Diss. ETH No. 17780

Automatic dietary monitoring using on-body sensors:

Detection of eating and drinking behaviour in healthy individuals

A dissertation submitted to

ETH Zurich

for the degree of

Doctor of Sciences

presented by

Oliver Amft

Dipl. Ing., Chemnitz University of Technology
Date of birth 16 August 1975
citizen of Germany

accepted on the recommendation of

Prof. Dr. Gerhard Tröster, examiner
Prof. Dr. Wolfgang Langhans, co-examiner
Dr. Serge Reichlin, co-examiner

2008

Oliver Amft

Automatic dietary monitoring using on-body sensors:
Detection of eating and drinking behaviour in healthy individuals

Diss. ETH No. 17780

First Edition 2008 1 2 3
Published by ETH Zürich, Switzerland

ISBN 978-3-909386-77-2

Printed by Lulu.com
Copies may be ordered online from http://www.lulu.com

To Corina

Contents

Abstract

Energy balance and healthy eating behaviour are essential aspects that determine health risks for chronic diseases and general morbidity. This crucial relation becomes most prominent with the increasing share of citizens developing excessive body fat and weight worldwide. While short-term clinical interventions can help obese patients, long-term strategies are sought to prevent overweight and obesity. Sustained success in prevention is expected by supporting individuals in changing personal lifestyle and maintaining an appropriate eating behaviour.

Current weight and diet coaching programs use self-reporting techniques of eating behaviour to adapt and personalise feedback to participants. However maintaining these reports is an additional burden for participants. Moreover the reports incur large bias and hence, limit program success. Novel tools and technical solutions are sought that alleviate the individual from manual eating behaviour reporting.

In this work a novel concept is introduced, called automatic dietary monitoring (ADM), that targets this goal. New ADM-based diet coaching solutions are supported by the constant trend in electronic miniaturisation. Miniaturisation permits to embed sensors and computers in everyday objects, including clothing, accessories, and buildings. Systems that leverage this paradigm of pervasive computing can support their user with personalised health status and diet coaching services. Moreover ADM-based solutions conceptually permit coaching program durations of several years in order to make the coaching most effective. The essential functions of ADM-based solutions are sensing and recognition of the user's eating behaviour. This work evaluates on-body sensing and pattern recognition solutions for ADM.

The thesis comprises eight scientific publications that address four specific goals of this work: (1) to review on-body sensing solutions and modalities, relevant for diet monitoring, (2) to evaluate recognition of intake activities from continuous sensor data, (3) to infer intake cycles from temporally distributed activity events, and (4) to estimate eating behaviour from recognised activities.

On-body sensing solutions were reviewed with respect to the physiology of eating and activities directly related to food intake. Three activities were selected for further evaluation: intake gestures (using inertial sensors at arms and torso), chewing (using an ear-worn microphone to record food breakdown sounds) and swallowing (using Electromyography, EMG, and a stethoscope microphone).

A procedure to recognise activity events in continuous data was developed. The procedure utilises explicit data segmentation (equidistant or data-adaptive), a pattern search based on feature similarity and an event fusion

step. All selected activities were evaluated using the recognition procedure. Data-adaptive segmentation was utilised for inertial sensors and EMG data. Equidistant segmentation was evaluated for all modalities. The feature similarity search was used to spot activity events of variable length. Finally, event fusion was used to combine several similarity search instances in one recognition result. The recognition performance for spotting and categorising activity events from all sensing solutions was quantitatively analysed in several studies. It was shown how individual recognition stages improve performance.

A model was developed to describe food intake cycles. The model is based on a temporal composition of intake gestures, chewing, and swallowing events. To implement the model an activity event parsing approach was used. Applicability of the implementation was evaluated for different types of intake (food categories, drinking). In a second investigation, acoustic chewing phases were identified in intake cycles using an exploratory search approach. An acoustic chewing sequence model was introduced to facilitate the search task. The results show that acoustic phasing structure depends on food texture.

