36. Haider, F.; Pollak, S.; Zarogianni, E.; Luz, S. SAAMEAT: Active Feature Transformation and Selection Methods for the Recognition of User Eating Conditions. In Proceedings of the Proceedings of the 20th ACM International Conference on Multimodal Interaction; Association for Computing Machinery: New York, NY, USA, October 2 2018; pp. 564–568.
37. Konstantinidis, D.; Dimitropoulos, K.; Ioakimidis, I.; Langlet, B.; Daras, P. A Deep Network for Automatic Video-Based Food Bite Detection. In Proceedings of the Computer Vision Systems; Tzovaras, D., Giakoumis, D., Vincze, M., Argyros, A., Eds.; Springer International Publishing: Cham, 2019; pp. 586–595.
38. Qiu, J.; Lo, F.P.-W.; Lo, B. Assessing Individual Dietary Intake in Food Sharing Scenarios with a 360 Camera and Deep Learning. In Proceedings of the 2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN); May 2019; pp. 1–4.
39. Hossain, D.; Ghosh, T.; Sazonov, E. Automatic Count of Bites and Chews From Videos of Eating Episodes. *IEEE Access* **2020**, *8*, 101934–101945, doi:10.1109/ACCESS.2020.2998716.
40. Rouast, P.V.; Adam, M.T.P. Learning Deep Representations for Video-Based Intake Gesture Detection. *IEEE Journal of Biomedical and Health Informatics* **2020**, *24*, 1727–1737, doi:10.1109/JBHI.2019.2942845.
41. 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* **2020**, *12*, 209, doi:10.3390/nu12010209.
42. Nour, M. Real-Time Detection and Motivation of Eating Activity in Elderly People with Dementia Using Pose Estimation with TensorFlow and OpenCV. *Advances in Social Sciences Research Journal* **2021**, *8*, 28–34, doi:10.14738/assrj.83.9763.

43. Park, D.; Hoshi, Y.; Mahajan, H.P.; Kim, H.K.; Erickson, Z.; Rogers, W.A.; Kemp, C.C. Active Robot-Assisted Feeding with a General-Purpose Mobile Manipulator: Design, Evaluation, and Lessons Learned. *arXiv:1904.03568 [cs]* **2019**.
44. Alshboul, S.; Fraiwan, M. Determination of Chewing Count from Video Recordings Using Discrete Wavelet Decomposition and Low Pass Filtration. *Sensors (Basel)* **2021**, *21*, 6806, doi:10.3390/s21206806.
45. Kato, Y.; Kikutani, T.; Tohara, T.; Takahashi, N.; Tamura, F. Masticatory Movements and Food Textures in Older Patients with Eating Difficulties. *Gerodontology n/a*, doi:10.1111/ger.12596.
46. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the Proceedings of the 18th ACM international conference on Multimedia; Association for Computing Machinery: New York, NY, USA, October 25 2010; pp. 1459–1462.
47. Baltrušaitis, T.; Robinson, P.; Morency, L.-P. OpenFace: An Open Source Facial Behavior Analysis Toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV); March 2016; pp. 1–10.
48. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *arXiv:1812.08008 [cs]* **2019**.
49. Viola, P.; Jones, M. *Rapid Object Detection Using a Boosted Cascade of Simple Features*; 2001; Vol. 1, p. 511; ISBN 978-0-7695-1272-3.
50. Kazemi, V.; Sullivan, J. One Millisecond Face Alignment with an Ensemble of Regression Trees. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; June 2014; pp. 1867–1874.
51. King, D.E. Dlib-Ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
52. Bradski, G. OpenCV 2000.
53. Deep Learning | Nature Available online: https://www.nature.com/articles/nature14539 (accessed on 25 October 2022).
54. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data* **2021**, *8*, 53, doi:10.1186/s40537-021-00444-8.
55. Review of Deep Learning Algorithms and Architectures | IEEE Journals & Magazine | IEEE Xplore Available online: https://ieeexplore.ieee.org/document/8694781 (accessed on 25 October 2022).
56. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks 2016.
57. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); October 2017; pp. 2980–2988.
58. Beauchemin, S.S.; Barron, J.L. The Computation of Optical Flow. *ACM Comput. Surv.* **1995**, *27*, 433–466, doi:10.1145/212094.212141.
59. Matsuo, K.; Palmer, J.B. Video Fluoroscopic Techniques for the Study of Oral Food Processing. *Curr Opin Food Sci* **2016**, *9*, 1–10, doi:10.1016/j.cofs.2016.03.004.
60. Li, X.; Liu, J.; Baron, J.; Luu, K.; Patterson, E. Evaluating Effects of Focal Length and Viewing Angle in a Comparison of Recent Face Landmark and Alignment Methods. *EURASIP Journal on Image and Video Processing* **2021**, *2021*, 9, doi:10.1186/s13640-021-00549-3.

61. FaceReader Methodology Note What Is Facereader ?; 2016.
62. Lewinski, P.; den Uyl, T.M.; Butler, C. Automated Facial Coding: Validation of Basic Emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics* **2014**, *7*, 227–236, doi:10.1037/npe0000028.
63. Álvarez-Pato, V.M.; Sánchez, C.N.; Domínguez-Soberanes, J.; Méndoza-Pérez, D.E.; Velázquez, R. A Multisensor Data Fusion Approach for Predicting Consumer Acceptance of Food Products. *Foods* **2020**, *9*, 774, doi:10.3390/foods9060774.
64. Farooq, M.; Doulah, A.; Parton, J.; McCrory, M.A.; Higgins, J.A.; Sazonov, E. Validation of Sensor-Based Food Intake Detection by Multicamera Video Observation in an Unconstrained Environment. *Nutrients* **2019**, *11*, 609, doi:10.3390/nu11030609.
65. Thomas, J.M.; Dourish, C.T.; Higgs, S. Effects of Awareness That Food Intake Is Being Measured by a Universal Eating Monitor on the Consumption of a Pasta Lunch and a Cookie Snack in Healthy Female Volunteers. *Appetite* **2015**, *92*, 247–251, doi:10.1016/j.appet.2015.05.034.
66. Suwalska, J.; Bogdański, P. Social Modeling and Eating Behavior—A Narrative Review. *Nutrients* **2021**, *13*, 1209, doi:10.3390/nu13041209.
67. Theodoridis, T.; Solachidis, V.; Dimitropoulos, K.; Gymnopoulos, L.; Daras, P. A Survey on AI Nutrition Recommender Systems. In Proceedings of the Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments; Association for Computing Machinery: New York, NY, USA, June 5 2019; pp. 540–546.

4

Rule-based systems to automatically count bites from meal videos

Michele Tufano[1], Marlou Lasschuijt[1], Aneesh Chauhan[2], Edith J.M. Feskens[1], Guido Camps[1,3]

[1]Division of Human Nutrition and Health, Wageningen University & Research,
[2]Wageningen Food and Biobased Research, Wageningen University & Research,
Stippeneng 4, 6708 WE Wageningen, The Netherlands
[3]OnePlanet Research Center, Plus Ultra II, Bronland 10,
6708 WE Wageningen, The Netherlands
Published in Frontiers (2024)
https://doi.org/10.3389/fnut.2024.1343868

ABSTRACT

Eating behavior is a key factor for nutritional intake and plays a significant role in the development of eating disorders and obesity. The standard methods to detect eating behavior events (i.e., bites and chews) from video recordings rely on manual annotation, which lacks objective assessment and standardization. Yet, video recordings of eating episodes provide a non-invasive and scalable source for automation. Here, we present a rule-based system to count bites automatically from video recordings with 468 3D facial key points. We tested the performance against manual annotation in 164 videos from 15 participants. The system can count bites with 79% accuracy when annotation is available, and 71.4% when annotation is unavailable. The system showed consistent performance across varying food textures. Eating behavior researchers can use this automated and objective system to replace manual bite count annotation, provided the system's error is acceptable for the purpose of their study. Utilizing our approach enables real-time bite counting, thereby promoting interventions for healthy eating behaviors. Future studies in this area should explore rule-based systems and machine learning methods with 3D facial key points to extend the automated analysis to other eating events while providing accuracy, interpretability, generalizability, and low computational requirements.

INTRODUCTION

Eating behavior plays a key role in determining nutritional intake of people and is defined by food choices, eating habits and food oral processing (bites, chews, and swallows). Eating behavior is shaped throughout life by a combination of parent-child interactions, peer influences, interaction with the food environment, and the textural properties of food [1]–[3]. An example of individual eating behavior is the eating rate (food in grams/min), which has been shown to impact food intake [4], energy intake [5], and weight gain [6], [7]. Indeed, a fast eating rate can increase the risk of obesity [8], [9] and metabolic diseases [10], [11]. Therefore, measuring individual eating behavior is needed to support individual interventions to decrease eating rates and the risk of obesity [12], [13].

To measure eating behavior, eating episodes (i.e., a meal) must be analyzed to count eating behavior events. To achieve this, video recordings are an essential and non-invasive source of information. Currently, two trained researchers must watch the videos to annotate every eating event manually and compare their results to ensure an acceptable level of consistency [14]. Due to the repetitive and time-consuming nature of this task, manual annotation is prone to subjectivity and attentional lapses [15]. The current method does not allow large prospective studies and real-time feedback on individual eating behavior. To improve eating behavior research and perform research in larger cohorts, the human annotation process should be automated with computer technologies, such as face detection [16].

Face detection methods are used to recognize human faces and facial features in an image or video sequences. Key points (or landmark) detection is a computer vision task to localize and track key points on the human face and body from a camera or videos. Several 2D key point detectors and packages are available for face and facial features detection: Viola-Jones face detector [17], Kazemi-Sullivan key point detector [18], OpenSMILE [19], OpenFace [20], OpenPose [21] , and dlib [22]. A novel 3D face detector, Mediapipe can detect human faces and apply 468 key points to it (compared to the 68 key points of 2D detectors) [23]. These open-source packages can be tailored to solve a given task through machine learning or a rule-based system.

Machine learning approaches with facial key points have been used to detect eating behavior events from video recordings of eating episodes. For example, to predict difficulty of speaking while eating, mouth key points can be used to train support vector machines and deep neural networks algorithms for classification [24], [25]. To count bites, the mouth corner between upper and lower

lips can be calculated through mouth key points, and deep neural networks can be trained for the classification task [26]. To detect bites, mouth corners and upper body key points can be identified through key points to develop a deep learning-based algorithm for classification [15].

Facial key points can be coupled with rule-based systems, which employ the conditional "if-then" logic sequences to process inputs and execute decisions within a set of predetermined rules. In rule-based systems, the Euclidean distance is often employed to quantify the separation between two specific reference key points. This measurement serves as the foundation for establishing rules that classify various eating behaviors. For instance, the Euclidean distance between a reference key point on the left eye and another on the jaw region can be calculated to detect chewing [27]. Similarly, the distance between key points on hands and mouth can be used to detect eating, when the hands are near the mouth [28].

Facial key points are the optimal method to automatically count bites from video recordings, [29]. While existing techniques primarily rely on 68 2D facial key points for this purpose, the advantages of utilizing 3D facial key points have not yet been explored for bite detection. Therefore, this study aimed to assess the accuracy of a system that uses 468 3D facial key points in detecting bites from video recordings. Our focus was on identifying the least computationally expensive and most versatile solution to implement, which led us to explore key-point-based decision models, and to assess their viability and effectiveness in detecting bites. To our knowledge this study is the first to apply rule-based system with 3D facial key points to a large video dataset.

METHODS

Study participants

The study was performed at Wageningen University and Research (the Netherlands), Human Nutrition Research Unit, between October and December 2020. Master thesis students and personnel of the Division of Human Nutrition and Health were not allowed to participate in the study. Healthy adults (18-55 years, BMI: 18.5-30 kg/m^2) were recruited as participants through the divisional volunteer database and advertisements on social media. The participants signed informed consent and filled in the inclusion/exclusion questionnaire. The inclusion criteria were English proficiency and normally eating three meals per day during weekdays. In contrast, the exclusion criteria were food allergies to the foods used in this study, a lack of appetite, chewing or swallowing problems,

following an energy-restricted diet, more than 5 kg weight change during the last two months, alcohol consumption higher than 21 glasses per week, being on medications influencing appetite, taste or smell, intensive exercise for more than 8 hours a week, and being a high restrained eater (according to the Dutch Eating Behavior Questionnaire [30]). In total, 58 participants were eligible to participate in the study, of which 18 participants (11 females) were selected according to the abovementioned criteria and included in the study. The Social Ethical Committee of Wageningen University (the Netherlands) approved this study (Lasschuijt, 2020-11). The participants received financial compensation.

Meals and Conditions

The participants ate four breakfasts, lunch, dinner, and desserts in the diner/dining room of the behavior research unit for four days. The participants were instructed to eat as much or as little as they wanted until they felt comfortably full. For the breakfast meals, the participants were provided with one of the four distinct options: 1) fresh mixed fruit, 2) homemade smoothie, 3) canned mixed fruit, and 4) store-bought smoothie. For the lunch meals, the participants were provided with one of the four distinct options: 1) fresh tagliatelle pasta with homemade tomato sauce, hard-steamed vegetables, and large pieces of chicken fillet, 2) fresh tagliatelle pasta with homemade tomato sauce, soft-steamed vegetables, and homemade pulled chicken, 3) store-bought pork meat tortellini with pre-canned tomato sauce, hard-cooked vegetables, and grated cheese, and 4) ready-to-eat macaroni Bolognese with grated cheese. For the dinner meals, the participants were provided with one of the four distinct options: 1) potato parts with large pieces of pork fillet and whole hard-steamed green beans, 2) homemade mashed potato with eggs small pieces of soft-steamed green beans, 3) pre-flavored and baked potato parts with large pieces of chicken schnitzel and whole, hard-steamed green beans, 4) mashed potato with chicken meatballs and small pieces of soft-steamed green beans. For the dessert meals, the participants were provided with one of the four distinct options: 1) dried figs and almonds, 2) curd with added honey and crushed pecan nuts, 3) mass produced fig bread, 4) walnut and honey flavored yoghurt. According to the NOVA classification [31], two food conditions were chosen: unprocessed (category 1), processed/ultra-processed (category 3 and 4). Food texture was manipulated to create two more conditions: slow and fast. The slow condition included foods that would require small bites, many chews and therefore long oro-sensory exposure duration. Conversely, the fast condition included foods that would require large bites, fewer chews and therefore shorten the oro-sensory exposure duration. Food texture manipulation included solid/liquid manipulation (i.e., fresh fruit vs smoothie), or hardness and piece size manipulation (i.e., potato parts vs homemade mashed potato).

Video recordings

While eating a meal, the participants were recorded with a video camera (Axis M1054, Axis Communications) in front of their seat. The camera was positioned at approximately 1.5 m from the participant, with the lower frame in line with the table, the upper frame above the top of the cranium, and the sides of the frame at shoulder width. The videos were recorded using the software Noldus Observer XT 11 (Noldus Information Technology, the Netherlands) on Windows 10, installed on laptops (Lenovo Thinkpad L380, Intel core i5). To support the video analysis, the participants were instructed to show a numbered card to the camera before they started to eat and to raise a hand once they finished eating. The video recordings were later used for manual annotation and video analysis.

Manual annotation

The video recordings were annotated by using the software Noldus Observer XT 11 (Noldus Information Technology, the Netherlands). Two human annotators watched the videos and annotated the following eating behaviors: meal duration (min), duration between the bites (s), oro-sensory exposure (s), number of chews, and number of bites during the meal. The annotators calculated the bite size by dividing the total amount of food eaten (in grams) by the number of bites per meal. The eating rate (g/min) was then calculated by dividing the amount eaten (g) by the meal duration (min). The videos recordings from 3 participants could not be analyzed due to technical errors. In total, 170 videos from 15 participants were annotated.

Video analysis

The video recordings files were opened in Microsoft Photos on Windows 10 (Microsoft, Reedmond, WA, USA), and the video clip editor function was used to cut the video recordings. The videos were cut precisely to capture the entire eating process, beginning the moment the participant showed a numbered card before and ending when they set the ear-sensor on the table. While the ear-sensor was used to track jaw movement, data from this device was not included in the present study. The video recordings from the liquid breakfast meals (homemade smoothie and store-bought smoothie) were discarded for all participants because they were liquid therefore did not include bites.

The software implementation was achieved using Windows 10 on a Lenovo laptop (00329-00000-00003-AA-1666) with an Intel® Core™ i5-8265U CPU @ 1.60 GHz 1.80 GHz, 8.00 GB RAM. The programming language used for video analysis was Python 3.9 [32], with PyCharm 2021.2.2 (Community Edition, version 11.0.12+7-b1504.28, JetBrains) as IDE [33]. We used Mediapipe for 3D facial recognition [23] using the OpenCV computer vision library [34], and CVZone for

visualization [35]. NumPy [36] was used to convert a list of facial coordinates into an array and Pandas [37] to store the outcome in a dataset.

Threshold search

We extracted mouth coordinates from the video using Mediapipe with 3D key points. The mouth ratio is calculated by dividing the Euclidean distance between the upper and lower lips by the distance between the left and right sides of the mouth (key points 0, 17, 61, and 291, respectively). The mouth ratio was extracted from every frame of the meal video recordings. After performing outlier removal based on Z-Scores per participant, we discarded the meals with less than 10 bites (e.g., desserts). We used grid search, random search, and Bayesian optimization to find a custom threshold per participant.

We designed a custom model encapsulated in a Python class, to predict bite counts based on a predefined threshold value. The model does not require training and uses the threshold to count the bites in each sample that exceeds this value. Hyperparameter tuning is performed using Grid Search, Randomized Search, and Bayesian Optimization Cross-Validation methods to optimize the threshold parameter for each unique participant in our dataset. We used cross-validation folds set between 2 and 5 based on the number of samples. The performance was evaluated using a custom scoring function that calculates the negative mean of the absolute differences between the predicted and manually annotated bite counts. The best threshold values obtained for each participant were recorded and stored in a dataset to calculate accuracy in counting bites (Supplementary Table 1).

Dataset

The final dataset consisted of the total bites count from the manual annotation and bites predictions over videos in 15 participants for 6719 annotated bites. Descriptive statistics per participant, eating condition, and manual annotation can be found in Tables 1, 2, and 3, respectively.

Table 1 – Number of videos, average bite, and standard deviation per participant

Participant	Videos	Avg Bites	Std bites
A	10	55.4	17.3
B	14	35.9	15.4
C	14	39.4	20.1
D	13	34.6	24.9
E	11	59.7	28.2
F	6	40.2	21.0

Table 1 – Number of videos, average bite, and standard deviation per participant (continued)

Participant	Videos	Avg Bites	Std bites
G	14	34.1	21.7
H	14	43.4	22.6
I	13	34.1	16.0
L	13	51.4	28.2
M	7	49.9	25.7
N	12	32.4	16.2
O	8	22.4	12.1
P	9	45.1	16.0
Q	6	40.7	11.5

Table 2 – Number of videos, average bite, and standard deviation per eating episode.