Finally, estimation of food type and amount from on-body sensor information was investigated. The relation of food, material texture and acoustic breakdown emissions was used to derive pattern models for discriminating up to 19 foods. Recognition of food-texture groups corresponding to nutritional recommendations (food pyramid) was evaluated. Furthermore, robust recognition of a fixed food set was demonstrated by combining the recognition procedure with intake cycle information. Food weight was estimated for individual bites of recognised foods. The weight estimation is based on timing and count variables of the chewing cycle structure. Predictive information of several variables was investigated. In a further investigation, bolus volume was classified from the swallowing reflex. While the latter approach provided categorical amount information, the approach based on chewing sequence variables allowed a continuous weight prediction.

Zusammenfassung

Energiegleichgewicht und gesundes Ernährungsverhalten sind zwei wesentliche Faktoren, die das Risiko für chronische Krankheiten und allgemeine Morbidität beeinflussen. Der weltweit wachsende Anteil von Personen mit überhöhtem Körperfett und -gewicht verdeutlicht diese Abhängigkeit. Während kurzfristige klinische Interventionen bei Adipositas-Patienten helfen, sind jedoch langfristige Strategien notwendig, um Übergewicht und Adipositas zu vermeiden. Einen nachhaltigen Erfolg versprechen Präventionsmassnahmen, die persönliche Lebensveränderung und adäquates Ernährungsverhalten unterstützen.

Aktuelle Beratungsprogramme für Gewicht und Ernährung benutzen Berichte über das Ernährungsverhalten, die vom Teilnehmer selbst verfasst wurden. Die Beratung wird entsprechend dieser Berichte angepasst und personalisiert. Für Programmteilnehmer ist das Ausfüllen der Berichte jedoch ein zusätzlicher Aufwand. Darüber hinaus, haben die Berichte einen hohen Bias und begrenzen damit den Programmerfolg. Neue Methoden und technische Lösungen sind nötig, um Programmteilnehmer von der manuelle Erfassung des Ernährungsverhaltens zu entlasten.

In dieser Arbeit wird ein neues Konzept eingeführt, genannt Automatic Dietary Monitoring (ADM), dass diese Entlastung zum Ziel hat. Neue, ADM-basierte Beratungsprogramme werden insbesondere durch den anhaltenden Miniaturisierungstrend bei elektronischen Systemen unterstützt. Die Miniaturisierung erlaubt es, Sensoren und Computer in alltägliche Objekte zu integrieren, wie zum Beispiel in Kleidung, Accessoires und in Gebäude. Systeme, die diesen Gedanken des Pervasive Computing verfolgen, können ihren Benutzer mit personalisierten Rückmeldungen zum Gesundheitsstatus und Ernährungsberatungsdiensten unterstützen. Darüber hinaus erlaubt das ADM Konzept eine Programmdauer von mehreren Jahren, um die Beratung wirkungsvoll zu gestalten. Die wesentlichen Funktionen ADM-basierter Lösungen sind die messtechnische Erfassung und Erkennung des individuellen Ernährungsverhaltens. Dieser Arbeit untersucht insbesondere körpergetragene Sensoren und Lösungen zur Mustererkennung für ADM.

Die Arbeit besteht aus acht wissenschaftlichen Publikationen, die vier spezifische Ziele verfolgen: 1. Evaluierung von körpergetragene Messlösungen und Modalitäten zur Ernährungsbeobachtung, 2. Untersuchung zur Mustererkennung in Aktivitäten der Nahrungsaufnahme, 3. Erkennung des Nahrungsaufnahmezykluses aus zeitlich verteilten Aktivitätsereignissen und 4. Bestimmung des Ernährungsverhaltens auf Basis der Aktivitätserkennung.

Es wurden körpergetragene Messlösungen im Hinblick auf die Ernährungsphysiologie und Aktivitäten untersucht, die direkt mit der Nahrungsaufnahme zusammen hängen. Drei Aktivitäten wurden für die weitere Untersuchung aus-

gewählt: Gesten zur Nahrungsaufnahme (mit Hilfe von Inertialsensoren an den Armen und am Rumpf), Kauen (mit Hilfe eines Ohrmikrophon zur Aufnahme von Kaugeräuschen) und Schlucken (mit Hilfe von Elektromyographie, EMG, und einem Stethoskop-Mikrophon).

Ein Verfahren zur Mustererkennung von Aktivitätsereignissen in kontinuierlichen Daten wurde entwickelt. Das Verfahren nutzt die drei Schritte Datensegmentierung (äquidistant oder daten-adaptiv), eine Mustersuche basierend auf Merkmalähnlichkeiten und eine Ereignis-Fusion. Alle ausgewählten Aktivitäten wurden mit diesem Erkennungsverfahren untersucht. Die daten-adaptive Segmentierung wurde für Inertialsensoren und EMG-Daten eingesetzt. Die äquidistante Segmentierung wurde auch für alle weiteren Sensormodalitäten untersucht. Die Merkmalähnlichkeitssuche wurde benutzt, um Aktivitätsereignisse mit variabler Länge zu erkennen. Schliesslich wurde die Ereignis-Fusion entwickelt, um mehrere Instanzen zur Merkmalähnlichkeitssuche in einem Erkennungsergebnis zu verknüpfen. Die Erkennungsleistung für Detektion und Kategorisierung von Aktivitätsereignissen aller Messlösungen wurde in mehreren Studien quantitativ untersucht. Die Arbeit zeigt, wie die Erkennungsschritte eine kontinuierliche Erkennungsleistung verbessern.

Ein Modell wurde entwickelt, um den Nahrungsaufnahmezyklus zu beschreiben. Das Modell basiert auf einem zeitlichen Verbund von Gesten, Kau- und Schluckereignissen. Zur Umsetzung des Modells wurde ein linguistischer Analyseansatz zur Verarbeitung von Aktivitätsereignissen benutzt. Die Anwendbarkeit wurde anhand von verschiedenen Ernährungsformen (Speisekategorien, Trinken) untersucht. In einer zweiten Untersuchung wurden akustische Phasen im Nahrungsaufnahmezyklus mit Hilfe einer explorativen Suche identifiziert. Ein akustisches Kausequenzmodell wurde eingeführt, um die Suche zu ermöglichen. Die Ergebnisse zeigen eine Phasenstruktur in Abhängigkeit von der Speisentextur.

Schliesslich wurde die Bestimmung von Speisetyp und -menge aus Informationen der körpergetragenen Sensoren untersucht. Die Beziehung von Speise, Materialtextur und akustischen Zerbrechemissionen wurde benutzt, um akustische Modelle für die Unterscheidung von 19 Speisen zu bestimmen. Die Erkennung von Speisen-Texturgruppen wurde in Anlehnung an Ernährungsempfehlungen (Ernährungspyramide) untersucht. Weiterhin wurde die stabile Erkennung einer festgelegten Speisenzahl gezeigt, indem das Erkennungsverfahren mit Informationen aus dem Nahrungsaufnahmezyklus ergänzt wurde. Das Speisengewicht wurde für einzelne Bissen einer Speise bestimmt. Diese Gewichtsschätzung basiert auf zeitlichen Variablen und Zählgrössen aus der Kausequenzstruktur. Die Schätzqualität verschiedener Variablen wurde untersucht. In einer weiteren Untersuchung wurde das Bolusvolumen während des Schluckreflexes diskriminiert. Während der zweite Ansatz kategorische Mengeninformationen liefert, erlaubt der Ansatz basierend auf Kausequenz-Variablen die Schätzung eines kontinuierlichen Gewichtswerts.

1

Introduction

An introduction on the relevance of nutrition in daily life is provided. The global struggle in fighting diet-related pandemics and the need for alternative diet monitoring solutions is summarised. This lack of adequate diet reporting solutions motivates the present work – the development of novel automatic on-body diet monitoring techniques.