Eating episode	Videos	Avg Bites	Std bites
Breakfast	25	43,4	18,48
Dessert	38	43,13	26,40
Dinner	50	37,82	20,36
Lunch	51	41,25	21,62

Table 3 – Descriptive statistics of the manual annotation of bites.

Mean	40.96
Median	40
Mode	50
Range	95
Variance	481.71
Standard Deviation	21.94
25th Percentile (Q1)	23
50th Percentile (Q2)	40
75th Percentile (Q3)	56

Counting Bites

We utilized a Python-based video processing pipeline to automatically count bites during eating events. The pipeline employs the CVZone library's "FaceMeshDetector" for facial landmark detection in each video frame [35]. Mouth key points are then extracted and used to calculate the mouth ratio. This ratio is stored and averaged over a window of 5 frames to smooth out noise. Bites are counted based on a predefined threshold for this mouth ratio, found as described above. To prevent double-counting, consecutive frames exceeding the threshold are considered as a single bite event, managed by an internal counter

within our custom class. Each bite event is timestamped and recorded for further analysis. The system could process multiple videos concurrently through the multiprocessing module. The accuracy was determined by comparing the number of bites predicted per video with the manual annotations, as outlined in [38]. The accuracy of total bite prediction is calculated on the overall number of bites per meal, rather than on the identification of each bite. To estimate the thresholds for participants lacking annotations, we used leave-one-out cross validation (LOOCV). We computed the mean threshold using values obtained through random selection from all other participants, excluding the one under evaluation. In each LOOCV iteration, the threshold was calculated by averaging the remaining 14 threshold values from the random search, omitting one value for validation. This yielded 15 different average thresholds, each leaving out a unique value.

RESULTS

This study used thresholds for facial key points to count bites from video recordings. We found thresholds per participant using grid search, random search, and Bayesian optimization and applied them to 164 video recordings of eating episodes from 15 study participants. The system counted a bite when the mouth ratio (the distance between the upper and lower lips to the width of the mouth) exceeded the threshold (Figure 1). The videos were manually annotated to provide the ground truth for the bite counts. We evaluated the accuracy of the three methods and determined the system's accuracy in a scenario where video annotations are unavailable. Further, we tested the system's accuracy on different meal textures.



Figure 1 - Example of the program counting a bite. On the left: the program displays the current bite count, the time from the start of the video, and the frames per second (FPS). The text is displayed on the video of a participant eating a meal with the facial key points applied to the face (in white), the lips highlighted (in orange), and the vertical and horizontal mouth lines that are used to calculate the mouth ratio (in white); On the right: the horizontal grey line represents the threshold for counting bites. The orange line represents the mouth ratio.

A) The mouth ratio is below the threshold. B) The participant takes a bite by inserting food in the mouth. During mouth opening, the mouth ratio increases and surpasses the threshold; the program counts one bite, and the bite count is updated to one. C) The participant closes the mouth after the bite therefore the mouth ratio is again below the threshold.

Overall, the threshold obtained with the random search achieved the highest accuracy of 79.8% (std: 15.2, min: 47.5%, max: 98.4%). The thresholds obtained with the grid search achieved 76.4% accuracy (std: 18.6, min: 33.9%, max: 99.2%), while the Bayesian optimization approach achieved 72.9% accuracy (std: 18.4, min: 28.4%, max: 99.6%) (Table 4).

Table 4 – Bite count and accuracy per participant and method, including manual annotation.

		Grid Search		*Random Search*		*Bayesian Optimization*	
Participant	Annotation	Predicted bites	Accuracy	Predicted bites	Accuracy	Predicted bites	Accuracy
A	554	524	94.6	532	96.0	369	66.6
B	502	662	68.1	647	71.1	1334	65.7
C	552	917	33.9	597	91.8	947	28.4
D	450	521	84.2	516	85.3	519	84.7
E	657	490	74.6	572	87.1	490	74.6
F	241	214	88.8	193	80.1	618	56.4
G	478	520	91.2	516	92.1	525	90.2
H	607	594	97.9	559	92.1	620	97.9
I	443	327	73.8	328	74.0	327	73.8
L	668	1031	45.7	1019	47.5	958	56.6
M	349	254	72.8	220	63.0	252	72.2
N	389	354	91.0	353	90.7	412	94.1
O	179	256	57.0	264	52.5	256	57.0
P	406	513	73.6	504	75.9	507	75.1
Q	244	246	99.2	240	98.4	243	99.6
Total	6719	7423		7060		8377	
Average	447.9	494.9	76.4	470.7	79.8	558.5	72.9
Standard Deviation			18.6		15.2		18.4

Figure 2 illustrates the range of performance outcomes across participants for the three methods under consideration. The system showed the highest performance in one participant, with an average accuracy of 99% across the three methods. This was followed by 3 participants, who achieved accuracies of 96%, 92%, and 91% respectively. On the other hand, some participants experienced subpar performance. Three out of 15 participants displayed below-average accuracy of 55.8%, 51.4% and 49.9%. Notably, one participant exhibited significant variability in the performance of the three methods: grid search yielded an

accuracy of 33.9%, random search achieved 91.8%, and Bayesian optimization resulted in 28.4%.

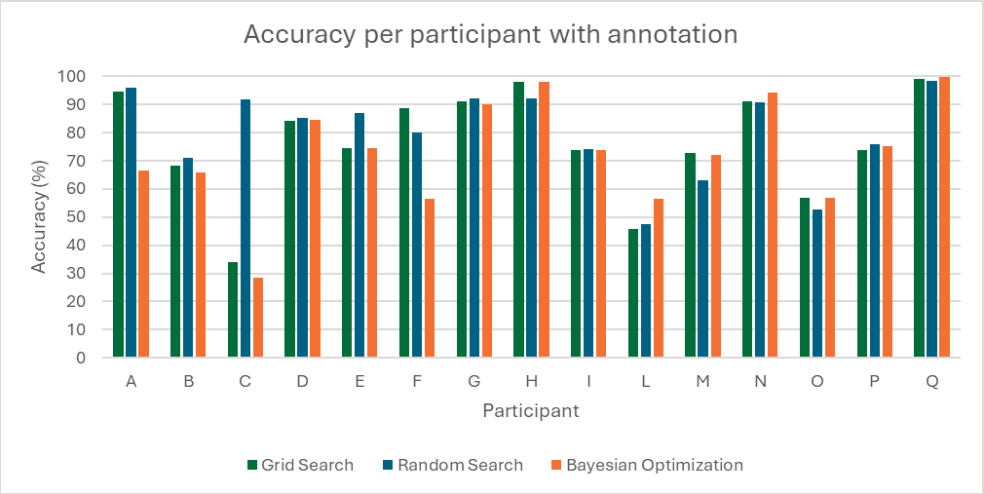


Figure 2 – Accuracy in percentage (y-axis) per participant (x-axis) with annotation for grid search (green), random search (blue), and Bayesian optimization (red).

To assess accuracy when video annotation is unavailable for a participant, we employed leave-one-out cross-validation. The overall accuracy per participant without annotation is 71.4% (std: 19.1, min: 36.6%, max: 94.6%). Half of the participants showed accuracy higher than 80%. The system performed poorly in three participants out of 15 with 46.6%, 40.5%, and 36.6% accuracy (Table 5, Figure 3).

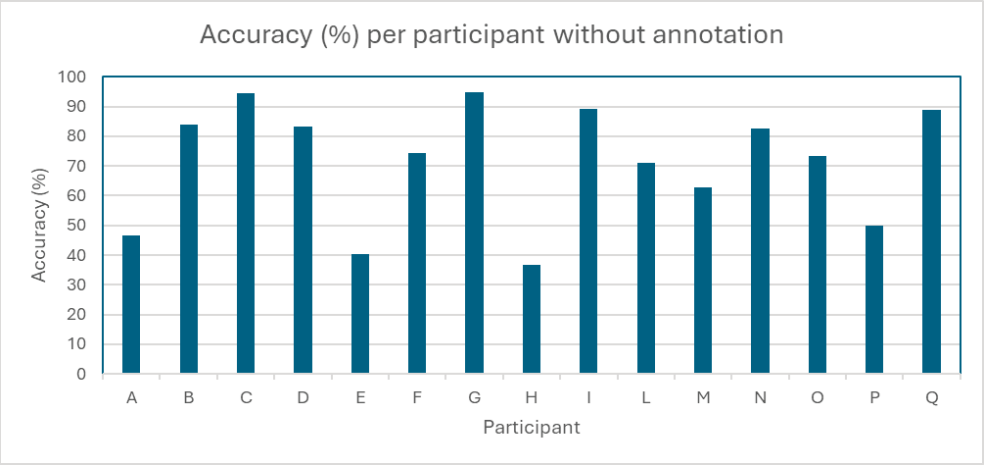


Figure 3 – Accuracy in percentage (y-axis) per participant (x-axis) without annotation.

Table 5 – Accuracy per participant without annotation, leave-one-out cross validation

Participant	Accuracy
A	46,6
B	83,9
C	94,6
D	83,1
E	40,5
F	74,3
G	94,8
H	36,6
I	89,2
L	71,0
M	62,8
N	82,5
O	73,2
P	49,8
Q	88,9
Average	71,4
Standard Deviation	19,1

To assess accuracy when food texture conditions are different, we calculated the system's accuracy for meals with soft or hard textures. The results revealed negligible differences: 93.5% accuracy for soft meals and 92.9% for hard meals (Table 6, Figure 4). Furthermore, Spearman's rank correlation coefficient for LOOCV versus manual annotation revealed a slightly stronger correlation in the hard texture than in the soft texture (ρ=0.7518 and ρ=0.7102, respectively, Figure 5).

Table 6 – Bite counts and accuracy per food texture condition

Condition	**Manual Annotation**	**Predicted Bites**	**Accuracy**
Fast (Soft texture)	2871	2684	93.5
Slow (Hard texture)	3848	4120	92.9

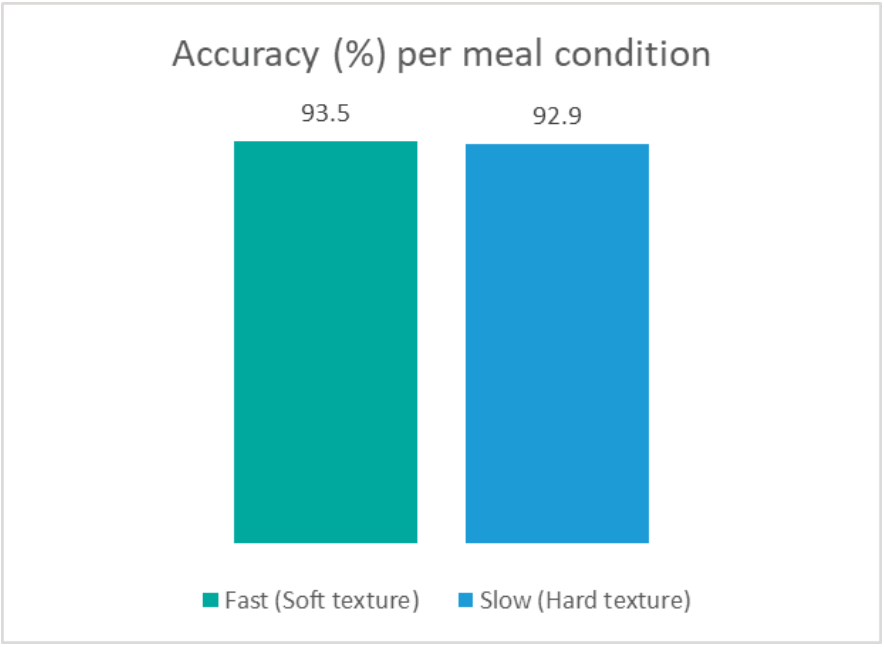


Figure 4 – Accuracy in percentage per meal texture condition in fast condition with soft texture (green) and slow condition with hard texture (blue).

A

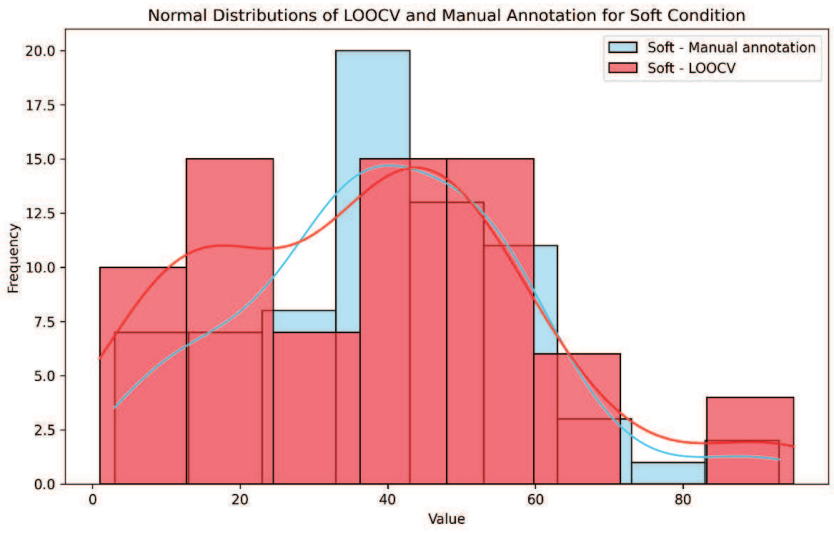


Figure 5 – (A) Normal distribution of LOOCV (light red) and manual annotation (light blue) for the soft texture. (B) Normal distribution of LOOCV (light red) and manual annotation (light blue) for the hard texture.

B

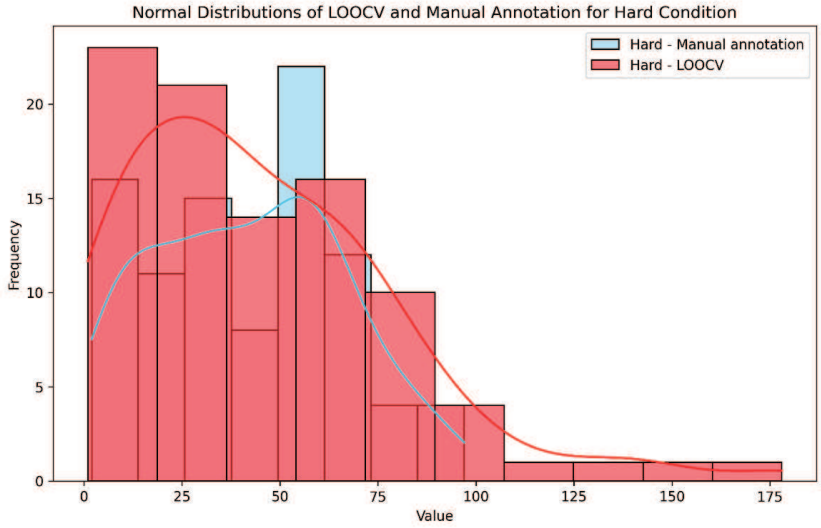


Figure 5 – (A) Normal distribution of LOOCV (light red) and manual annotation (light blue) for the soft texture. (B) Normal distribution of LOOCV (light red) and manual annotation (light blue) for the hard texture. (*continued*)

DISCUSSION

This study aimed to equip eating behavior researchers with an automatic, fast, standard, and objective method to automatically count bites from video recordings of eating episodes, using 3D facial key points. Our results demonstrate that bites can be detected without the use of deep learning methods. We tested three methods to find the threshold to count bites per participant: grid search, random search, and Bayesian optimization. With available annotation, the thresholds found with random search achieved the highest accuracy of 79.8%, followed by grid search (76.4%), and Bayesian optimization (72.9%). When annotated bites per participant are unavailable, the system achieved 71.4% accuracy. The system performed with comparable accuracy when the participants consumed a meal with soft or hard texture, thus across a slow and fast eating rate.

The system achieves improved accuracy with the availability of bite annotations in videos. In an eating behavior study, if the researchers count the bites per meal video of a participant, the system can achieve 99% accuracy. For one participant, six meal video recordings were annotated for total bite count, leading to 99% accuracy. This level of precision suggests that for some individuals

the system could serve as a reliable substitute for manual annotation methods, offering rapid and standardized measurements of bite counts. Another strength of this study is the system's consistent performance across hard and soft food textures. In the domain of eating behavior research, it is common to examine how participants respond to foods with different textural properties. The system we developed was not affected by variations in food texture and thus eating rate, thereby functioning as an unbiased tool for counting bites within the context of eating behavior studies. Yet, the system seems to slightly overpredict videos with hard texture food. The accuracy analysis and Spearman's correlation coefficient showed negligible differences in the system's accuracy when adapted to different food textures.

There are limitations of this study that should be considered. Firstly, the system showed inconsistent accuracy across participants. For three participants out of 15, a below-average accuracy of 55.8%, 51.4% and 49.9% was achieved. Such limitations render the system an unreliable substitute to replace manual annotation in those cases. Without annotated bites, the system's accuracy plunges from 79.8% to 71.4%. This could be a limitation, when precise bite counts are essential.

The performance variability among participants can be attributed to several factors. The participants for which the system achieved high accuracy (91.17%, 95.93%, and 91.93%) have a relatively high number of videos, which might provide a stable and generalized dataset to find the threshold. Furthermore, these participants have moderate bites standard deviation (21.7, 22.6, and 16.2, Table 1) suggesting that their eating behavior is not too erratic, which might decrease false bite predictions. Three participants out of 15 showed less accurate results (51.37, 49.93, and 55.50). Two of those participants had a high standard deviation, which suggests that the eating behavior is more variable and hence, more difficult to predict.

Moreover, one participant has an average bite count that is significantly lower than the average of all other participants (22.4 vs 42.6), which could affect the system's ability to effectively find the threshold for this case. Overall, the system's accuracy seems to be less effective (55.5%) for participants with less than 177 annotated bites. However, we did not find a correlation between accuracy and video duration or number of videos (Table S3), which prevents setting a minimum video duration requirement for our method. Lastly, it is possible that some external factors affect the performance of the system. For example, the lightning condition, angle of the camera, and presence of other people in the video frame might affect the system's ability to count bites. We recommend us-

ing this method after advising participants to avoid covering their mouth when eating and looking away from the camera. Future studies should explore several camera angles and lighting conditions.

Until now, researchers have used rule-based systems to detect chewing or eating activity, by calculating distances between facial key points or body parts [27], [28]. The previous research employed 68 2D facial key points. Here, we used 468 3D facial key points to count bites per meal. We tested the system in 164 meal videos from 15 participants. To our knowledge, this is the first study to examine the performance of 468 3D facial key points and a rule-based system to automatically count bites from meal video recordings. An advantage of our study was the opportunity to investigate such a system on a large video dataset of 164 videos from 15 participants. Compared to deep learning methods to count bites [38], our system offers greater adaptability and versatility for practical implementations.

Our study raises several opportunities for future research.

For instance, 3D facial key points could be used in combination with threshold search to count other eating events, and apply thresholds search to facial features different from the mouth ratio. Furthermore, 3D facial key points could be used in combination with machine learning and deep learning models to count bites and other eating behavior events. For instance, deep learning can predict chewing, an unfeasible task with our approach given the uniform mouth ratios associated with chewing behavior However, deep learning systems have high computational requirements, low interpretability, limited generalizability, and large datasets requirements. These limitations could make deep neural networks less accessible for researchers in eating behavior studies and limit their applicability across different research scenarios.