By reviewing state-of-the-art diet assessments, vital requirements for such new systems are presented. Moreover, initial monitoring attempts, originating in the area of pervasive healthcare, are discussed. Finally, the aims and outline of this thesis on on-body monitoring are presented.

1.1. The need for diet monitoring

Food intake aims at compensating energy expenditure, hence to attain a balanced metabolism. However, intake is more than that – it involves an enjoyable stimulus which cues eating. In the 20th century, the original challenge to acquire food became a commodity and many foods were designed with high energy content, oils, fat and caloric sweeteners [35, 42]. Concurrently, energy expenditure has decayed [28]. This is primarily due to reduced physical activity, required to accomplish everyday tasks and work. The World Health Organisation (WHO) reported a global rise in body fat, determined by the body mass index[1] (BMI) as consequence of energy imbalance [41]. According to [40, 41], BMI is used to identify overweight (BMI>25), or more severe, obesity (BMI>30). Both, overweight and obesity are a predispose for cardiovascular diseases, diabetes mellitus type 2 and further health risks [36], all eventually increasing morbidity [5, 29]. For 2005, WHO estimated a pandemic of 1.6 billion overweight and 400 million obese adults worldwide [42]. An even increasing trend was projected for 2015 that emphasises a surge in child and adolescent obesity.

The prevalence of obesity in the US population (aged 20 years and over) increased from 14% in 1980 to 23% in 1994 and reached 30% in 2000 [9]. Moreover, by 2000 the ratio of overweight US-citizens exceeded 64%. In 2004, 17% of US-children and adolescents aged 2 to 19 years were overweight [24]. National prevalence among adults in Europe ranges from 8% in Switzerland and 10% in Italy and the Netherlands to 25% in England and nearly 30% in Greece and Croatia. In 2008 about 21% of the German adult population was obese [16].

Estimations for economic cost of obesity range between 2% and 8% of total healthcare costs in several developed countries. While these estimates are conservative, obesity represents one of the largest cost items in national healthcare budgets [41].

Moreover, overweight and obesity is not restricted to high-industrialised regions and is even faster growing in developing countries. In 2002 about 15% of the Chinese citizens were overweight [16]. – Fighting these epidemic dimensions is a critical challenge for the success of our species!

Besides energy balance[2], food intake provides unique access to nutrients that cannot be sufficiently synthesised by the body, such as vitamins, minerals and water [44]. This explains the enjoyable stimulus and motivation for

[1] Body weight normalised by the squared body height; depends on age and body composition. Initially proposed by Adolphe Quetelet between 1830–1850. It is persistently used to assess body fat, despite its shortcomings (http://en.wikipedia.org/wiki/Body_mass_index and [7, 10, 21]).

[2] Substances providing energy include proteins, fats, carbohydrates.

a diversified diet composition. It warrants sufficient amount of all required nutrients [13]. Hence, any form of eating disorder is detrimental for health [4].

Several further eating disorders exist [3, 12], including crash dieting, anorexia nervosa, binge eating, bulimia and orthorexia (see page 219 for descriptions). Hence, there are many aspects influencing individual food intake, including genetic and physiologic as well as psychological and social constraints [17]. The result is an individual *eating behaviour* described by specific food choice and restraint, portion size, energy intake and meal intake frequency. Eating behaviour is reported in temporal resolutions ranging from individual snacks and meals every day to averages over several years, depending on the type of investigation [45].

Monitoring eating behaviour is the prerequisite for research on disease intervention and epidemiology as well as in deployed prevention, such as weight loss coaching programs [25, 46]. In order to systematically reduce disease risks, these programs target a modification of accustomed lifestyle. However, this is a tough challenge for the individual. It requires continuous, potentially life-long, everyday support and coaching [25]. To support the coach and individual with actual information, eating behaviour reporting must provide a similar temporal resolution, thus requires tracking of every individual meal intake. Such actual information is particularly vital to adapt feedback in coaching programs and has been identified to improve success rates [23].

Current studies on eating behaviour and weight loss consider typical intervention periods of six months [37]. However only 20% of the individuals that initially lost at least 10% of their weight, can maintain the new weight one year after discharge [47]. This result suggests an even longer coaching phase of two to five years.

To date, most investigations and weight loss programs assess daily eating behaviour (intake schedule, food composition, amount and energy content) with the help of questionnaires [48]. Questionnaire assessments in the form of self-reports *could* capture eating behaviour in the required temporal resolution and information detail. However, they fail due to the burden of manual logging [48]. All diet assessments (see Section 1.2 below) are either laboratory-based or require a considerable effort by the respondent.

Novel tools and technical solutions are sought that alleviate the individual from manual food intake logging. The vision for such solutions is to provide eating behaviour information in the conceptual quality of daily self-reports. This is the goal of *automatic dietary monitoring* (ADM). These solutions will remove the inter-individual estimation error and increase user compliance in interventions. Moreover, they would permit novel risk-prevention programs through long-term personalised coaching [6] – clearly infeasible using manual monitoring.

1.2. Classic assessment of metabolism and diet

Various monitoring solutions have been developed that targeted the understanding of human metabolism, its modification through food choices and intake patterns. The approaches can be grouped into metabolism-focused assessments, eating-rate monitoring and questionnaires.

Direct and indirect calorimetry assessments using metabolic chambers and the doubly-labelled water test represent the most accurate solutions to measure metabolic rate. The highest standard of metabolic assessment is achieved through heat or gas exchange measurement in metabolic chambers, hence by monitoring the effect of ingested energy [34]. However, this procedure is neither feasible for monitoring behaviour under the impact of natural environments nor acceptable for investigations spanning months. In contrast, the doubly-labelled water test [30] is particularly useful for measuring average metabolic rate while following normal lifestyle. It is performed by tracking the loss of deuterium and oxygen-18 from body fluids (saliva, urine, or blood) after administering dose of water labelled with the heavy isotopes. It is typically used for studies with durations of two weeks or less [27].

In order to specifically assess food weight and eating rate (intake weight over time in g/s) an "Universal Eating Monitor" (UEM) was introduced [18]. The approach utilises a table with an integrated scale to measure the plate or bowl weight. The system was used for assessments of fluid intake or prepared solid food pieces mostly. The table can track potential deceleration in the intake speed. The UEM is applicable for laboratory studies and was used in clinical assessments of obesity [19] and, more recently, for investigations on psychological aspects of eating behaviour [15].

Dietary assessment based on questionnaires measure food intake directly. They can be utilised without activity-restricting supervision or laboratory environments. Three techniques exist: food-frequency history, 24 h recall and food records. Food-frequency assessments have been designed for epidemiology studies, capturing food consummation history (food item from a list and calendaring frequency) for long time periods (months to several years) [43]. The 24 h recall quantifies consumption of a single day through specific questions of an interviewer (food type and qualitative portion size) [48]. Food records are daily self-reports maintained by the respondent for up to one week, recording food type, time of consumption and weighted amounts [48]. Energy intake is assessed through manual analysis of reported food products by a dietitian.

Based on their temporal resolution of individual days to weeks 24 h recall and food records are used in eating behaviour studies and weight coaching programs [37]. However, both suffer from a number of shortcomings, such as motivation, intake awareness as well as memorising and literate capabilities of the respondent [48]. Moreover, respondents are influenced by changing perceptions of desirability and increasing self-awareness due to the reporting. In turn, food details that could be interpreted as abnormal are omitted, snacks

are forgotten. Reporting errors varied between 50% under- and overestimation [14, 31, 39]. Due to the required effort and large errors, food records are not applicable for monitoring durations longer than one week. Similarly, the effort of respondent and interviewer to maintain twenty-four hour recalls render the method infeasible for longer periods [48].