CONCLUSIONS

Our study shows that rule-based systems that use 468 3D facial key points can be used to count bites from video recordings. The system can count bites with 79% accuracy when annotation is available for a small video subset (i.e., six videos) , and 71.4% if it is applied to a new participant when annotation is unavailable. The system showed consistent performance across varying food with soft and hard textures. Compared to deep learning approaches, rule-based methods demand lower computational requirements, offer higher interpretability and generalizability, and require less data. For eating behavior researchers,

rule-based methods can serve as an accessible tool that can be applied across different research contexts. Eating behavior researchers can use this tool to replace manual annotation of bite count if the accuracy of 79% with available annotation and 71.4% with unavailable annotation is acceptable for the purpose of their study.

ACKNOWLEDGMENTS

The authors wish to thank Els Siebelink, Lise Heuven, Maressa Viana de Lima, Ilse Zwanenburg, Julia Roelofs, Femke de Gooijer, Kyra Konings, and Odette Paling for helping with the meal preparation, data collection and annotation, and Dr. Peter Griffith for his teachings on scientific writing. ChatGPT Model 4 (OpenAI) was used to rephrase the text, to design the Python class for threshold search, and to debug code. This work was supported by the grant PROJ-01041/ Province of Gelderland - Op Oost - EFRO InToEat.

REFERENCES

[1] S. Gahagan, "The Development of Eating Behavior - Biology and Context," *J. Dev. Behav. Pediatr. JDBP*, vol. 33, no. 3, pp. 261–271, Apr. 2012, doi: 10.1097/DBP.0b013e31824a7baa.

[2] C. G. Forde and K. de Graaf, "Influence of Sensory Properties in Moderating Eating Behaviors and Food Intake," *Front. Nutr.*, vol. 9, 2022, Accessed: May 25, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnut.2022.841444

[3] C. G. Forde and D. Bolhuis, "Interrelations Between Food Form, Texture, and Matrix Influence Energy Intake and Metabolic Responses," *Curr. Nutr. Rep.*, Mar. 2022, doi: 10.1007/s13668-022-00413-4.

[4] M. Viskaal-van Dongen, F. J. Kok, and C. de Graaf, "Eating rate of commonly consumed foods promotes food and energy intake," *Appetite*, vol. 56, no. 1, pp. 25–31, Feb. 2011, doi: 10.1016/j.appet.2010.11.141.

[5] E. Robinson *et al.*, "A systematic review and meta-analysis examining the effect of eating rate on energy intake and hunger," *Am. J. Clin. Nutr.*, vol. 100, no. 1, pp. 123–151, Jul. 2014, doi: 10.3945/ajcn.113.081745.

[6] A. Fogel *et al.*, "Eating behaviors moderate the associations between risk factors in the first 1000 days and adiposity outcomes at 6 years of age," *Am. J. Clin. Nutr.*, vol. 111, no. 5, pp. 997–1006, May 2020, doi: 10.1093/ajcn/nqaa052.

[7] J. H. W. van den Boer, J. Kranendonk, A. van de Wiel, E. J. M. Feskens, A. Geelen, and M. Mars, "Self-reported eating rate is associated with weight status in a Dutch population: a validation study and a cross-sectional study," *Int. J. Behav. Nutr. Phys. Act.*, vol. 14, no. 1, p. 121, Sep. 2017, doi: 10.1186/s12966-017-0580-1.

[8] R. Otsuka *et al.*, "Eating fast leads to obesity: findings based on self-administered questionnaires among middle-aged Japanese men and women," *J. Epidemiol.*, vol. 16, no. 3, pp. 117–124, May 2006, doi: 10.2188/jea.16.117.

[9] T. Ohkuma, Y. Hirakawa, U. Nakamura, Y. Kiyohara, T. Kitazono, and T. Ninomiya, "Association between eating rate and obesity: a systematic review and meta-analysis," *Int. J. Obes.*, vol. 39, no. 11, Art. no. 11, Nov. 2015, doi: 10.1038/ijo.2015.96.

[10] M. Sakurai *et al.*, "Self-reported speed of eating and 7-year risk of type 2 diabetes mellitus in middle-aged Japanese men," *Metabolism*, vol. 61, no. 11, pp. 1566–1571, Nov. 2012, doi: 10.1016/j.metabol.2012.04.005.

[11] B. Zhu, Y. Haruyama, T. Muto, and T. Yamazaki, "Association Between Eating Speed and Metabolic Syndrome in a Three-Year Population-Based Cohort Study," *J. Epidemiol.*, vol. 25, no. 4, pp. 332–336, 2015, doi: 10.2188/jea.JE20140131.

[12] N. Z. Abidin, M. Mamat, B. Dangerfield, J. H. Zulkepli, M. A. Baten, and A. Wibowo, "Combating Obesity through Healthy Eating Behavior: A Call for System Dynamics Optimization," *PLOS ONE*, vol. 9, no. 12, p. e114135, Dec. 2014, doi: 10.1371/journal.pone.0114135.

[13] Y. Shavit, Y. Roth, and K. Teodorescu, "Promoting Healthy Eating Behaviors by Incentivizing Exploration of Healthy Alternatives," *Front. Nutr.*, vol. 8, p. 277, 2021, doi: 10.3389/fnut.2021.658793.

[14] C. G. Forde, C. Leong, E. Chia-Ming, and K. McCrickerd, "Fast or slow-foods? Describing natural variations in oral processing characteristics across a wide range of Asian foods," *Food Funct.*, vol. 8, no. 2, pp. 595–606, Feb. 2017, doi: 10.1039/c6fo01286h.

[15] D. Konstantinidis, K. Dimitropoulos, B. Langlet, P. Daras, and I. Ioakimidis, "Validation of a Deep Learning System for the Full Automation of Bite and Meal Duration Analysis of Experimental Meal Videos," *Nutrients*, vol. 12, no. 1, Art. no. 1, Jan. 2020, doi: 10.3390/nu12010209.

[16] I. Ioakimidis, M. Zandian, L. Eriksson-Marklund, C. Bergh, A. Grigoriadis, and P. Sodersten, "Description of chewing and food intake over the course of a meal," *Physiol. Behav.*, vol. 104, pp. 761–9, Jul. 2011, doi: 10.1016/j.physbeh.2011.07.021.

[17] P. Viola and M. Jones, *Rapid Object Detection using a Boosted Cascade of Simple Features*, vol. 1. 2001, p. 511. doi: 10.1109/CVPR.2001.990517.

[18] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1867–1874. doi: 10.1109/CVPR.2014.241.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, in MM '10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1459–1462. doi: 10.1145/1873951.1874246.

[20] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–10. doi: 10.1109/WACV.2016.7477553.

[21] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *ArXiv181208008 Cs*, May 2019, Accessed: Dec. 15, 2021. [Online]. Available: http://arxiv.org/abs/1812.08008

[22] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[23] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," *ArXiv190608172 Cs*, Jun. 2019, Accessed: Sep. 09, 2021. [Online]. Available: http://arxiv.org/abs/1906.08172

[24] S. Hantke, M. Schmitt, P. Tzirakis, and B. Schuller, "EAT -: The ICMI 2018 Eating Analysis and Tracking Challenge," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, in ICMI '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 559–563. doi: 10.1145/3242969.3243681.

[25] F. Haider, S. Pollak, E. Zarogianni, and S. Luz, "SAAMEAT: Active Feature Transformation and Selection Methods for the Recognition of User Eating Conditions," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, in ICMI '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 564–568. doi: 10.1145/3242969.3243685.

[26] D. Konstantinidis, K. Dimitropoulos, I. Ioakimidis, B. Langlet, and P. Daras, "A Deep Network for Automatic Video-Based Food Bite Detection," in *Computer Vision Systems*, D. Tzovaras, D. Giakoumis, M. Vincze, and A. Argyros, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 586–595. doi: 10.1007/978-3-030-34995-0_53.

[27] S. Alshboul and M. Fraiwan, "Determination of Chewing Count from Video Recordings Using Discrete Wavelet Decomposition and Low Pass Filtration," *Sensors*, vol. 21, no. 20, p. 6806, Oct. 2021, doi: 10.3390/s21206806.

[28] M. Nour, "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.*, vol. 8, no. 3, Art. no. 3, Mar. 2021, doi: 10.14738/assrj.83.9763.

[29] M. Tufano, M. Lasschuijt, A. Chauhan, E. J. M. Feskens, and G. Camps, "Capturing Eating Behavior from Video Analysis: A Systematic Review," *Nutrients*, vol. 14, no. 22, p. 4847, Nov. 2022, doi: 10.3390/nu14224847.

[30] T. van Strien, J. E. R. Frijters, G. P. A. Bergers, and P. B. Defares, "The Dutch Eating Behavior Questionnaire (DEBQ) for assessment of restrained, emotional, and external eating behavior," *Int. J. Eat. Disord.*, vol. 5, no. 2, pp. 295–315, 1986, doi: 10.1002/1098-108X(198602)5:2<295::AID-EAT2260050209>3.0.CO;2-T.

[31] "Monteiro, C.A., Cannon, G., Lawrence, M., Costa Louzada, M.L. and Pereira Machado, P. 2019. Ultra-processed foods, diet quality, and health using the NOVA classification system. Rome, FAO." Accessed: Jun. 13, 2023. [Online]. Available: https://www.fao.org/3/ca5644en/ca5644en.pdf

[32] "Welcome to Python.org," Python.org. Accessed: Aug. 14, 2022. [Online]. Available: https://www.python.org/

[33] "PyCharm: the Python IDE for Professional Developers by JetBrains," JetBrains. Accessed: Aug. 14, 2022. [Online]. Available: https://www.jetbrains.com/pycharm/

[34] Bradski G., "The OpenCV Library," *Dr Dobb's Journal of Software Tools*, 2000.

[35] CVZone, "CVZone." Aug. 14, 2022. Accessed: Aug. 14, 2022. [Online]. Available: https://github.com/cvzone/cvzone

[36] "Array programming with NumPy | Nature." Accessed: Aug. 14, 2022. [Online]. Available: https://www.nature.com/articles/s41586-020-2649-2

[37] The pandas development team, "Pandas." Feb. 2020.

[38] D. Hossain, T. Ghosh, and E. Sazonov, "Automatic Count of Bites and Chews From Videos of Eating Episodes," *IEEE Access*, vol. 8, pp. 101934–101945, 2020, doi: 10.1109/ACCESS.2020.2998716.

SUPPLEMENTARY MATERIAL

Table S1 – Threshold used for counting bites found with random search and LOOCV methods.

Participant	Random Search	LOOCV
A	75.4	61.5
B	56.1	62.9
C	61.7	62.5
D	63.4	62.4
E	50.5	63.3
F	75.4	61.5
G	61.0	62.6
H	72.8	61.7
I	65.6	62.2
L	52.9	63.1
M	61.7	62.5
N	69.4	62.0
O	62.2	62.5
P	49.7	63.4
Q	59.4	62.7

Table S2 – Videos used to find the threshold per participant, with information on food texture, eating episode, total bites per video, and video duration (in MM:SS format)

Participant	Food texture	Eating Episode	Manual annotated bites	Duration (mm:ss)
A	soft	dessert	50	11:15
A	soft	dinner	54	15:52
A	soft	lunch	66	26:55
A	hard	breakfast	61	21:20
A	hard	dessert	13	08:50
A	hard	dinner	78	27:27
A	hard	lunch	65	31:15
A	soft	dinner	60	24:21
A	soft	lunch	60	24:35
B	soft	dinner	40	14:20
B	hard	dinner	52	23:21
B	soft	dinner	56	16:16
B	soft	lunch	54	18:29
C	soft	dinner	42	10:14
C	soft	lunch	38	14:29

Table S2 – Videos used to find the threshold per participant, with information on food texture, eating episode, total bites per video, and video duration (in MM:SS format) (*continued*)

Participant	Food texture	Eating Episode	Manual annotated bites	Duration (mm:ss)
C	hard	dinner	73	10:31
C	hard	lunch	52	20:57
C	soft	dinner	41	16:39
C	soft	lunch	58	16:22
C	hard	breakfast	36	09:57
C	hard	dinner	59	21:59
C	hard	lunch	63	16:48
D	soft	dinner	50	13:33
D	soft	lunch	55	13:10
D	hard	breakfast	33	13:04
D	hard	dessert	10	07:14
D	hard	dinner	83	20:30
D	hard	lunch	57	23:32
D	soft	dinner	37	09:30
D	soft	lunch	33	13:14
D	hard	breakfast	12	07:06
D	hard	dessert	11	07:57
D	hard	dinner	58	18:55
E	soft	dessert	35	05:23
E	soft	dinner	58	08:19
E	soft	lunch	49	09:19
E	hard	breakfast	23	05:34
E	hard	dinner	97	17:59
E	hard	lunch	62	13:37
E	soft	dessert	23	05:49
E	soft	dinner	90	16:20
E	soft	lunch	93	14:11
E	hard	breakfast	40	06:24
E	hard	lunch	87	14:29
F	soft	dinner	50	14:42
F	soft	lunch	22	18:07
F	soft	lunch	40	15:08
F	hard	lunch	63	19:51
F	hard	dinner	57	19:51
G	soft	dessert	15	08:03
G	soft	dinner	36	14:11

Table S2 – Videos used to find the threshold per participant, with information on food texture, eating episode, total bites per video, and video duration (in MM:SS format) (*continued*)

Participant	Food texture	Eating Episode	Manual annotated bites	Duration (mm:ss)
G	soft	lunch	32	17:15
G	hard	breakfast	30	16:17
G	hard	dinner	66	19:55
G	hard	lunch	55	21:02
G	soft	dessert	14	06:28
G	soft	dinner	42	12:52
G	soft	lunch	31	13:04
G	hard	breakfast	16	12:58
G	hard	dinner	72	27:28
G	hard	lunch	56	20:44
H	soft	dessert	36	09:28
H	soft	dinner	36	15:05
H	soft	lunch	52	23:46
H	hard	breakfast	51	19:51
H	hard	dinner	32	19:40
H	hard	lunch	55	25:07
H	soft	dessert	35	08:01
H	soft	dinner	67	12:23
H	soft	lunch	49	18:40
H	hard	dinner	62	26:26
I	soft	dessert	27	11:06
I	soft	dinner	48	12:34
I	soft	lunch	27	14:32
I	hard	breakfast	32	09:01
I	hard	dessert	11	07:19
I	hard	dinner	56	18:02
I	hard	lunch	39	18:18
I	soft	dessert	20	06:24
I	soft	dinner	32	08:08
I	soft	lunch	37	12:35
I	hard	breakfast	33	12:06
I	hard	dinner	61	29:16
I	hard	lunch	41	20:31
L	soft	dessert	10	02:21
L	soft	dinner	52	11:09
L	soft	lunch	81	17:58

Table S2 – Videos used to find the threshold per participant, with information on food texture, eating episode, total bites per video, and video duration (in MM:SS format) (*continued*)

Participant	Food texture	Eating Episode	Manual annotated bites	Duration (mm:ss)
L	hard	breakfast	50	24:14
L	hard	dinner	72	16:06
L	hard	lunch	88	26:02
L	soft	dessert	30	04:48
L	soft	lunch	65	14:56
L	hard	breakfast	62	17:37
L	hard	dinner	71	21:24
L	hard	lunch	70	17:11
M	hard	breakfast	35	10:53
M	hard	dessert	14	04:58
M	hard	lunch	79	13:44
M	soft	lunch	51	08:19
M	hard	breakfast	28	10:30
M	hard	dinner	60	24:46
M	hard	lunch	82	18:58
N	soft	dinner	40	08:58
N	soft	lunch	43	09:14
N	hard	breakfast	18	06:35
N	hard	dinner	40	15:05
N	hard	lunch	52	11:45
N	soft	dinner	44	08:32
N	soft	lunch	54	09:04
N	hard	breakfast	16	06:20
N	hard	dinner	36	10:53
N	hard	lunch	30	07:17
O	soft	lunch	26	12:52
O	hard	dinner	23	14:39
O	hard	lunch	22	17:11
O	soft	dinner	33	12:04
O	hard	breakfast	11	07:34
O	hard	dinner	41	11:59
O	hard	lunch	21	13:50
P	soft	dinner	54	10:52
P	soft	lunch	33	09:07
P	hard	breakfast	58	18:12
P	hard	dinner	55	19:03

Table S2 – Videos used to find the threshold per participant, with information on food texture, eating episode, total bites per video, and video duration (in MM:SS format) (*continued*)

Participant	Food texture	Eating Episode	Manual annotated bites	Duration (mm:ss)
P	soft	dinner	58	11:12
P	soft	lunch	39	08:54
P	hard	dinner	50	13:41
P	hard	lunch	50	09:24
Q	soft	dinner	50	08:54
Q	soft	lunch	38	12:05
Q	hard	breakfast	21	08:14
Q	hard	dinner	54	15:39
Q	hard	dinner	42	12:45
Q	hard	lunch	39	10:55

Table S3 – Correlation matrix between the grid search, random search and Bayesian optimization accuracy and, total bites, total video duration, and number of videos per participant

	Grid Search	Random Search	Bayesian Optimization	Average Accuracy	Total Duration	Total Bites	Number of Videos
Grid Search	1.00	0.60	0.82	0.95	0.03	-0.15	-0.01
Random Search	0.60	1.00	0.41	0.75	0.15	0.10	0.13
Bayesian Optimization	0.82	0.41	1.00	0.88	0.03	-0.01	0.21
Average Accuracy	0.95	0.75	0.88	1.00	0.08	-0.03	0.13
Total Duration Seconds	0.03	0.15	0.03	0.08	1.00	0.71	0.78
Total Bites	-0.15	0.10	-0.01	-0.03	0.71	1.00	0.79
Number of Videos	-0.01	0.13	0.21	0.13	0.78	0.79	1.00

Video S1: A video demo of the Bitecounter software can be found at: https://youtu.be/f8TBIe4TJno

5

Automatic count of bites, chews, and sips from meal videos using vision transformers

Michele Tufano[1], Aneesh Chauhan[2], Edith J.M. Feskens[1], Guido Camps[1,3]

[1]Division of Human Nutrition and Health, Wageningen University & Research, Stippeneng 4, 6708 WE Wageningen, The Netherlands
[2]Wageningen Food and Biobased Research, Wageningen University & Research, Stippeneng 4, 6708 WE Wageningen, The Netherlands
[3]OnePlanet Research Center, Plus Ultra II, Bronland 10, 6708 WE Wageningen, The Netherlands
In preparation for submission

ABSTRACT

Eating behavior analysis requires labor-intensive manual video annotation, limiting large-scale dietary studies. We developed automated methods using transformer architectures to detect and quantify eating behaviors from meal videos. Using 103 annotated videos from 36 participants consuming standardized meals, we trained Vision Transformer (ViT) and CNN-LSTM models for bite, chew, and sip detection The ViT model achieved 98.45% frame-classification accuracy for bites under participant-independent testing, supporting per-participant bite-count accuracy of 86.2%. For chew detection, CNN-LSTM outperformed ViT (85.56% vs 69.21% window-classification accuracy). Chew counting reached 80% per-video count accuracy (r=0.86 against ground truth). Sip detection using pretrained VideoMAE reached 61% aggregate count accuracy, limited by dataset size and annotation challenges. Transformer architectures effectively captured global spatial relationships for bite detection through self-attention mechanisms, while sequential models better captured temporal dynamics for chewing behaviors. Performance degraded with downward head positions and varied eating utensils, highlighting requirements for diverse training data. These automated approaches enable replacement of manual annotation for large-scale eating behavior studies. Future work should expand datasets to include diverse eating scenarios and integrate object detection for improved drinking behavior quantification. This study demonstrates that transformer-based architectures can effectively automate eating behavior analysis, supporting future eating behavior interventions and prevention strategies.