1.3. Diet monitoring using personal assistants

The constant trend of electronic miniaturisation has enabled sensors and computers to be embedded in everyday objects, including clothing, tools and buildings. Systems that leverage this pervasive computing paradigm can support their user with personal in-time health status and coaching services. The core functions for such personal assistants are (1) sensing and recognising the user's state and activity, (2) inferring health state as well as tracking tasks and actions relevant for the targeted service, and (3) providing adequate feedback. Minimising the user's disturbance by the system is a core property that affects the entire design.

In the light of dietary monitoring, sensing is a difficult challenge due to the complexity of eating behaviour. No single sensor or observable effect exist that would specifically resemble a manual self-report in natural environments. For this reason investigators and commercial solutions omit the sensing step and rather use classic self-reporting approaches instead. Diet monitoring research has focused on the translation of paper-based self-reporting into electronic diaries, such as PDA- or smartphone-based solutions e. g. [32]. Latest investigations and discussions indicate that PDAs cannot improve the validity of manual self-reporting assessments. They may even introduce new challenges to untrained users [1, 49, 50]. Investigations of alternate data entry methods, such as voice logs, bar-code and shopping receipt scanning resulted in similar estimation and validity errors [20, 33]. Commercialised solutions include many Internet-based coaching platforms using self-reports e. g. [2]. Moreover, services based on alternate reporting solutions have been established, such as MyFoodPhone [22] that uses mobile phone pictures for diet tracking.

Research has made sporadic attempts towards ADM. Typically, these works have focused on single activities and modalities. Patterson et al. [26] used radio-frequency-identification RFID tags on 60 household objects and a reader worn at the user's hand to track morning activities, including breakfast preparation and consumption. Chang et al. [8] used a table equipped with RFID readers to identify food containers and weight sensors and tracked food transport from containers to personal plates. While the first approach has potential to assess the meal timing and food type, the latter can additionally record food weight. Both approaches require specific labelling of objects and food to identify it. Finally, Gao et al. [11] deployed an computer vision approach to identify hand motions towards the head ("dining motions") of patients at a nursing home.

Their approach required cameras installed in the room. In addition to the behaviour sensing and recognition challenge, all purely infrastructure-based sensing approaches need to identify the person, e. g. the camera surveillance system requires an additional face recognition to robustly assign the movements to a specific person. Consequently, the development of wearable sensing approaches seems advantageous to eliminate these shortcomings.

1.4. Aims of the work

The aim of this work was to develop and evaluate new sensing and recognition solutions for ADM. With these solutions eating behaviour was inferred. All investigations focus on on-body sensing solutions. In order to emphasise the technical system development at this early stage, all studies considered healthy individuals in individual recording sessions up to 3 h duration for each participant. Specifically, the following goals were investigated:

Review of relevant on-body sensing solutions and modalities.

The lack of ADM solutions stems from the absence of unimodal sensing opportunities for food intake. All previous approaches required an instrumented environment, e. g. RFID, weight tables or cameras as information source. This work investigates the applicability of sensors worn at the body or attached to garments. As these sensors reside close to the body, detailed information regarding the eating behaviour is expected. In example, on-body sensors allow the recognition of individual food intake gesture types, rather than the unspecific to-head movement obtained from surveillance cameras. Body-worn sensors can provide information originating from physiologic responses to eating as well as activities preparing food absorption. In this work, behaviour sensing solutions are considered both aspects. Out of all solutions considered, a subsequent selection and detailed evaluation was made. The selection covered intake gesture, chewing and swallowing activities in order to describe the complete food intake cycle.

Recognition of intake activities in continuous sensor data.

Recognition of patterns in sensor data provides the basis for estimating eating behaviour information in this work. The focus was set on short-term (up to a several seconds in length), non-repetitive patterns in user activity. These units of activity are referenced as *activity events* throughout this work.