INTRODUCTION

Eating behavior is defined by food choices, eating habits, and oral processing actions such as biting, chewing, and swallowing, which overall determine our nutritional intake. This behavior is influenced throughout life by the surrounding food environment, interactions between parents and children, peer influences, and the textural properties of food [1–3]. An important element of eating behavior is the individual's eating rate, which measures the amount of food consumed in grams per minute (gr/min). Studies have shown that eating rate affects food intake [4] ,energy intake [5], and weight gain [6,7]. Specifically, a faster eating rate can increase the risk of obesity [8] and metabolic diseases [9,10]. Therefore, it is crucial to measure individual eating behavior to support interventions aimed at decreasing eating rates and reducing obesity risk [11,12].

Video recordings of meal episodes are a critical, non-invasive method to measure individual eating behavior. Currently, the golden standard consists of two trained researchers who manually review the videos to annotate each eating event (i.e., bite, chew, sip), then cross-check their results for consistency [13]. This manual process is not only repetitive, but also labor- and time-intensive, requires training, and is susceptible to subjectivity and lapses in attention [14]. Consequently, the current approach is impractical for large-scale studies or real-time feedback on individual eating behaviors. Hence, to advance research in eating behavior and facilitate studies with larger cohorts, it is essential to automate the video analysis process using novel technologies such as facial landmarks and deep learning [15].

Deep learning has revolutionized the automation of eating behavior detection from video data, significantly outperforming manual methods. Recent research has applied these methodologies across multiple eating activities, with Faster R-CNN and AlexNet combinations achieving over 85% accuracy for bite and chew detection[16] , while the CNN-LSTM RABiD algorithm shows near-perfect agreement with human annotations for measuring meal duration (κ = 0.894; F1-score: 0.948)[17]. For broader intake gesture recognition, neural network approaches (CNN-LSTM, 3D CNN, SlowFast) have achieved F1 scores exceeding 0.85, underscoring the importance of temporal context for optimal model performance[18]. While CNN and similar models have dominated research in this area, transformers architecture has not yet been applied to tracking eating and drinking gestures, despite its proven success in other video classification tasks.

The advent of transformer architecture, has significantly advanced the field of video classification, offering several benefits over traditional convolutional neu-

ral networks (CNNs). Transformers excel in capturing long-range dependencies and global context, which are critical for understanding temporal dynamics in videos [19, 20]. This ability to handle sequential data with high flexibility has led to improved performance in video classification tasks [21]. Despite these improvements, transformers require higher computational costs and the need for larger datasets to achieve optimal performance [22].

Previous studies on video-based automatic eating behavior analysis have acknowledged the temporal dimension of eating episodes but have not utilized transformer architectures, which excel at processing temporal and sequential data. This study investigates transformer models with temporal dependencies for counting bites, chews, and sips in meal videos.

METHODS

Study participants

The study took place between April and July 2022 at the Human Nutrition Research Unit, Wageningen University and Research, The Netherlands, and was approved by the Social Ethical Committee of Wageningen University (Lasschuijt, 2022-03). Participants, aged 18-55 with a BMI between 18.5 and 30 kg/m², were recruited from Wageningen University's Division of Human Nutrition volunteer database, supplemented by social media advertisements. After an online information session, potential participants provided digital informed consent and completed an online inclusion and exclusion questionnaire. The inclusion criteria required participants to be healthy (self-reported), English-speaking, and typically eat three meals a day at regular times. Exclusion criteria included having a BMI below 18.5 or above 30 kg/m², food allergies to any study foods, appetite, chewing, or swallowing issues, following a calorie-restricted diet, significant weight changes (>5 kg) in the past two months, excessive alcohol consumption (more than 21 drinks per week), taking medications affecting appetite, taste, or smell, engaging in more than 8 hours of intense exercise per week, or being a highly restrained eater according to the Dutch Eating Behaviour Questionnaire (DEBQ) with cut-off scores above 2.89 for men and 3.39 for women. Personnel and Master's students from the Division of Human Nutrition and Health were excluded from participating. Out of 137 invited individuals, 77 were eligible, and 69 completed all study sessions. Participants were financially compensated and were instructed to eat until comfortably full.

Table 1 - Participant Recruitment and Eligibility Summary

Individuals invited	137
Excluded (did not meet inclusion criteria)	60
Eligible participants	77
Participants who completed all study sessions	69
Age range (years)	18–55
BMI range (kg/m²)	18.5–30
Financial compensation provided	Yes

Meals and conditions

The original study used a 2×2 randomized controlled crossover design varying eating rate and energy density, with food and energy intake as primary outcomes[23]. The present analysis uses the meal-video recordings collected during the original study to develop automated detection of bites, chews, and sips, which were annotated as secondary outcomes of the original protocol.Eating rates were modified through varying food textures that required different levels of oral processing, while energy densities were adjusted using different sandwich ingredients. All meals comprised commercially available items, including bread buns, chicken, cheese, mayonnaise, carrot, lettuce, and butter. Meals were served in large portions, about 150% of the average size (~900 grams). The texture of the meals was varied by using different types of bread buns (soft-fast, average, hard-slow) and adjusting the size of raw carrot pieces in the sandwiches. Each meal was served with a glass of water (150 g) and a small side salad (60 g).

Video recordings

Participants were recorded during their 4 meals using a Logitech C310 Webcam positioned approximately 1.5 meters in front of them. The camera was configured to capture the area from the table surface to above the participant's head, with the frame width aligned to their shoulder width. Videos were recorded with a software build to record videos using OpenCV (version 4.12.0.88)[24] Windows 10 laptops (Lenovo Thinkpad L380, Intel Core i5). To facilitate video analysis, participants were instructed to hold a numbered card in view of the camera before eating and to raise a hand when they finished. These recordings were later used for manual annotation and analysis. Because of recording errors, 18 videos are missing, so the eating behaviors during those meals could not be annotated.

Manual annotation

The video recordings were analyzed using ELAN Software version 6.4 (Max Planck Institute for Psycholinguistics) [25]. Two human annotators reviewed the videos to document specific eating behaviors, including meal duration (min), time interval between bites (s), oro-sensory exposure duration (s), number of chews, number of sips, and number of bites per meal. Bite size was determined by dividing the total food intake (in grams) by the number of bites, while eating rate (g/min) was calculated by dividing the food consumed (g) by the meal duration (min). Of the 276 videos from 69 participants who completed all study sessions, 103 videos from 36 participants were annotated by a single annotator and used for further processing to ensure consistency in annotation and data extraction The annotation files from ELAN were exported to JSON files. A summary of the observed eating behavior can be found in Table 2.

Table 2 - Summary of observed eating behaviors across 104 video recordings from 36 participants. Data show considerable inter-individual variation in all measured behaviors, with chewing frequency showing the greatest absolute variability.

Eating Behavior	Mean ± SD	5th-95th Percentile
Total Chews	917.2 ± 309.3	494-1465
Total Sips	8.9 ± 4.0	4-18
Total Bites	43.2 ± 14.4	25-73

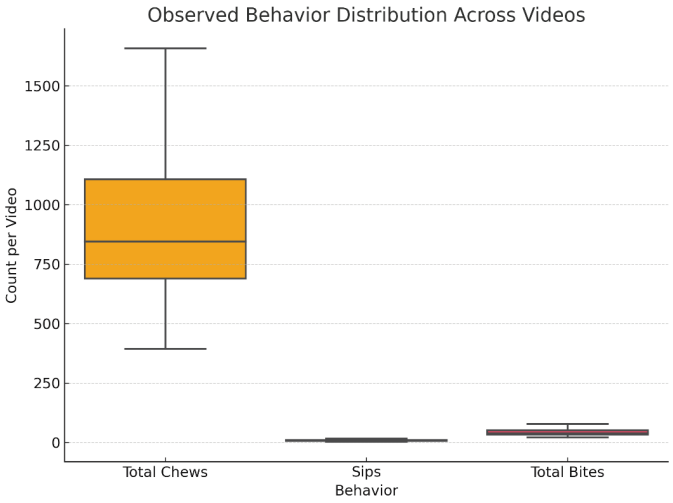


Figure 1: Distribution of observed eating behaviors across all videos. Boxplots show the per-video distribution of total *chews*, *sips*, and *bites*. The central line indicates the median; boxes represent the interquartile range.

Video pre-processing

The video analysis involved processing a set of MP4 video files and corresponding JSON annotations to extract specific events, such as chewing, biting, and drinking, along with background activity. The files were organized, linking each video with its annotation file. For each video-annotation pair, labelled events corresponding to bites, chews, and sips were identified, and relevant video frames were extracted using OpenCV[24], converted to RGB format, and compiled into GIFs using imageio[26], with a maximum of 100 GIFs per event per video to ensure dataset balance. Background activity was also captured by generating GIFs from non-annotated video segments. All GIFs were subsequently converted to MP4 format using moviepy`[27]. Finally, participants were randomly split into training and testing sets with an 80/20 split, ensuring participant-level independence, and files were moved into corresponding directories for further analysis.

The video analysis to achieve the final dataset was performed using Windows 10 Pro on a Lenovo laptop (00330-80000-00000-AA904) with 12th Gen Intel(R) Core(TM) i9-12950HX (2.30 GHz) and 32 GB of RAM. The programming language used for video analysis was Python 3.8 [28] with Visual Studio Code version 1.92.2 as IDE [29].

Performance Metrics

We report two families of metrics. Classification accuracy is the proportion of correctly labeled frames (bite ViT, chew ViT) or sequence windows (chew CNN-LSTM) on the held-out test set, with participants disjoint from training. Count accuracy compares the algorithm's count of events (bites, chews, or sips) per video against the manual annotation, defined as 1 − |predicted − ground truth| / ground truth. Count accuracy is reported at two levels: per-video (averaged over all videos) and per-participant (averaged within each participant before averaging across participants).

Bites Dataset

For bite detection, a binary classification dataset was created using frame extraction from videos. Source videos were organized into train/test sets with ‘background’ and ‘bites’ classes. From each video, frames were randomly sampled (two per video) using OpenCV, with each frame saved as a JPEG with the original filename and frame index appended. All extracted frames were visually examined to verify correct classification of bites and background.

The training set comprised 8458 background images and 5490 bite-class images from 29 participants. The test set contained 1481 background images and

777 bite-class images from 7 different participants. Test set participants were entirely separate from those in the training set to evaluate the model's generalization capabilities across different individuals.

Training Bite Classifier

A Vision Transformer (ViT) model was trained for binary bite classification using a transfer learning approach. A pre-trained "google/vit-base-patch16-224" model was fine-tuned with appropriate label mappings for 'background' and 'bites' classes. Image preprocessing included data augmentation for training samples (random resized crops to 224×224 pixels and horizontal flips) and deterministic resizing for test samples, both followed by normalization using model-specific parameters. Training utilized the HuggingFace Transformers python library with 20 training epochs, batch size of 64, gradient accumulation steps of 4, learning rate of $5e\text{-}^{5}$, and 10% warmup ratio. Model performance was evaluated after each epoch using accuracy metrics, with automatic retention of the best-performing checkpoint.

Hyperparameter optimization for the Vision Transformer (ViT) bite classifier was performed using Optuna framework. The search space included learning rates (10^-6 to 10^-4), batch sizes (16, 32, 64), and weight decay values (10^-6 to 10^-3). Ten optimization trials were conducted using a 5-epoch training cycle per configuration with cosine annealing learning rate scheduling. Image preprocessing included standard data augmentation techniques and ImageNet normalization parameters. The final model was trained for 20 epochs using the optimal hyperparameters, with model checkpoints saved based on validation accuracy. GPU acceleration was utilized when available to maximize computational efficiency.

Chews Dataset

Video recordings were collected from participants in controlled eating sessions as described above. Raw video data was paired with corresponding JSON annotation files containing time-stamped labels for feeding behaviors (bites, chews, and sips). A custom extraction pipeline processed these pairs to generate annotated video sequences of at least 10 seconds duration, with each sequence capturing complete feeding events.

The dataset generation process prioritized sequences starting with bites or sips, extending each segment to include subsequent feeding behaviors until reaching minimum 10-second duration. For frame-level labeling, when multiple behaviors occurred simultaneously, a hierarchical priority system was applied (bite > sip > chew > background) to assign a single dominant label per frame. All

sequences underwent verification to ensure frame-count consistency between video files and corresponding label files.

Video recordings were processed using our custom LandmarkAnalysis module. The analysis pipeline converts input frames to RGB color space and extracts landmark coordinates through MediaPipe's holistic framework with detection and tracking confidence thresholds set at 0.5. The system processes either individual video files or batches from specified directories, with support for common video formats.

Multiple biomechanical metrics were computed from the detected landmarks:

- *Oral-facial measurements* included mouth aperture, lip curvature, and mouth-to-chin ratio calculated using specific landmark coordinates. Mouth movement was quantified through horizontal-to-vertical ratio calculations in both 2D and 3D space.
- *Jaw metrics* comprised jaw opening, rotation angle, and excursion determined using spatial relationships between jaw landmarks relative to reference facial points.
- *Hand proximity* analysis calculated the Euclidean distance between hand landmarks and oral regions to quantify hand-to-mouth proximity.
- *Temporal dynamics* measured frame-to-frame displacements of landmarks to track movement velocities and trajectories.

Time series data from landmark measurements underwent Savitzky-Golay filtering (window length=7, polynomial order=3) to reduce noise while preserving signal characteristics. Movement peaks were identified using combined first and second derivatives, with prominence threshold of 0.5 and minimum width of 3 frames.

Extracted landmark metrics were transformed into multi-channel feature images suitable for computer vision algorithms. Raw and filtered landmark metrics were normalized and encoded as separate RGB channels (filtered metrics, raw metrics, and peak indicators). This representation captures both temporal dynamics and spatial relationships within fixed-size windows of 64 frames. A class-balanced sampling approach was implemented to generate equal proportions of feeding behavior classes, with labels consolidated into "chew" and "non-chew" categories. The feature images underwent normalization and dimensionality standardization to 224×224 pixels to maintain compatibility with standard vision model architectures.

The code was implemented in Python 3.8 with CUDA acceleration[30] for GPU-compatible operations. Key dependencies included MediaPipe[31] for landmark detection, OpenCV[24] for image processing, NumPy[32] for numerical operations, and PyTorch[33] for video handling.

Training Chew Classifier

We implemented a Vision Transformer (ViT) architecture for classifying chewing behaviors from the generated feature images. The model utilized a pretrained ViT "google/vit-base-patch16-224" with patch size 16×16 and base configuration, fine-tuned on our domain-specific data. The dataset was partitioned into training (60%), validation (20%), and test (20%) sets with stratified sampling to maintain class balance. This 60/20/20 split was chosen to enable effective hyperparameter tuning via the validation set, as the chew classification task involved higher intra-class variability and greater risk of overfitting. Data augmentation techniques including random horizontal flips, rotations, and color jittering were applied during training to enhance model generalization. Training used the Adam optimizer with weight decay (1e-4) and cosine annealing learning rate scheduler, minimizing cross-entropy loss over 100 epochs with early stopping based on validation performance.

We also developed a spatiotemporal architecture combining convolutional and recurrent components. The CNN feature extractor comprised four convolutional blocks with batch normalization, ReLU activation, and progressive feature map depth (32-64-128-128 channels). The LSTM component consisted of two stacked bidirectional layers with 128 hidden units, enabling sequential processing of CNN-extracted features. The model included multiple dropout layers (0.25-0.5) and was trained using Adam optimization with learning rate reduction on validation loss plateau.

Hyperparameter optimization for the CNN-LSTM model was conducted using Optuna with Tree-structured Parzen estimator (TPE) sampling across 20 trials. The search space included batch sizes (8, 16, 32), learning rates (10^-5 to 5×10^-4), weight decay (10^-6 to 10^-3), model architecture variations, and training epochs (30-100). Each configuration was evaluated based on maximum validation accuracy. The process incorporated dataset validation, fixed random seeds for reproducibility, and GPU acceleration. Best parameters were systematically recorded with optimization history and parameter importance visualizations generated for analysis.

Sip detection

We implemented a sip detection system using a transformer-based approach with the VideoMAE architecture pre-trained on Kinetics. The model processes video frames at 3 FPS temporal overlap between consecutive windows. Frames undergo standard preprocessing with the VideoMAEImageProcessor and are analyzed using the VideoMAEForVideoClassification model, which employs half-precision floating-point (FP16) for efficient GPU utilization. A targeted whitelist approach identifies drinking activities as four specific class indices (100, 101, 102, 352, corresponding respectively to drinking, drinking beer, drinking shots, tasting beer) from the Kinetics taxonomy[34].

Counting bites

A bite counting algorithm was implemented to process videos using the trained Vision Transformer model. The system extracts frames at 10 FPS and generates prediction logits for each frame. Multiple temporal smoothing windows (10-40 frames) were tested to optimize detection robustness. The algorithm applies sliding window averaging over prediction logits, followed by SoftMax transformation to generate class probabilities. Bite events are identified through threshold-based binarization (threshold=0.5) and detection of 0-to-1 transitions in the probability time series. Results for each smoothing window configuration were systematically recorded, with bite counts for all processed videos exported to Excel for comparative analysis. The implementation leveraged GPU acceleration when available and included efficient tensor operations for processing long video sequences.

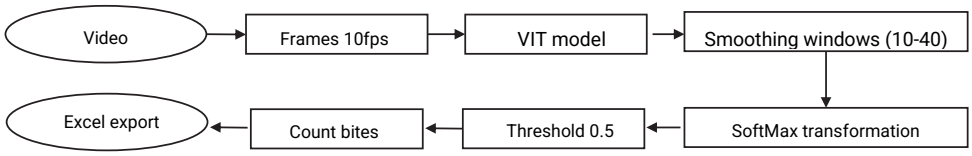


Figure 2: Flowchart of the bite counting algorithm

Counting chews

A chew counting algorithm was developed based on peak detection and signal processing techniques applied to CNN-LSTM model outputs. The system processes video frames at 10 FPS and generates classification logits for each sequence window. A Savitzky-Golay filter was applied to the chew probability signal for noise reduction, with adaptive polynomial order selection based on sequence length. Multiple smoothing windows (25, 55, 85 frames) were evaluated to optimize detection robustness. Peak detection utilized prominence-based thresholding with distance constraints to identify individual chewing events.