The challenge of activity event recognition originates from the requirements in practical deployment: (1) to spot activity events embedded in other, unknown data (NULL class) and (2) to categorise events according to a application-specific definition. The spotting can be viewed as a time-domain search problem with the aim to determine time of occurrence and length of every relevant event in the data. For the detection of variable, habitual and partly unconscious activities, as in eating behaviour, many classic recognition solutions fail. Sliding a fixed observation window over the data is not feasible, as certain activities, such as gestures, are varying approx. 100 percent in length. Moreover, evaluating the observation window for every new data sample is an inefficient processing approach. Finally, the presence of unknown embedding data prevents the generalisation of a naive binary classification (correct activity event or NULL class). For this problem [38] proposed the restriction on an activity subset. However this closed-set approach does not extend to one-class problems, such as the spotting of swallowing or a single food. Finally, the combination of detection and classification algorithms need to be evaluated for the class-specific independent recognition used in the work.

Fusion of temporally distributed activity events to infer intake cycles.

The complexity of eating behaviour cannot be captured in single activity events. By partitioning the activity recognition problem, individual results (activity event streams) are obtained that can be viewed as independent services. In this work, a temporal partitioning of dietary activities is considered. In order to infer eating behaviour, an intake cycle model is sought that permits the temporal combination of multiple recognition services in composite activities. In order to verify the model, an implementation is required that can parse activity event streams and permits recursive relations, such as consecutive sequences of chewing and swallowing events. The temporal fusion of activity events is a prerequisite for the food type and amount estimation.

Estimation of eating behaviour from activity recognition.

Most prominent diet monitoring goals include intake schedule, food composition, amount and energy content. In this work, the ADM approach is evaluated regarding the estimation of food type as well as amount. The recognised activity events and composite activities are used to quantitatively evaluate the food identification performance and estimate food amount.

1.5. Thesis outline

This thesis comprises eight scientific publications addressing the aims summarised above (Chapters 3 to 10). In Chapter 2 achievements and conclusions of the thesis are summarised. Finally, Chapter 2 provides an outlook onto open and new research challenges.

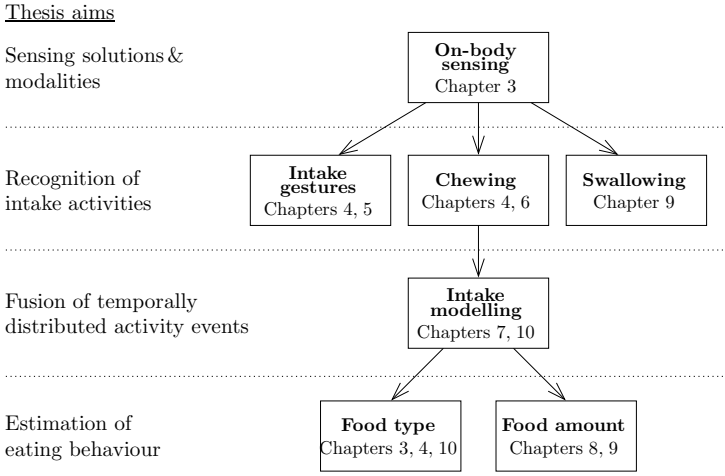


Figure 1.1. Outline of the scientific contributions included in the thesis according to the aims presented in Section 1.4. Arrows indicate result relations.

Table 1.1 lists the included publications and the chapter organisation. The publications are grouped according to the thesis aims presented in Section 1.4. Originating from the review of sensing solutions for diet monitoring in Chapter 3, three activity-based sensing approaches were evaluated in Chapter 4 and further in Chapter 5 (intake gestures), Chapter 6 (chewing), Chapter 9 (swallowing).

The fusion of activity events and recognition of intake cycles are discussed in Chapter 7 and 10. Chapter 7 targets the clustering of chewing cycles within chewing sequences. Chapter 10 presents an intake cycle modelling approach covering all three selected sensing solutions.

Eating behaviour was assessed regarding food type in Chapter 3 (food classification), Chapter 4 (texture group recognition) and Chapter 10 (intake cycle identification). Furthermore, food amount estimation was investigated in Chapter 8 (bite weight) and Chapter 9 (swallowing volume).