We implemented parameter optimization using Optuna[35] with TPE sampling across 100 trials to tune five critical parameters: prominence factor (0.01-0.5), minimum peak distance (1-15 frames), peak width constraints (0.5-20.0), and calibration factor (0.5-10.0). The objective function minimized mean absolute error between predicted and ground-truth chew counts across a validation set of 50 manually annotated videos. Results were systematically recorded with performance metrics including mean absolute error, mean percentage error, and per-video deviation statistics. The implementation incorporated memory optimization techniques, parallel processing, and GPU acceleration for handling extended video sequences efficiently.

Counting sips

To detect individual sip events in the continuous frame-wise predictions, we developed a post-processing algorithm with temporal grouping and threshold-based classification. The system processes prediction CSV files containing timestamps, class labels, and confidence scores for each analyzed frame. First, frame-level predictions are filtered to retain only drinking-related classes with confidence scores exceeding a minimum threshold. These filtered frames are then temporally sorted and grouped into discrete sipping events using a temporal continuity approach.

Continuous frame sequences are considered part of the same sipping event when the temporal gap between consecutive drinking frames is below a configurable maximum gap threshold. A drinking sequence is classified as a valid sip only when its duration exceeds a minimum frame count threshold, effectively filtering out brief motion artifacts or detection noise.

We implemented a two-stage parameter optimization process. Initially, we performed an exhaustive grid search across 175 parameter combinations (confidence threshold: 0.1-0.7, minimum sip duration: 1-7 frames, maximum gap: 1-10 seconds). This preliminary optimization identified promising parameter regions. Subsequently, we refined these parameters using Optuna with TPE sampler across 50 trials, focusing on narrower parameter ranges (confidence threshold: 0.15-0.35, minimum sip duration: 6-12 frames, maximum gap: 6.0-11.0 seconds).

Performance was evaluated against a manually annotated ground truth dataset of 105 eating episodes with precise sip counts. We utilized cross-validation when sufficient ground truth data was available, with the objective function maximizing accuracy, defined as one minus the ratio of mean absolute error to the average number of ground truth sips.

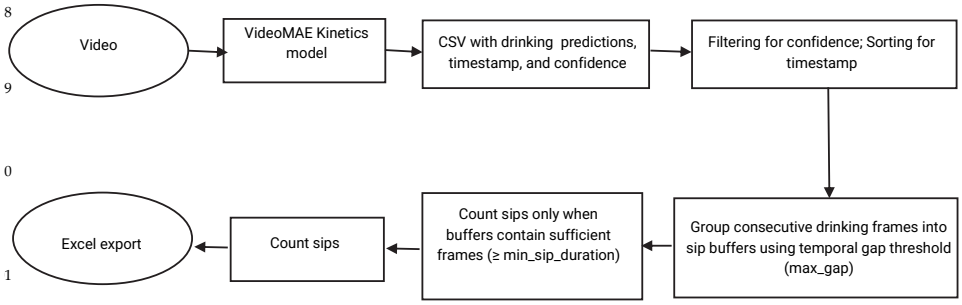


Figure 3: Flowchart of the sip counting algorithm

RESULTS

Bite detection model

The Vision Transformer (ViT) model trained for binary bite classification distinguished bite from background events with 98.05% accuracy. Using a participant-independent evaluation approach with separate individuals in the test set (n=7) from those in the training set (n=29), the model demonstrated generalization capabilities. Hyperparameter optimization through Optuna identified optimal parameters of learning rate=6.81×10^{-6}, batch size=16, and weight decay=9.63×10^{-5}, yielding 99.11% validation accuracy (Table 1). Examples of background and bite event frames from the training dataset are shown below (Figure 4 and 5, respectively). On the test dataset, the model maintained 98.05% accuracy by epoch 1, improving to 98.45% by epoch 2, with training accuracy reaching 96.94%. This rapid convergence indicates the model captured essential features for bite classification with minimal training.

Table 1: Performance metrics of the bite detection model

Metric	Value
Test Accuracy	98.45%
Validation Loss	0.0473
Training Loss	0.0825
Best Hyperparameters	Learning rate: 6.81×10^{-6}, Batch size: 16, Weight decay: 9.63×10^{-5}
Dataset Size	13,948 training images (8,458 background, 5,490 bite-class), 29 participants
Test Set Size	2,258 images (1,481 background, 777 bite-class), 7 participants



Figure 4: Representative background event frames from study participants. Images masked to preserve participant anonymity.



Figure 5: Representative bite event frames from study participants. Images masked to preserve participant anonymity.

Bite count

The bite-counting algorithm achieved per-participant count accuracy of 86.2% (n=36). With the 25-frame smoothing window, performance was: mean error 2.93, MAE 10.23, MSE 216.26, RMSE 14.71, mean percentage error 24.68%, and per-video count accuracy of 80.39% across 102 videos. At the participant level,

37.8% of participants exceeded 90% count accuracy with the original parameters. Per-participant count accuracy degraded with longer windows (55-frame: 63.2%; 85-frame: 53.2%; 100-frame: 49.9%).

Chew Detection Performance

The initial Vision Transformer (ViT) model trained for chew classification achieved modest performance with 69.21% window-classification accuracy on the test dataset. Despite using the same participant-independent evaluation approach that was successful for bite classification, the ViT model demonstrated limited effectiveness in distinguishing chewing from non-chewing events. The confusion matrix revealed balanced precision and recall metrics (0.71 for chew class, 0.68 for non-chew class), but overall performance remained insufficient for reliable chew detection (Figure 6). This prompted a transition to an alternative architecture. The CNN-LSTM model, optimized through Optuna trials, demonstrated significantly improved performance with 85.56% test window-classificationaccuracy. Hyperparameter optimization identified optimal parameters of learning rate=1.10×10^{-4}, batch size=16, epochs=50, and weight decay=2.62×10^{-6}, yielding improved classification metrics (Table 3). The CNN-LSTM architecture effectively captured both spatial and temporal features essential for chew detection, with precision of 0.86 and recall of 0.88 for the chew class, and precision of 0.85 and recall of 0.83 for the non-chew class (Figure 7).

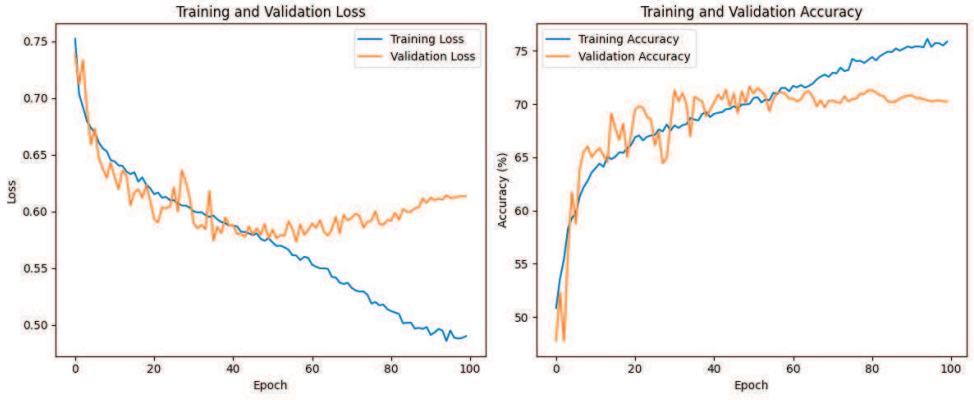


Figure 6 – Training and Validation Loss and Accuracy curves for the ViT-based chew detection model (best model at epoch 50).

Table 2: Performance metrics of the initial ViT-based chew detection model

Metric	Value
Test window classification Accuracy	69.21%

Table 2: Performance metrics of the initial ViT-based chew detection model (*continued*)

Metric	Value
Precision (Chew/Non-chew)	0.71/0.68
Recall (Chew/Non-chew)	0.71/0.68
F1-score (Chew/Non-chew)	0.71/0.68
Dataset Size	1,910 test images (998 chew, 912 non-chew)
Model Architecture	Vision Transformer (ViT) with image-based feature extraction

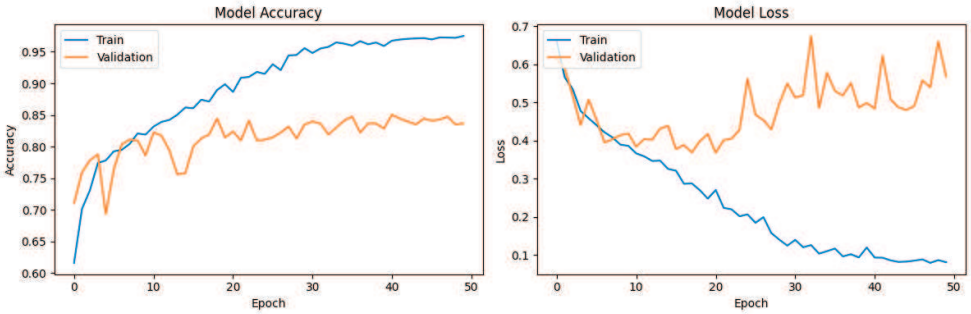


Figure 7 – Training and Validation Loss and Accuracy curves for the CNN-LSTM chew predictor over 50 epochs (peak validation accuracy: 85.06% at epoch 41).

Table 3: Performance metrics of the CNN-LSTM chew detection model

Metric	Value
Test window classification accuracy	85.56%
Precision (Chew/Non-chew)	0.86/0.85
Recall (Chew/Non-chew)	0.88/0.83
F1-score (Chew/Non-chew)	0.87/0.84
Best Hyperparameters	Learning rate: 1.10×10^{-4}, Batch size: 16, Epochs: 50, Weight decay: 2.62×10^{-6}
Dataset Composition	Multiple participants with class-balanced sampling of chew and non-chew events (350 chew, 287 non-chew in test set)
Model Architecture	CNN-LSTM with sequence-based temporal features

Sips count

The sip counting algorithm achieved 61.48%aggregated count accuracy through Optuna hyperparameter optimization across 102 videos. Optimal

parameters were identified as: confidence threshold=0.29, minimum sip duration=10 frames, and maximum temporal gap=6.0 seconds between consecutive drinking frames. The algorithm employed pretrained VideoMAE-based activity recognition to detect drinking-related frames, followed by temporal segmentation that grouped consecutive detections within the gap threshold (maximum_gap). Sips were validated only when frame groups exceeded the minimum duration requirement (min_sip_duration). The modest accuracy reflects the inherent complexity of collecting data and distinguishing discrete sipping events from continuous drinking behavior in video.

DISCUSSION

This study demonstrated the effectiveness of transformers architectures for automated detection and quantification of eating behaviors from video recordings. The Vision Transformer (ViT) architecture achieved 98.45% frame-classification accuracy performance in bite detection, while a CNN-LSTM approach demonstrated 85.56% window-classification accuracy for chew detection. These classifiers supported event counting with per-participant count accuracy of 86.2% for bites and per-video count accuracy of 80% for chews

The sip detection algorithm achieved 61% accuracy using VideoMAE-based activity recognition with temporal segmentation parameters optimized through Optuna.

The participant-independent testing approach confirms the ViT model successfully transfers learned features across individuals, maintaining high accuracy despite variations in eating behaviors and physical characteristics. This strong generalization capacity indicates the model captures invariant features of bite gestures while remaining robust to individual differences, a crucial requirement for automatic dietary monitoring applications. The use of transformer architecture for bite detection represents a methodological advancement over previous CNN-based approaches, leveraging the self-attention mechanism's ability to process global context within consecutive frames.

Our results extend previous work in automated eating behavior detection. While earlier studies utilizing Faster R-CNN and AlexNet combinations achieved over 85% accuracy for bite detection, our ViT-based approach demonstrated superiorframe-classification accuracy performance (98.45%). Similarly, our CNN-LSTM model for chew detection (85.56% window-classification accuracy) compares favorably with prior studies, though slightly below the CNN-LSTM

RABiD algorithm's reported near-perfect agreement with human annotations (κ = 0.894; F1-score: 0.948). These comparisons suggest that both transformer-based and hybrid CNN-LSTM architectures offer viable approaches for automated eating behavior analysis.

The superior performance of the ViT model for bite detection likely stems from its ability to process global spatial relationships within frames through self-attention mechanisms. This aligns with the visual nature of bite detection, where contextual understanding of hand-to-mouth relationships provides strong discriminative features. In contrast, the relatively modest performance of the ViT model for chew detection (69.21%) compared to the CNN-LSTM architecture (85.56%) underscores that temporal dynamics, when properly modeled with sequence-based architectures, are crucial for accurate chew detection. Chewing represents a rhythmic behavior that requires temporal context to distinguish from other facial movements.

The correlation between predicted and actual chew counts (r=0.86) demonstrates that despite the complexity of chewing behaviors, automated quantification approaches can achieve clinical utility for measuring eating behavior. The observed higher accuracy for moderate chewing rates (600-1000 chews) suggests potential calibration issues at extreme values that may require further refinement.

Several limitations merit consideration. First, insufficient sip detection data prevented robust model development for this behavior. The drinking detection approach utilized pre-trained VideoMAE classification models with minimal fine-tuning, potentially explaining reduced performance relative to bite and chew detection. The pretrained model provided drinking frame identification, but temporal grouping parameters (confidence threshold=0.29, minimum duration=10 frames, maximum gap=6.0 seconds) required extensive optimization across participant clusters.

Second, accurate ground truth annotation for chew events proved challenging due to subtle jaw movements and video occlusion. This annotation difficulty affected model training, contributing to lower chew detection accuracy compared to bite detection. Bite detection benefits clearer visual cues (hand-mouth proximity, mouth opening), while chewing presents fewer distinguishable frames and subtle inter-frame jaw position variations. Alternative annotation tools optimized for deep learning training (e.g., Darwin[36] or CVAT[37]) may provide superior precision for capturing correct chewing frames compared to ELAN, which targets behavioral quantification.

Third, transformer models require substantial computational resources for extended video sequences, limiting real-time deployment on consumer hardware without optimization. While participant-independent testing demonstrated acceptable generalization, applicability remains restricted due to dataset limitations. Training data consisted of participants eating salads with forks and sandwiches by hand, excluding chopsticks. This limitation affects generalizability, particularly since chopsticks tend to lean toward plates, facing downward rather than toward cameras, increasing occlusion. Model accuracy drops when participants look down, especially chewing detection, as optimal performance requires camera-facing orientation. These constraints highlight requirements for expanded datasets, improved annotation methodologies, and computational optimization for practical deployment.

Despite these technical and data-related limitations, transformer architectures are a promising technology to achieve automated eating behavior analysis. Successful automation of eating behavior detection enables large-scale dietary pattern analysis and targeted consumption reduction interventions. Real-time feedback systems support behavioral modification strategies for obesity prevention through immediate intervention delivery via lightweight edge computing architectures.

Current limitations highlight the need for targeted data collection to improve drinking behavior analysis. Key development goals include integrating YOLO (You Only Look Once)-based cup detection to enhance sip quantification and employing Vision-Language Transformers for more robust activity recognition. These advancements aim to increase detection accuracy across diverse consumption scenarios and behavioral patterns.

CONCLUSION

This study demonstrates that transformer-based and temporal sequence models can effectively automate the detection and quantification of eating behaviors from video recordings. The high accuracy achieved for participant-independent bite detection represents a significant advance toward practical applications in dietary monitoring. Despite challenges in chew and sip detection, the overall performance suggests these approaches could replace labor-intensive manual annotation processes, enabling large-scale eating behavior studies and supporting future interventions.

REFERENCES

1. Gahagan, S. Development of Eating Behavior: Biology and Context. *J. Dev. Behav. Pediatr.* **33**, 261 (2012).
2. Forde, C. G. & de Graaf, K. Influence of Sensory Properties in Moderating Eating Behaviors and Food Intake. *Front. Nutr.* **9**, 841444 (2022).
3. Forde, C. G. & Bolhuis, D. Interrelations Between Food Form, Texture, and Matrix Influence Energy Intake and Metabolic Responses. *Curr. Nutr. Rep.* **11**, 124–132 (2022).
4. Viskaal-van Dongen, M., Kok, F. J. & de Graaf, C. Eating rate of commonly consumed foods promotes food and energy intake. *Appetite* **56**, 25–31 (2011).
5. Robinson, E. *et al.* A systematic review and meta-analysis examining the effect of eating rate on energy intake and hunger. *Am. J. Clin. Nutr.* **100**, 123–151 (2014).
6. Fogel, A. *et al.* Eating behaviors moderate the associations between risk factors in the first 1000 days and adiposity outcomes at 6 years of age. *Am. J. Clin. Nutr.* **111**, 997–1006 (2020).
7. van den Boer, J. H. W. *et al.* Self-reported eating rate is associated with weight status in a Dutch population: a validation study and a cross-sectional study. *Int. J. Behav. Nutr. Phys. Act.* **14**, 121 (2017).
8. Ohkuma, T. *et al.* Association between eating rate and obesity: a systematic review and meta-analysis. *Int. J. Obes. 2005* **39**, 1589–1596 (2015).
9. Zhu, B., Haruyama, Y., Muto, T. & Yamazaki, T. Association between eating speed and metabolic syndrome in a three-year population-based cohort study. *J. Epidemiol.* **25**, 332–336 (2015).
10. Sakurai, M. *et al.* Self-reported speed of eating and 7-year risk of type 2 diabetes mellitus in middle-aged Japanese men. *Metabolism.* **61**, 1566–1571 (2012).
11. Abidin, N. Z. *et al.* Combating obesity through healthy eating behavior: a call for system dynamics optimization. *PloS One* **9**, e114135 (2014).
12. Shavit, Y., Roth, Y. & Teodorescu, K. Promoting Healthy Eating Behaviors by Incentivizing Exploration of Healthy Alternatives. *Front. Nutr.* **8**, 658793 (2021).
13. Forde, C. G., Leong, C., Chia-Ming, E. & McCrickerd, K. Fast or slow-foods? Describing natural variations in oral processing characteristics across a wide range of Asian foods. *Food Funct.* **8**, 595–606 (2017).
14. 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**, 209 (2020).
15. Tufano, M., Lasschuijt, M., Chauhan, A., Feskens, E. J. M. & Camps, G. Capturing Eating Behavior from Video Analysis: A Systematic Review. *Nutrients* **14**, 4847 (2022).
16. Hossain, D., Ghosh, T. & Sazonov, E. Automatic Count of Bites and Chews From Videos of Eating Episodes. *IEEE Access* **8**, 101934–101945 (2020).
17. 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**, 209 (2020).
18. Rouast, P. V. & Adam, M. T. P. Learning Deep Representations for Video-Based Intake Gesture Detection. *IEEE J. Biomed. Health Inform.* **24**, 1727–1737 (2020).
19. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).

20. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
21. Arnab, A. *et al.* ViViT: A Video Vision Transformer. Preprint at https://doi.org/10.48550/arXiv.2103.15691 (2021).
22. Khan, S. *et al.* Transformers in Vision: A Survey. *ACM Comput. Surv.* **54**, 1–41 (2022).
23. Lasschuijt, M. P. *et al.* The Effect of Energy Density and Eating Rate on *Ad Libitum* Energy Intake in Healthy Adults—A Randomized Controlled Study. *J. Nutr.* **155**, 2602–2610 (2025).
24. opencv-python: Wrapper package for OpenCV python bindings.
25. ELAN | The Language Archive. https://archive.mpi.nl/tla/elan.
26. imageio: Library for reading and writing a wide range of image, video, scientific, and volumetric data formats.
27. moviepy: Video editing with Python.
28. Python Release Python 3.8.0. *Python.org* https://www.python.org/downloads/release/python-380/.
29. July 2024 (version 1.92). https://code.visualstudio.com/updates/v1_92.
30. CUDA Toolkit 12.1 Downloads. *NVIDIA Developer* https://developer.nvidia.com/cuda-downloads.
31. MediaPipe Solutions guide | Google AI Edge. *Google AI for Developers* https://ai.google.dev/edge/mediapipe/solutions/guide.
32. NumPy. https://numpy.org/.
33. PyTorch. *PyTorch* https://pytorch.org/.
34. Kay, W. *et al.* The Kinetics Human Action Video Dataset. Preprint at https://doi.org/10.48550/arXiv.1705.06950 (2017).
35. Optuna - A hyperparameter optimization framework. *Optuna* https://optuna.org/.
36. V7 Darwin | AI Data Labeling & ML Training Data Platform. https://www.v7labs.com/darwin.
37. Leading Image & Video Data Annotation Platform | CVAT. *CVAT.ai* https://www.cvat.ai.

6

A Cloud-Based Toolkit for Automated Meal Video Analysis: Enabling Large-Scale Eating Behavior Research

Michele Tufano[1*], Simon Hengeveld[1], Marlou Lasschuijt[1], Edith J.M. Feskens[1], Guido Camps[1]

[1] Division of Human Nutrition and Health, Wageningen University & Research
In preparation for submission

ABSTRACT

Manual annotation of eating behavior from video recordings remains labor-intensive, time-consuming, and susceptible to observer bias, limiting scalable eating behavior research applications.

We present the Automated Meal Video Analysis (AMVA) Toolkit, an open-source cloud-native platform designed for the research community to automate video-based eating behavior analysis. This publicly available system leverages AWS serverless architecture with transformer-based computer vision models for bite and chew detection, integrated with privacy-preserving facial masking capabilities. The containerized implementation utilizes ECS Fargate with automatic resource scaling and GDPR-compliant data handling.

Performance evaluation across 12 real-world meal videos demonstrated processing times scaling linearly with video duration (3 seconds to 26 minutes). The system achieved 98% accuracy in bite frame prediction and 86% in bite counting, with processing rates of 30-40 milliseconds per frame on 16 vCPU containers. Manual annotation time decreased from six weeks to six hours for 118 videos, representing a 40-fold efficiency improvement.

We provide AMVA Toolkit as a freely accessible platform enabling researchers to implement scalable, objective eating behavior quantification through automated analysis. The toolkit eliminates manual annotation dependencies while maintaining measurement precision, supporting standardized output formats for both retrospective analyses and prospective study designs in eating behavior research.

BACKGROUND

The manual annotation of eating behavior from video recordings remains a critical yet labor-intensive task in nutrition research[1,2]. Researchers must review footage of meal videos frame by frame to quantify the number of bites, chews, and sips, a process that is not only time-consuming but also susceptible to inconsistencies and observer bias[3]. Annotation software such as ELAN[4] can facilitate this task, but it still requires trained researchers and adherence to privacy agreements. Furthermore, existing methodologies do not support prospective study designs due to the substantial data volumes requiring extensive processing time, precluding dynamic study condition modifications and longitudinal participant tracking, while introducing significant temporal delays between data collection and analytical output delivery.

Automated systems to track eating behavior enable scalable analysis of large cohorts with objective, standardized measurements that eliminate inter-rater variability and human bias[5]. These systems provide real-time parameter extraction, implicit behavioral monitoring, and simultaneous quantification of multiple variables (intake events, chewing frequency, mastication duration) [6,7]. Operational advantages include batch processing capabilities, cost-effective implementation through reduced labor requirements, and millisecond-level temporal precision with standardized output formats compatible with statistical analysis pipelines.

Several automated approaches have been developed for oral processing behavior assessment from recorded videos. FaceReader (Noldus Information Technology)[8] represents the primary commercial solution, employing facial expression analysis to automatically extract chewing parameters including intake timing, chewing motions, and mastication duration[7,9]. However, the system demonstrates systematic measurement underestimation and reduced accuracy when deployed from controlled laboratory settings to real-world environments[7]. Current experimental status necessitates extensive per-individual calibration, with optimal performance restricted to trained panels rather than general consumer populations, creating significant scalability constraints for population-level dietary monitoring applications.

AMVA is a novel cloud-based toolkit developed to automate video-based analysis of eating behavior. The platform is scientifically validated through peer-reviewed publication of its underlying models. It provides estimates of bites, chews, and sips per video, along with their temporal distributions. Users can process individual videos or analyze entire datasets. By automating event de-

tection, AMVA improves both the efficiency and reproducibility of eating behavior analysis, reducing reliance on manual annotation while enabling real-time insights. The platform allows researchers to upload videos to the cloud, specify target events (i.e., bites, chews, and sips), and receive an automatically generated Excel file containing event predictions with corresponding timestamps. Additionally, a consolidated report summarizing event counts across all analyzed videos is provided. To ensure data privacy, all video files are automatically deleted immediately after analysis completion. Additionally, any uploaded video files that remain unanalyzed are deleted after 24 hours, making sure that video files are never stored on our servers for extended periods of time.

AMVA incorporates key advancements in privacy and data security. It enables eating behavior recording without capturing facial features through an integrated facial masking system, preserving participant anonymity (Figure 1). This approach minimizes extensive privacy agreement requirements, facilitating large-scale, unidentified eating behavior studies.

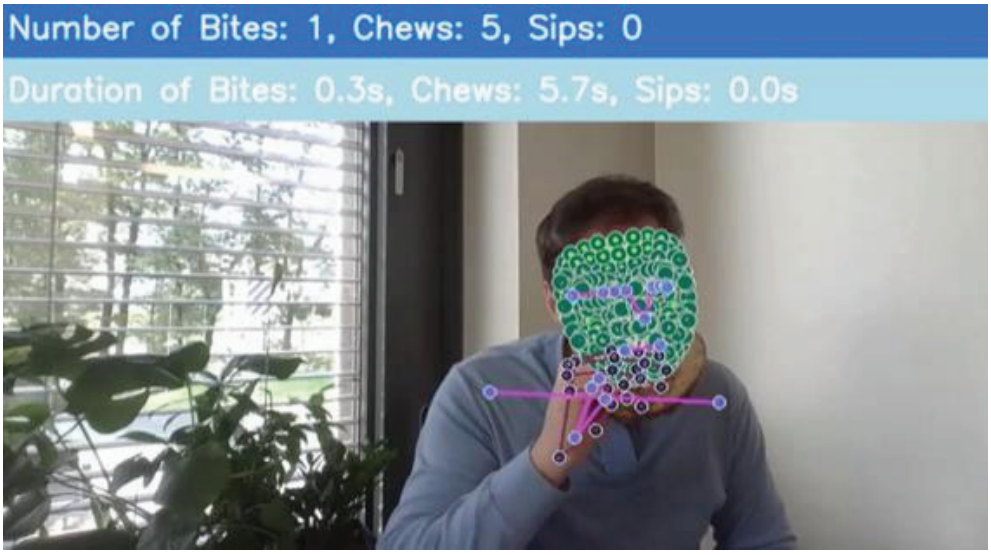


Figure 1. Demonstration of AMVA research software tracking bite and chew detection while maintaining privacy through facial masking (shown in green).

AMVA toolkit demonstrates significant efficiency gains in eating behavior analysis. Manual annotation time decreased from six weeks to six hours for 118 videos—a 10-fold reduction. The system achieves 98% accuracy in bite frame prediction and 86% in bite counting, with 85.5% accuracy for chewing frame detection and 80% for chew counting, as previously reported (Chapter 5 of this thesis).

The system's real-time processing capabilities and standardized output formats provide immediate analytical feedback, supporting both retrospective analyses and prospective study designs. This automation framework establishes a scalable foundation for objective eating behavior quantification across diverse research contexts, reducing manual annotation dependencies while preserving measurement precision.

IMPLEMENTATION

The AMVA Toolkit is implemented as a cloud-native application that leverages Amazon Web Services (AWS) to provide secure authentication, scalable video processing, and efficient data management. All AWS resources are deployed in the *eu-west-3 (Paris)* region, which means all data and its processing remain within the European Union to ensure compliance with data protection regulations such as the General Data Protection Regulation (GDPR)[10]. The system employs a serverless backend architecture integrated with a web-based frontend interface. This configuration enables researchers to upload video data, execute automated eating behavior analysis algorithms, and access computational results through a streamlined workflow. Figure 2 provides the complete system architecture overview.

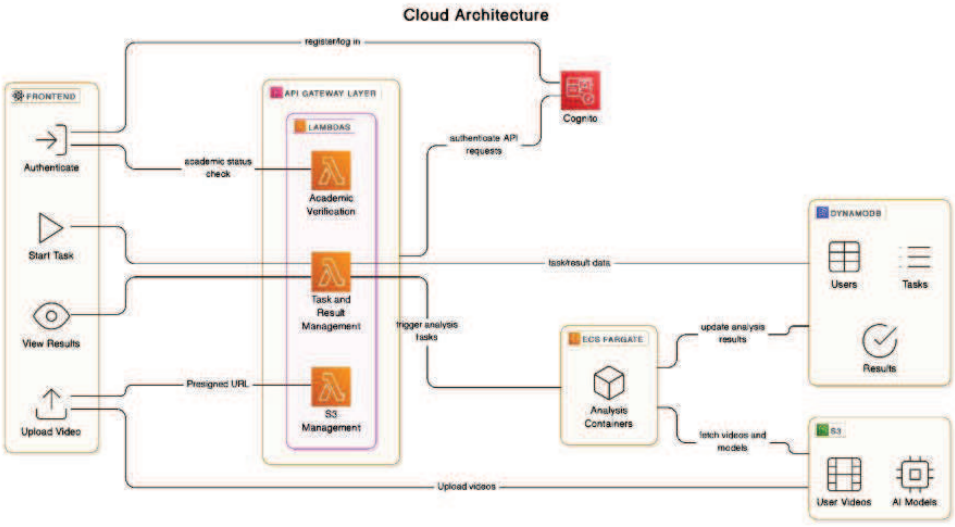


Figure 2: Schematic overview of the cloud-based architecture for the video analysis workflow. The system enables users to authenticate via a frontend interface and interact with cloud services through an API Gateway and Lambda functions. Authentication is managed by Amazon Cognito. Users can uploads videos using a presigned URLs to Ama-

zon S3, and task initiation triggers analysis in ECS Fargate containers. These containers retrieve video and model data from S3, perform the analysis, and store results in Amazon DynamoDB. Lambda functions manage academic status checks, task orchestration, and result handling, ensuring efficient and secure processing.

The system architecture consists of three core layers: (1) a web-based frontend, (2) an AWS Lambda API, and (3) a containerized computer vision pipeline on ECS Fargate. Both Lambda and ECS Fargate operate as serverless services, consuming resources only during active workloads without persistent server instances. This design delivers environmental sustainability through reduced energy consumption and economic efficiency through pay-per-use resource allocation.

Two storage services support these components. Amazon DynamoDB stores minimal user metadata including task status, analysis results, and credit balances, with records linked to AWS Cognito-managed user identities. Amazon S3 handles larger assets including computer vision models and uploaded video files. For privacy and storage optimization, video files undergo automatic deletion upon task completion or after 24 hours if unprocessed.

WEB-BASED FRONTEND (NEXT.JS)

The frontend is implemented using Next.js, a React-based framework that provides the user interface for all cloud backend interactions. AWS Cognito handles authentication through secure token-based access control via login and registration forms. Once authenticated, users access a personal dashboard where they can upload video files and initiate analyses. The system operates on a credit-based model, charging users based on video duration, with initial credits provided upon registration. Academic users may apply for research credits, while direct credit purchasing functionality remains under development. Users monitor task progress through a dedicated overview interface, and completed analyses generate tabular result summaries with Microsoft Excel export capability for offline processing. The interface integrates authentication, file management, progress tracking, and result visualization within a unified workflow that streamlines the research process from data upload to analysis export.

SERVERLESS API ARCHITECTURE (AWS API GATEWAY + LAMBDA)

Frontend requests route through AWS API Gateway, which exposes RESTful endpoints connected to dedicated AWS Lambda functions. Lambda functions provide stateless, event-driven compute units that scale automatically with demand without persistent server requirements. The API Gateway and Lambda functions bridge frontend operations with backend AWS services. Core backend operations include generating pre-signed URLs for direct S3 video uploads, querying DynamoDB for task status and user data retrieval/updates, and initiating analysis tasks that trigger the containerized video processing pipeline via ECS Fargate.

THE CONTAINERIZED VIDEO ANALYSIS (AWS ECS FARGATE)

AWS ECS Fargate orchestrates video analysis through isolated containers allocated 16 vCPUs per task. The containerized pipeline downloads video files and pre-trained models from S3, executes vision transformer-based bite detection algorithms, maintains periodic DynamoDB status updates, stores analysis results, and automatically deletes input videos to ensure compliance with privacy regulations. This serverless architecture employs transient, per-task containers that consume computational resources exclusively during active analysis periods, reducing operational costs and environmental impact through demand-based scaling.

COMPUTER VISION PROCESSING FRAMEWORK

The AMVA Toolkit integrates transformer-based computer vision (presented in Chapter 5) through a Video Processor class that implements an efficient bite and chew detection pipeline. The system extracts frames at configurable intervals (default: 10 fps) and processes each frame using a pre-trained vision transformer model optimized for temporal sequence analysis in nutritional behavior assessment. This architecture enables automated analysis of eating patterns through frame-level feature extraction and sequential pattern recognition across video timeseries data.

The video processing workflow begins with frame extraction and color space transformation, converting BGR to RGB color representation for compatibility

with pre-trained visual models. Each extracted frame undergoes normalization and tensor conversion before being processed through the model. Frame processing occurs at reduced frequency relative to the source video frame rate to maintain computational efficiency while preserving temporal resolution sufficient for behavior analysis.

A critical innovation in this implementation is the temporal smoothing algorithm that mitigates prediction variance through a sliding window approach. The system maintains a fixed-length collection of model predictions (default window size: 25 frames) and calculates time-averaged probabilities to improve detection stability across video sequences. This approach effectively addresses temporal noise and intermittent detection issues common in video-based behavior analysis. The implementation employs a double-ended queue data structure to efficiently manage the sliding window, maintaining a running sum of logits to optimize computational performance during real-time analysis.

The bite detection logic implements a state-transition detection system that identifies events when probability thresholds are exceeded (default: 0.5) and the previous frame was in a negative state. Each detected event is recorded with millisecond-precision timestamps formatted as minutes, seconds, milliseconds, enabling detailed temporal analysis of eating behaviors. The system maintains continuous progress monitoring during processing, providing percentage-based feedback to facilitate processing of extended video sequences. scenarios, supporting flexible deployment models from local development to cloud-based production environments.

The AMVA Toolkit extends its computer vision capabilities through a CNN-LSTM hybrid architecture for chew detection and quantification. The Video Processor class integrates both bite and chew detection pipelines, utilizing configurable frame sampling rates and batch processing for computational efficiency. The chew detection workflow employs signal processing techniques as previously described (Chapter 5 of this thesis) with model predictions undergoing temporal smoothing and calibrated quantification.

RESULTS AND DISCUSSION

Performance evaluation utilized 12 real-world meal consumption videos processed through the AWS ECS Fargate containerized pipeline. Test videos maintained consistent 640×480 resolution while exhibiting variable duration (6 seconds to 26 minutes), frame count (174-38,022 frames), and frame rate (24-30 FPS).

Processing performance demonstrated 3-second to 26-minute execution times, achieving average frame processing rates of 30-40 milliseconds per frame. Container nodes operated with 16 vCPU and 32GB RAM allocation. Comprehensive results are described in Table 1.

Table 1. Performance evaluation of containerized bite and chew detection pipelines on AWS ECS Fargate across 12 real-world meal videos. Videos ranged from 6 seconds to 26 minutes (3.2-496.8 MB), with processing times scaling proportionally to video duration.

Video	File size	Video duration	Frames	Bite analysis time	Chew analysis time	Container start-up time	Bites detected	Chews detected
1	3.2 MB	11s	257	8s	25s	44s	1	10
2	5.6 MB	6s	174	4s	10s	43s	1	19
3	8.2 MB	27s	658	18s	1m 3s	42s	2	19
4	37.6 MB	2m 5s	3010	1m 34s	5m 36s	48s	8	77
5	261.7 MB	13m 39s	19644	9m 23s	33m 35s	45s	31	635
6	270.8 MB	17m 50s	25644	13m 28s	47m 56s	46s	18	548
7	273.0 MB	16m 8s	23215	10m 34s	39m 53s	45s	34	577
8	280.0 MB	16m 6s	23152	11m 52s	43m 51s	46s	40	567
9	291.3 MB	17m 24s	25031	13m 11s	47m 23s	46s	17	769
10	301.9 MB	15m 7s	21747	10m 50s	39m 55s	44s	19	491
11	480.3 MB	26m 2s	37443	18m 37s	1h 10m 8s	46s	54	1048
12	496.8 MB	21m 7s	38022	12m 7s	45m 1s	50s	25	616

The data shown in Table 1 demonstrate an approximate linear correlation between frame count and processing time, with notable exceptions indicating additional complexity factors. For example, V12 processed faster than V11 despite

higher frame count (bite analysis: 12m vs 18m), correlating with bite detection density differences (25 vs 54 bites). Similar patterns appear between V9 and V8, suggesting computational overhead scales with detected eating events rather than frame count alone. Container initialization requires 42-50 seconds, creating negligible overhead for extended recordings but proportionally significant impact on brief video analyses. Chew detection is considerably slower than bite detection. This is to be expected considering the increased complexity of the CNN-LSTM architecture that is used for chew detection when compared to the state-transition approach for the bite analysis.

The current CPU-optimized ECS Fargate environment provides stable performance but suboptimal efficiency for machine learning inference and OpenCV video processing workloads, which benefit substantially from GPU acceleration[11]. Amazon's lack of GPU-enabled Fargate instances necessitates future migration to NVIDIA-powered EC2 instances[12]. Implementation of spot instances and optimized work queues will enhance computational sustainability and cost efficiency. Despite current limitations, the system demonstrates sufficient performance for scalable eating behavior analysis in production environments.

The AMVA Toolkit demonstrates several key advantages for eating behavior analysis. The cloud-native architecture enables scalable processing with automatic resource allocation, eliminating local computational constraints. The serverless implementation provides cost-effective operation through pay-per-use resource consumption and automatic scaling. Standardized output formats facilitate integration with statistical analysis pipelines, while the credit-based access model accommodates diverse research budgets.

Yet, current implementation constraints include CPU-only processing on ECS Fargate, resulting in suboptimal performance for machine learning inference workloads. Container initialization overhead (40-53 seconds) creates inefficiencies for brief video analyses. The system requires stable internet connectivity for cloud-based processing, potentially limiting field deployment scenarios. Processing costs may accumulate for large-scale studies with extensive video datasets. Model performance exhibits sensitivity to video quality parameters and recording conditions, with accuracy degradation observed under suboptimal lighting conditions or non-frontal camera positioning relative to participants.

CONCLUSION

The AMVA Toolkit represents a significant advancement in automated eating behavior analysis, addressing critical limitations of manual annotation methodologies through cloud-native implementation and computer vision automation. The system demonstrates substantial efficiency gains, reducing analysis time by 10-fold while maintaining high accuracy rates for bite detection (98% frame prediction, 86% counting accuracy), and for chew detection (85% frame prediction, 80% counting accuracy). Privacy-preserving facial masking capabilities enable large-scale studies without extensive consent protocols, while serverless architecture provides cost-effective scaling and GDPR compliance.

Performance evaluation confirms linear processing scalability with video duration, though computational overhead correlates with detected eating events rather than frame count alone. Current CPU-optimized implementation provides stable performance, but future GPU migration will enhance processing efficiency for machine learning inference workloads. The platform's capabilities and standardized output formats support diverse research applications, from controlled laboratory studies to population-level dietary monitoring.

AMVA establishes a robust foundation for objective eating behavior quantification, enabling reproducible measurements across diverse research contexts while eliminating inter-rater variability. The cloud-based architecture facilitates collaborative research initiatives and longitudinal studies through automated data processing pipelines. This automation framework advances eating behavior research capabilities by providing scalable, precise, and privacy-compliant tools for eating behavior analysis.

AVAILABILITY AND REQUIREMENTS

The AMVA Toolkit is accessible as a web-based application hosted on AWS infrastructure within the EU-West-3 region for GDPR compliance. The platform requires user registration with email verification and operates on a credit-based system. Academic users can apply for free research credits, while commercial licensing arrangements are under development.

Users need standard web browsers with JavaScript enabled and stable internet connectivity for video uploads. Supported video formats include standard MP4 encodings with recommended resolutions of 640×480 or higher. Processing costs scale with video duration, requiring credit allocation prior to analysis initiation. No local software installation or specialized hardware is required.

The AMVA Toolkit operates under a proprietary license model with academic research exemptions. Video data processing complies with GDPR regulations through automated deletion protocols and EU-based data residency requirements.

REFERENCES

1. Forde, C. G., Leong, C., Chia-Ming, E. & McCrickerd, K. Fast or slow-foods? Describing natural variations in oral processing characteristics across a wide range of Asian foods. *Food Funct.* **8**, 595–606 (2017).
2. Ioakimidis, I. *et al.* Description of chewing and food intake over the course of a meal. *Physiol. Behav.* **104**, 761–769 (2011).
3. 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**, 209 (2020).
4. ELAN | The Language Archive. https://archive.mpi.nl/tla/elan.
5. Tufano, M., Lasschuijt, M., Chauhan, A., Feskens, E. J. M. & Camps, G. Capturing Eating Behavior from Video Analysis: A Systematic Review. *Nutrients* **14**, 4847 (2022).
6. Bell, B. M. *et al.* Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review. *Npj Digit. Med.* **3**, 1–14 (2020).
7. Katsikari, A., Pedersen, M. E., Berget, I. & Varela, P. Use of face reading to measure oral processing behaviour and its relation to product perception. *Food Qual. Prefer.* **119**, 105209 (2024).
8. FaceReader | Facial expression analysis. *Noldus* https://noldus.com/facereader.
9. Lewinski, P., den Uyl, T. M. & Butler, C. Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *J. Neurosci. Psychol. Econ.* **7**, 227–236 (2014).
10. Union, P. O. of the E. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). *Publications Office of the EU* https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en (2016).
11. Mittal, S. & Vaishay, S. A survey of techniques for optimizing deep learning on GPUs. *J. Syst. Archit.* **99**, 101635 (2019).
12. Recommended GPU Instances - AWS Deep Learning AMIs. https://docs.aws.amazon.com/dlami/latest/devguide/gpu.html.

7

Discussion

MAIN FINDINGS

This thesis aimed to determine how artificial intelligence can be used to automate assessment of eating behavior and urban food landscapes to support nutrition research on the individual and population level. Through five interconnected studies, we established computational frameworks that enable scalable, objective measurement of dietary patterns at both the individual and food environment level.

URBAN FOOD LANDSCAPE

Urban food landscapes shape dietary habits and health outcomes, with limited access to nutritious food contributing to obesity and cardiometabolic outcomes, especially in socioeconomically disadvantaged areas [1–5]. In **chapter 2**, we introduced a data-driven method to assess urban food landscapes using restaurant menus from online delivery platforms in three cities: Boston, Dubai, and London. Menu items were matched to the U.S. FoodData Central database[6] using machine learning, allowing for the calculation of nutritional indices and neighborhood-level nutrient averages. Results showed strong links between food landscapes, socioeconomic status, and obesity. In London and Boston, wealthier neighborhoods had greater access to nutrient-rich foods, with dietary fiber showing a strong inverse relationship with obesity ($p=0.001$ and $p=0.004$, respectively).

Our data-driven approach extends beyond conventional categorical classifications (i.e., fast food vs. traditional restaurants) by analyzing individual menu items. Traditional food environment assessments often rely on observational methods—such as food diaries and in-store audits—that focus on availability, quality, and pricing, but lack item-level nutritional detail[7,8].

To address this gap, we applied AI-based nutritional profiling by matching menu items to the USDA FoodData Central database[6], allowing for detailed assessments of nutrient availability across urban settings. Furthermore, by integrating data from multiple online food delivery platforms, our approach captures a more diverse range of food outlets. This scalable method can complement traditional food environment studies and help policymakers identify areas at risk of poor nutrition, obesity, and cardiometabolic outcomes.

AUTOMATED EATING BEHAVIOR DETECTION

Eating behavior is a critical determinant of nutritional intake and is closely linked to the development of eating disorders and obesity[9–12]. Current methods for identifying eating events—such as bites and chews—in video recordings primarily rely on manual annotation, which is labor-intensive, time-consuming, subjective, and lacks standardization[13]. However, video recordings of eating episodes offer a non-invasive and scalable data source, making them well-suited for automated analysis.

In **chapter 3**, we examined the existing methods for automatically detecting eating behavior events from video, screening scientific publications following PRISMA guidelines. Active appearance model, optical flow, video fluoroscopy, deep learning, and facial landmarks were the methods found in this review. Among those, facial landmarks offered the most effective method to track facial movements to model and count bites and chews.

In **chapter 4**, we developed a rule-based computationally efficient method for bite detection using 3D facial landmarks. Prior studies have primarily relied on rule-based approaches using 68 2D facial keypoints, achieving 75–80% accuracy under controlled conditions[14]. Our approach leverages 468 3D facial keypoints, offering improved spatial resolution and robustness to partial occlusions which is an important advantage in naturalistic eating settings. We evaluated the system's performance against manual annotations across 164 videos from 15 participants. The model achieved a bite detection accuracy of 79.0% when annotations were available, and 71.4% in fully automated settings without annotation input. Importantly, the system maintained consistent performance across various food textures, demonstrating its generalizability and practical utility. Yet, the accuracy of our rule-based program was not high enough for all applications which led us to combine the 2 most promising methods found in chapter 3: deep learning and facial landmarks.

In **chapter 5**, we developed automated methods using transformer-based architectures to detect and quantify eating behaviors from meal videos. Using 103 annotated recordings from 36 participants consuming standardized meals, we trained Vision Transformer (ViT) and CNN-LSTM models for bite, chew, and sip detection. The transformer-based approach for bite detection constitutes a methodological advancement over the previously used CNN-based methods, leveraging self-attention mechanisms for global context processing within frame sequences. Unlike convolutional approaches limited by local receptive

fields, transformers capture long-range temporal dependencies critical for distinguishing eating activities from similar facial movements.

While earlier studies using Faster R-CNN and AlexNet combinations achieved >85% bite detection accuracy in controlled settings, our ViT-based approach demonstrated superior performance (98.45%) across diverse eating scenarios. The CNN-LSTM model for chew detection (85.56% accuracy) compares favorably with prior studies, though slightly below the CNN-LSTM RABiD algorithm which reported near-perfect agreement ($\kappa = 0.894$; F1-score: 0.948) achieved with manual feature engineering[15]. Sip detection using the pretrained VideoMAE model achieved 61% accuracy, constrained by limited data and annotation variability.

Once the AI models were developed and validated, we focused on making them accessible to the research community for automated analysis of eating behavior from video recordings. In **chapter 6**, we introduced the AMVA Toolkit, an open-source, cloud-native platform for automated video-based eating behavior analysis. Designed for the research community, it integrates deep learning models for bite and chew detection, deployed on AWS serverless infrastructure with containerized ECS Fargate instances featuring auto-scaling and GDPR-compliant data management.

In tests on 12 real-world meal videos, processing time scaled linearly with video length (3 seconds to 26 minutes). The system achieved 98% accuracy in bite frame detection and 86% in bite counting, processing at 30–40 ms per frame on 16 vCPU containers. Manual annotation time was reduced 40-fold—from six weeks to six hours for 118 videos. AMVA offers a scalable, objective alternative to manual annotation, enabling standardized outputs for both retrospective and prospective eating behavior studies.

METHODOLOGICAL CONSIDERATIONS AND VALIDITY

Geographic and Cultural Constraints in Food Environment Assessment

Our urban food landscape methodology exhibits limited applicability beyond U.S.-centric nutritional databases. Database coverage varied significantly: Boston (71%), London (56%), Dubai (42%), reflecting the absence of region-specific food composition data. This limitation affects generalizability and highlights

systematic bias in available nutritional datasets favoring Western dietary patterns.

The reliance on English-language menu descriptions excludes significant restaurant populations in multicultural urban environments. Dubai analysis missed 38% of restaurants due to Arabic-only menus, potentially biasing results toward Western-oriented establishments. This linguistic limitation compounds existing coverage gaps, creating systematic underrepresentation of traditional cuisine options.

Menu standardization assumptions may not reflect actual preparation methods varying by location. Our methodology currently lacks mechanisms for capturing intra-item variability, potentially misrepresenting nutritional availability in specific neighborhoods. Importantly, our dataset excludes consumption patterns and restaurants absent from online delivery platforms, creating potential gaps in neighborhood-level dietary assessment.

Ecological Validity and Individual-Level Variability

While our ecological, cross-sectional design provides valuable insights into neighborhood-level food environments and eating behaviors, it inherently limits the granularity of analysis by focusing on aggregated data rather than individual experiences. This aggregation hides significant within-area heterogeneity; for example, the nutritional availability metrics often display standard deviations larger than the mean values, indicating wide variability among residents' actual access and consumption patterns. Such variability challenges the suitability of applying uniform area-level interventions without accounting for individual differences in dietary choices, economic resources, cultural preferences, and mobility constraints.

Furthermore, the static nature of our assessment captures only a temporal snapshot, failing to reflect the dynamic characteristics of urban food landscapes. Seasonal menu rotations, special promotions, and fluctuating delivery areas can significantly alter food accessibility and availability over time. Yet, these temporal dynamics remain unaccounted for in our methodology. This limitation could lead to inaccurate or outdated representations of the urban food landscape, reducing the ecological validity of our findings when applied beyond the immediate data collection period (i.e., May 2023).

Additionally, individual-level factors such as personal dietary restrictions, health conditions, and habitual eating behaviors, which influence food selection, are not captured by area-level metrics. This gap underscores the need for

complementary research approaches incorporating longitudinal tracking and individualized data collection to better understand how neighborhood food environments translate into actual consumption and health outcomes.

In summary, while our study advances understanding of urban food landscape at the community scale, future work must integrate individual-level variability and temporal dynamics to enhance ecological validity and inform targeted, effective public health interventions.

Challenges and Data Collection Constraints in Eating Behavior

A major challenge in eating-behavior datasets arises from the severe scarcity of sip and drinking events, which creates a structural imbalance that weakens model reliability. With only n=1,847 sip instances compared to n=15,234 bites, the dataset reflects natural consumption patterns that inherently favor solid foods and thereby restrict the development of robust drinking-related classifiers.

Furthermore, accurate chew annotation remains difficult due to subtle jaw movements and video occlusion that reduce detection reliability The absence of clearly defined temporal boundaries for chewing compounds this issue, propagating inconsistency into model training. Additionally, annotation software commonly used for eating behavior analysis, such as ELAN[16], are suboptimal for capturing subtle jaw movements during chewing sequences. More specialized computer vision annotation tools like CVAT[17] or Darwin[18] would better accommodate the precision required for fine-grained chewing behavior analysis.

Beyond annotation challenges, the training process itself demands considerable computational resources and time investment. Deep learning models, particularly transformer architectures, require extensive hyperparameter tuning, large memory allocations, and prolonged training cycles to achieve optimal performance. Within the scope of this thesis, we constrained our methodological exploration to rule-based approaches, transformer architectures, and CNN-LSTM models. However, the landscape of applicable architectures extends far beyond these selected methods. Notably, we did not investigate object detection frameworks such as YOLO, which could potentially enhance eating event detection by simultaneously localizing food items, utensils, and hand-to-mouth movements within video frames. Similarly, hybrid approaches combining our rule-based temporal segmentation with alternative deep learning backbones—such as 3D convolutional networks, graph neural networks for facial landmark modeling, or attention-based temporal fusion modules—remain unexplored. The systematic evaluation of these architectural combinations, and their po-

tential synergies with existing rule-based preprocessing pipelines, represents a valuable direction for future research that was beyond the resource constraints of the present work.

After training, using AI models for inference requires specialized hardware. Real-time inference imposes substantial computational demands, as current implementations still rely on GPU acceleration to achieve sub-second latency. This dependency restricts deployment to high-performance systems and effectively excludes mobile and low-power edge devices. At the state of the art, hardware requirements limit the real-time eating behavior analysis to GPU-powered laptop, which undermines the applicability of such technology to real-time feedback aimed at reducing eating speed.

Lastly, restricted training data (salads/sandwiches only) affect the models' generalizability when eating behavior includes chopstick usage or when the person is in a downward-facing orientation. The dataset employed in this work excludes diverse eating implements and cultural practices, limiting cross-cultural applicability. Soup consumption and finger foods remain entirely unaddressed, representing significant gaps in eating behavior coverage.

IMPLICATIONS AND FURTHER RESEARCH

Automated Urban Food Landscape Assessment for Public Health Intervention

Data-driven assessment of urban food landscapes using AI enables rapid identification of neighborhoods with inadequate access to nutritious foods. Unlike labor-intensive observational methods requiring multiple audits across seasons, our approach provides comprehensive city-scale analysis with minimal manual intervention. The methodology processes thousands of menu items simultaneously, extracting nutritional information unavailable through traditional storefront observations.

This methodology offers policymakers an efficient tool for targeted nutritional interventions, as demonstrated by identifying the Al Ttay area in Dubai as potentially at-risk for obesity due to low dietary fiber availability (2.8g per average menu item versus 4.5g city average). The granular analysis reveals specific nutritional deficits actionable through targeted interventions: increased whole grain options, vegetable-forward menu items, and fiber-enriched alternatives.

Municipal health departments can deploy our data-driven approach for systematic food landscape assessment, enabling identification of neighborhoods requiring nutritional intervention. Successful implementation depends on forming data-sharing partnerships with food delivery platforms, establishing infrastructure for real-time analysis, and training staff to interpret and act on the results.

Following successful system-level deployment, this infrastructure facilitates continuous, data-driven surveillance of urban food environments. Routine monitoring offers a powerful mechanism to evaluate the impact of public health policies and optimize resource distribution. By conducting quarterly assessments, authorities can respond swiftly to shifts in the food landscape. Automated alerts flag neighborhoods where nutritional quality is declining, allowing for timely interventions. This real-time, responsive model represents a significant improvement over conventional evaluation cycles, supporting more adaptive and effective public health strategies.

To maximize the translational impact of this framework, future research priorities should address scalability and generalizability across diverse urban contexts. Researchers could expand automated assessment methods to diverse cities worldwide, particularly underrepresented South American and African locations, by partnering with regional universities. This expansion needs to address language barriers, cultural food differences, and varying platform ecosystems. Implementing continuous data mining with machine learning models could enable early detection of neighborhoods at risk for nutritional decline.

Development of Open-Source Nutritional Databases and Standardization

To overcome current fragmentation that hinders cross-country comparisons and global health initiatives international health organizations should prioritize development of comprehensive, region-specific nutritional databases with standardized food classifications. A unified framework, including hierarchical food categories, nutrient density metrics, and normalized portion sizes, would enable scalable, automated assessments of food environments and support evidence-based policy worldwide.

This infrastructure would enable global scalability of automated food environment assessment and supports evidence-based nutritional policy development. Standardized APIs and data formats would facilitate cross-platform compatibility and research reproducibility. Open-source development models would ensure equitable access while enabling community-driven improvements.

Database development should incorporate cultural food practices and regional variations. Current Western-centric databases inadequately represent global dietary diversity, limiting assessment validity in non-Western contexts. Collaborative development involving regional nutrition experts would ensure cultural appropriateness while maintaining scientific rigor.

Automated Eating Behavior Analysis for Healthy Eating Behavior

Manual video annotation of eating behaviors is a major bottleneck for scaling research, with a 20-minute meal video requiring 40–60 minutes of expert annotation—depending on the complexity of oral processing behaviors. Our automated approach compresses this timeline, achieving 40x faster processing that makes previously impossible study designs feasible. This dramatic speedup shifts research capabilities from small retrospective studies to the possibility to run large-scale prospective population research.

The system's capability to deliver behavioral feedback within 30 seconds of eating events enables just-in-time interventions, shifting intervention paradigms from post-hoc analysis to real-time behavioral modification. This temporal precision supports interventions targeting at reducing eating rate as a strategy to reduce eating intake.

We recommend research institutions to integrate automated eating behavior analysis systems to enhance analytical throughput and enable prospective study designs focused on healthy eating behavior promotion. Implementation requires establishing standardized video collection protocols, training personnel in system operation, and developing study-specific analytical pipelines.

The 40-fold efficiency gain enables longitudinal cohort studies of 1,000+ participants over multi-year periods, previously impractical due to annotation bottlenecks. This capacity supports robust evaluation of eating behavior interventions and facilitates trials investigating long-term behavioral modification through automated feedback systems.

Scalable eating behavior analysis requires four core technologies: transformer architectures for temporal pattern recognition, rule-based systems for interpretable baselines, facial landmark detection for precise activity localization, and cloud infrastructure for accessibility.

The serverless AWS deployment we used in Chapter 6 ensures GDPR compliance through automatic file deletion and EU-based processing. Lambda func-

tions enable parallel video processing that scales automatically with demand, eliminating infrastructure management for research teams.

This web-based system democratizes access to advanced eating behavior analysis, requiring only internet connectivity and standard browsers—enabling researchers without specialized computational resources to utilize state-of-the-art capabilities.

Future research should focus on multi-camera systems using synchronized 4-camera setups (frontal, profile, overhead views) to overcome occlusion limitations through multi-view fusion algorithms that maintain performance when individual cameras are blocked. Researchers should prioritize collecting diverse utensil data from Asian countries to reduce Western bias and enable culture-specific model selection, while also gathering more drinking behavior videos using new protocols where participants only drink without eating to create better datasets for training AI models.

Understanding how people eat, beyond what they eat, carries significant implications for human health. Eating rate, bite frequency, and chewing thoroughness have been independently associated with satiation, glycemic response, and long-term weight management. Rapid eating, characterized by large bites and insufficient mastication, has been linked to overconsumption and increased obesity risk. By developing automated methods to quantify these behavioral parameters, this thesis provides tools that can support both clinical interventions targeting eating behavior modification and epidemiological research examining behavioral determinants of diet-related disease.

Traditional dietary assessment relies heavily on self-report instruments such as food frequency questionnaires, 24-hour recalls, and food diaries. These methods are susceptible to systematic biases including underreporting of energy intake, social desirability effects, and recall errors. Furthermore, self-report captures what individuals eat but poorly characterizes how they eat—missing behavioral dimensions such as eating speed, bite patterns, and meal duration. The computational approaches developed in this thesis address these limitations by providing objective, continuous measurement that does not depend on participant memory or willingness to accurately report. This transition from subjective self-report methods to automated behavioral detection represents a methodological advancement with potential to improve the validity of eating behavior data in nutrition research.

While the urban food landscape analysis in Chapter 2 characterized what foods are available across neighborhoods, the eating behavior studies in subsequent chapters shift focus to how individuals consume food. This progression reflects the thesis structure outlined in the introduction—moving from population-level environmental factors to individual-level behavioral processes. Both dimensions matter for nutritional outcomes: access determines the range of dietary choices, while eating behavior influences the physiological consequences of those choices. An individual consuming identical foods may experience different satiation and metabolic responses depending on their eating speed and mastication patterns.

The two analytical levels examined in this thesis—urban food environments and individual eating behavior—are conceptually linked through their combined influence on nutritional health outcomes. Neighborhood food landscapes shape the availability and accessibility of dietary options, while individual eating behaviors determine how consumed foods are processed and experienced. Future research could integrate these levels by examining whether food environment characteristics influence not only what people eat but how they eat it. For instance, time-pressured urban environments with high fast-food density may promote both unhealthy food choices and accelerated eating patterns, compounding cardiometabolic risk. The computational methods developed here—automated nutritional profiling of food environments and objective eating behavior assessment—provide complementary tools for investigating such multilevel interactions.

CONCLUSIONS

This thesis demonstrates that artificial intelligence can successfully automate the assessment of eating behavior and urban food landscapes, providing scalable solutions to critical challenges in nutrition research. Through five interconnected studies, we established computational frameworks that transform labor-intensive manual processes into efficient, objective measurement systems capable of supporting population-scale monitoring and intervention.

Our urban food landscape assessment methodology enables rapid, comprehensive analysis of neighborhood-level nutritional availability using data from online food delivery platforms. By matching menu items to nutritional databases through machine learning, we identified significant associations between food environment quality, socioeconomic status, and obesity rates across Boston and London. This approach provides policymakers with actionable tools for identify-

ing at-risk neighborhoods and targeting nutritional interventions, representing a substantial advancement over traditional observational methods.

The automated eating behavior detection systems developed in this work achieve relevant accuracy while reducing annotation time by 40-fold. Our transformer-based bite detection model (98.45% accuracy) and CNN-LSTM chew detection system (85.56% accuracy) demonstrate superior performance compared to existing methods. The AMVA Toolkit makes these capabilities accessible to the broader research community through a cloud-native platform that processes videos with GDPR-compliant infrastructure, enabling real-time behavioral feedback and just-in-time interventions.

Despite current limitations—including geographic constraints in nutritional databases, cultural bias toward Western dietary patterns, and insufficient data for drinking behaviors—the foundational infrastructure established here enables researchers to transition from small-scale manual studies to large-scale automated research for both urban food landscapes and eating behavior.

This thesis demonstrates a methodological shift in nutrition research: from small-scale, manually coded studies toward AI-enabled systems supporting continuous, high-resolution, and ecologically valid measurement. Across both application domains, AI augments rather than replaces nutritional theory by enabling analyses previously infeasible due to cost, time, and scalability constraints.

Future work should extend beyond improving model accuracy to developing integrated, multimodal AI systems that support hypothesis-driven research and real-time behavioral interventions across diverse populations. By bridging food environments, eating behavior, and health outcomes, this thesis positions artificial intelligence as a foundational methodology for scalable, data-driven nutrition science.

Future development should prioritize three areas: expanding cultural data representation, improving model generalizability across diverse eating and drinking contexts, and establishing standardized protocols for global collaboration. Additionally, integrating novel deep learning technologies and multimodal approaches will be essential. Rather than focusing solely on discrete tasks like counting bites and chews, future systems should combine multiple behavioral signals—including contextual factors, emotional responses, and audio cues—to enable more comprehensive tracking of eating behavior.

REFERENCES

1. Egger, G. & Swinburn, B. An 'ecological' approach to the obesity pandemic. *BMJ* **315**, 477–480 (1997).
2. Cummins, S. & Macintyre, S. Food environments and obesity--neighbourhood or nation? *Int. J. Epidemiol.* **35**, 100–104 (2006).
3. Pineda, E. & Mindell, J. Association of the food environment with obesity: a systematic review of geographical and statistical methods. *The Lancet* **388**, S89 (2016).
4. Powell, L. M., Chaloupka, F. J. & Bao, Y. The availability of fast-food and full-service restaurants in the United States: associations with neighborhood characteristics. *Am. J. Prev. Med.* **33**, S240-245 (2007).
5. Poelman, M. *et al.* Relations between the residential fast-food environment and the individual risk of cardiovascular diseases in The Netherlands: A nationwide follow-up study. *Eur. J. Prev. Cardiol.* **25**, 1397–1405 (2018).
6. USDA FoodData Central. https://fdc.nal.usda.gov/.
7. Gamba, R. J., Schuchter, J., Rutt, C. & Seto, E. Y. W. Measuring the food environment and its effects on obesity in the United States: a systematic review of methods and results. *J. Community Health* **40**, 464–475 (2015).
8. McKinnon, R. A., Reedy, J., Morrissette, M. A., Lytle, L. A. & Yaroch, A. L. Measures of the food environment: a compilation of the literature, 1990-2007. *Am. J. Prev. Med.* **36**, S124-133 (2009).
9. Gahagan, S. Development of Eating Behavior: Biology and Context. *J. Dev. Behav. Pediatr.* **33**, 261 (2012).
10. Forde, C. G. & de Graaf, K. Influence of Sensory Properties in Moderating Eating Behaviors and Food Intake. *Front. Nutr.* **9**, 841444 (2022).
11. Otsuka, R. *et al.* Eating fast leads to obesity: findings based on self-administered questionnaires among middle-aged Japanese men and women. *J. Epidemiol.* **16**, 117–124 (2006).
12. Ohkuma, T. *et al.* Association between eating rate and obesity: a systematic review and meta-analysis. *Int. J. Obes. 2005* **39**, 1589–1596 (2015).
13. Forde, C. G., Leong, C., Chia-Ming, E. & McCrickerd, K. Fast or slow-foods? Describing natural variations in oral processing characteristics across a wide range of Asian foods. *Food Funct.* **8**, 595–606 (2017).
14. Nour, M. 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**, 28–34 (2021).
15. 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**, 209 (2020).
16. ELAN | The Language Archive. https://archive.mpi.nl/tla/elan.
17. Leading Image & Video Data Annotation Platform | CVAT. *CVAT.ai* https://www.cvat.ai.
18. V7 Darwin | AI Data Labeling & ML Training Data Platform. https://www.v7labs.com/darwin.

SUMMARY

Eating behavior and food environment accessibility are fundamental determinants of nutritional intake and health outcomes, yet traditional assessment methods—manual video annotation for eating behavior and observational audits for food environments—remain labor-intensive, time-consuming, and difficult to scale. This thesis investigated how artificial intelligence can automate assessment of eating behavior and urban food landscapes to support nutrition research at both individual and population levels.

At the population level, a data-driven methodology was developed to assess urban food landscapes using restaurant menu data from online delivery platforms in Boston, London, and Dubai (Chapter 2). Machine learning matched menu items to the U.S. FoodData Central database, enabling calculation of nutritional indices at neighborhood level. Database coverage varied substantially by city—Boston (71%), London (56%), and Dubai (42%)—reflecting limitations in region-specific nutritional data availability. Dietary fiber demonstrated significant inverse associations with obesity in both London ($p=0.001$) and Boston ($p=0.004$), while higher socioeconomic neighborhoods consistently showed better access to nutrient-rich foods. This methodology provides a scalable alternative to traditional food environment assessment, enabling policymakers to identify neighborhoods at risk for inadequate nutritional access.

At the individual level, a systematic review following PRISMA guidelines identified five methodological categories for automated eating behavior detection from video recordings: facial landmarks, deep learning, optical flow, active appearance model, and video fluoroscopy (Chapter 3). Facial landmarks emerged as the most promising approach for detecting both bites and chews. Building on these findings, a computationally efficient rule-based system utilizing 468 3D facial keypoints was developed for automated bite counting, achieving 79% accuracy with available annotation and 71.4% accuracy without annotation input across 164 videos from 15 participants, with consistent performance across varying food textures (Chapter 4).

To improve detection accuracy, transformer-based architectures were subsequently developed using 103 annotated videos from 36 participants consuming standardized meals (Chapter 5). The Vision Transformer model achieved 98.45% frame accuracy and 86.2% counting accuracy for bite detection, effectively capturing global spatial relationships through self-attention mechanisms. For chew detection, a CNN-LSTM architecture outperformed the Vision Transformer (85.56% versus 69.21% accuracy), as sequential models better captured

the temporal dynamics inherent to chewing behaviors. Sip detection using pretrained VideoMAE achieved 61% accuracy, constrained by limited training data and annotation challenges.

These validated models were integrated into the Automated Meal Video Analysis (AMVA) Toolkit, an open-source cloud-native platform deployed on AWS serverless infrastructure with GDPR-compliant data handling and privacy-preserving facial masking capabilities (Chapter 6). The system reduced manual annotation time 40-fold—from six weeks to six hours for 118 videos—while maintaining processing scalability linear with video duration.

This thesis demonstrates that artificial intelligence can successfully automate assessment of eating behavior and urban food landscapes, transforming labor-intensive manual processes into scalable, objective measurement systems. Current limitations include geographic constraints in nutritional databases, cultural bias toward Western dietary patterns, and insufficient training data for drinking behaviors (Chapter 7). Future work should prioritize expanding cultural data representation, improving model generalizability across diverse eating contexts, and developing integrated multimodal AI systems for comprehensive behavioral assessment. By bridging food environments, eating behavior, and health outcomes, this thesis positions artificial intelligence as a foundational methodology for scalable, data-driven nutrition science, enabling analyses previously infeasible due to cost, time, and scalability constraints.

PHD CANDIDATE'S AUTHORSHIP STATEMENT

Chapter 1 I discussed the topic and scope with my promotor and supervisors, defined the research questions, and wrote the full text. They provided conceptual and editorial guidance, which I incorporated. I am responsible for the final version. Claude Sonnet 4.5 assisted with text rephrasing.

Chapter 2 is a published article where I am first author. I co-formulated the research questions and study design, and took primary responsibility for the computational methodology, data processing pipelines, and nutritional database integration. I played a central role in interpreting results and co-developed figures and text with co-authors. I revised the manuscript on their feedback and am responsible for the final content. ChatGPT (GPT-4) was used for code refactoring and text rephrasing.

Chapter 3 I led the study design, literature review, analyses, and initial draft. I structured the review, synthesized findings, and prepared all tables and figures. Co-authors reviewed and edited; I revised accordingly and am responsible for the final content.

Chapter 4 I formulated the research question and, with my supervisors, shaped the study design and algorithms. I designed the rule-based framework, implemented bite-detection algorithms, supervised data collection and annotation, conducted all analyses, and wrote the manuscript. ChatGPT (GPT-4) assisted with text rephrasing, Python class design, and debugging. I bear full responsibility for the final content.

Chapter 5 I defined the research questions and, with my supervisor, developed and trained deep learning and transformer-based models. I performed all analyses, prepared figures, and wrote the chapter, incorporating feedback from my promotor and supervisor. ChatGPT (Model 4) and Claude Code (Sonnet 4.5) were used for code development, refactoring, and text rephrasing.

Chapter 6 Developed collaboratively. I wrote the introduction and discussion, and helped define the research questions. With my supervisor and co-authors, I contributed to the study design, methodology, system deployment, and interpretation. I participated in software deployment, revised based on feedback, and am responsible for the final version. ChatGPT (Model 4) and Claude Code (Sonnet 4.5) assisted with code tasks and text rephrasing.

Chapter 7 General Discussion I drafted the discussion, synthesizing findings across all chapters. My promotor and supervisor provided perspectives and edits, which I incorporated. I am responsible for the final text. ChatGPT (Model 4) and Claude Code (Sonnet 4.5) assisted with text rephrasing.

ACKNOWLEDGEMENTS

First and foremost, I want to thank Guido for everything he taught me during this Ph.D. journey. It's funny how, when you work with your leaders, you start to emulate them, subconsciously most of the time.

The first time we met was in Groningen, when I was doing an MSc extracurricular course about Leadership, Making the Difference. You were one of the lecturers. During the lecture on tone and speech delivery, you made us stand up and say a vegetable word with different intonations. I chose pumpkin, I don't know why. After repeating "pumpkin" 20 times trying to sound like a politician, you burst into a loud laugh.

It's great to emulate you, Guido! Thank you for not hiring me (at first)! I remember when you called me on a lazy Italian Sunday afternoon, asking how my Ph.D. applications were going. Then you decided to bring me to this little town in the middle of the Netherlands, giving me a chance as a research assistant. You've always been so supportive and kind, and sometimes so inappropriate. You faked firing me every other meeting with phrases like "We're sorry to let you go..." Very supportive, as always.

It's been a blessing working with you! I learned so much it's hard to describe. I'm extremely grateful that I met you and had the chance to share all these years, through the challenges, laughs, inappropriate jokes, and sarcasm. I'm also very happy that I met Danse and your kids, and that we shared our time at MIT. The perfect ending to this Ph.D. journey is, indeed... continuing to work together on the next venture, Eatpol. It's great that you're part of this journey too, so that we can keep working together! I'm also very excited to see how it will go.

Thank you so much for everything you did for me, all the teachings, and sometimes the subtle nudges. I've been the guinea pig for all your persuasion tricks, and I've started to recognize them year after year. "Oh, now he used this Cialdini trick to convince me to do this." It worked. Also, I had no choice. Thank you, thank you, thank you.

Thank you, Marlou! It's been great working with you. You taught me so much about eating behavior, project management, and communication, but most importantly I really appreciate that you were always there for any question, issue, or problem. It's been great working with you during these years and seeing how you've also developed as a supervisor. I wish you all the best for the future. Thank you.

Thank you, Edith! I really admire your leadership and guidance. I appreciate that, despite how busy you are, you've always created a safe space for juniors to express themselves, experiment, and grow. Thank you for always bringing the good spirit and leading me during this journey. I try to mimic some of your teachings in my work at Eatpol.

Thank you, Aneesh! You're a savior. Kind, fun, clever. Meeting you during my Ph.D. was really life changing. You taught me so much; I have great memories of our time spent discussing AI at Axis, and your guidance was extremely useful.

Thank you, Martina, Fabio, and Carlo. Thank you so much for everything you taught me and for all the amazing opportunities and fun we had together. To this day, I think "what would Martina or Fabio do?" when I'm approaching a problem.

Thank you, Kee, Jimmy, Ash, Diego, Jody, Paulina, Pei, Sadi, Tom, and all the SCL people! You guys made Boston and Senseable City Lab feel like home immediately. I have the fondest memories with you and can't wait to see you again.

Thank you, Femke, Esther, Zoë, Simon, and all the Hungry Robots! I'm very happy I've been part of this group of amazing tech people. Thank you for all the outings and fun times we've shared! It's been great sharing this journey with fellow robots.

Thank you to everyone at Global Nutrition and Sensory Science and Eating Behavior, and in particular Pari, Son, Zhiyao, Claudia, Eva, Matjaz, Kirsten, Giulia, Karin, Annaleen, Hans, Alida, Els, Ciaran, Victorie, Monica, and Markus. It's been great to share the office with you.

Thank you to Gitte, Robin, Robbin, and Stephan, and the whole StartHub community and Center of Value Creation for helping me in the transition from academia to entrepreneurship.

Thank you to my amazing paranymphs Hongy and Ina, for your friendship, and for just being the way you are! It's been great to have friends like you at the office.

Thank you, Nicolò, Zeno, Davide, Carlos, Kasper, and the Sphinx basketball team! My time in Wageningen has been great thanks to you!

Grazie alla mia famiglia, mamma, papà e Manu, per essere sempre di supporto e per essere le persone migliori del mondo! Vi amo e mi mancate sempre!

Thank you, Lucrezia! We met right after my COVID quarantine, after I moved to Wageningen. You've witnessed every step of the work in this book, always being supportive, kind, and there for me, and these years with you have been the best of my life. Thank you for your help with Chapter 2; your nutritional advice was essential for shaping the research. We had a blast in Boston, in Dubai, in all the places we've been together, and in our everyday life. I'm blessed to have met you, you're the best human ever, and I can't wait to live all the future adventures with you. Thank you for being the person you are. I love you.

ABOUT THE AUTHOR



Born in Naples, Italy, Michele Tufano studied at Sapienza University of Rome for his bachelor's in biotechnology before moving to Groningen, the Netherlands, in 2016 to pursue an MSc in Biomolecular Sciences. During this period, he conducted laboratory research on the genetic response of bacteria to water contaminants, while developing a growing interest in programming and computational biology.

An exchange program at Genome Institute of Singapore marked the beginning of his specialization in bioinformatics, where he started working with R and Python for data analysis. After spending a year at University of Illinois at Chicago, he decided to transition from wet-lab research to computational sciences, which led him to pursue a PhD in computer vision at Wageningen University & Research in the Netherlands. During a research exchange at Massachusetts Institute of Technology, Boston, USA, he applied artificial intelligence methods to quantify and analyze urban food environments across three cities (chapter 2 of this thesis).

Towards the completion of his PhD, Michele became motivated to translate his research into practical tools for eating behavior scientists and the food industry. After following entrepreneurship courses at Wageningen University, he recognized the broader potential of AI-driven food behavior analysis for product development and consumer research. This led to the founding of Eatpol, an official spin-off of Wageningen University & Research, focused on helping food companies develop more successful products by using AI to better understand consumer behavior. Today, he leads product development and strategy at Eatpol.

More information can be found at www.micheletufano.com and www.eatpol.com.

The research described in this thesis was financially supported by the grant PROJ-01041/Province of Gelderland -Op Oost -EFRO InToEat.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Michele Tufano (using Gemini Nanobanana 2) and Lindy Zuurmond

Printed by ProefschriftMaken