Figure 1.1 visualises the thesis contributions according to the aims presented in Section 1.4. Arrows indicate result relations.

Table 1.1. Publications included in the thesis (Chapters 3 to 10).

Chapter	Publication
3	Automatic Dietary Monitoring: On-body sensing solutions for eating behavior monitoring. O. Amft and G. Tröster. Submitted to *IEEE Pervasive Computing*, submitted June 2008.
4	Recognition of dietary activity events using on-body sensors. O. Amft and G. Tröster. *Artificial Intelligence in Medicine*, 42(2), 121–136, February 2008.
5	Gesture spotting with body-worn inertial sensors to detect user activities. H. Junker, O. Amft, P. Lukowicz, and G. Tröster. *Pattern Recognition*, 41(6), 2010–2024, June 2008.
6	Analysis of chewing sounds for dietary monitoring. O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. UBICOMP 2005: *Proceedings of the 7th International Conference on Ubiquitous Computing*, LNCS Vol. 3660, 56–72, Springer Berlin, Heidelberg, 2005.
7	Automatic identification of temporal sequences in chewing sounds. O. Amft, M. Kusserow, and G. Tröster. BIBM 2007: *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 194–201, IEEE Press, 2007.
8	Bite weight estimation using acoustic recognition of chewing. O. Amft, M. Kusserow, and G. Tröster. Submitted to *IEEE Transactions on Biomedical Engineering*, submitted June 2008.
9	Methods for detection and classification of normal swallowing from muscle activation and sound. O. Amft and G. Tröster. PHC 2006: *Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare*, ICST, 1–10, 2006.
10	Probabilistic parsing of dietary activity events. O. Amft, M. Kusserow, and G. Tröster. BSN 2007: *Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks*, IFMBE Proceedings Vol. 13, Springer, 242–247, 2007.

Bibliography

[1] J. Beasley. The pros and cons of using pdas for dietary self-monitoring. *J Am Diet Assoc*, 107(5):739, May 2007. doi:10.1016/j.jada.2007.03.023. author reply 739–739; author reply 740.

[2] J. Beidler, A. Insogna, N. Cappobianco, Y. Bi, and M. Borja. The PNA project. *J Comput Small Coll*, 16(4):276–284, May 2001.

[3] N. D. Berkman, C. M. Bulik, K. A. Brownley, K. N. Lohr, J. A. Sedway, A. Rooks, and G. Gartlehner. Management of eating disorders. Evidence report/technology assessment 135, Agency for Healthcare Research and Quality, US Department of Health and Human Services, 540 Gaither Road, Rockville, MD 20850, Apr 2006. prepared by: RTI-UNC Evidence-Based Practice Center, Research Triangle Park, NC.

[4] J. M. Bourre. Effects of nutrients (in food) on the structure and function of the nervous system: update on dietary requirements for brain. part 1: micronutrients. *Journal of Nutrition, Health & Aging*, 10(5):377–385, 2006.

[5] G. A. Bray. Medical consequences of obesity. *J Clin Endocrinol Metab*, 89(6):2583–2589, Jun 2004. doi:10.1210/jc.2004-0535.

[6] L. E. Burke, M. Warziski, T. Starrett, J. Choo, E. Music, S. Sereika, S. Stark, and M. A. Sevick. Self-monitoring dietary intake: current and future practices. *J Ren Nutr*, 15(3): 281–290, Jul 2005. doi:10.1016/j.jrn.2005.04.002.

[7] R. V. Burkhauser and J. Cawley. Beyond bmi: The value of more accurate measures of fatness and obesity in social science research. *J Health Econ*, 27(2):519–529, Mar 2008. doi:10.1016/j.jhealeco.2007.05.005.

[8] K.-H. Chang, S.-Y. Liu, H.-H. Chu, J. Y. Hsu, C. Chen, T.-Y. Lin, C.-Y. Chen, and P. Huang. The diet-aware dining table: Observing dietary behaviors over a tabletop surface. In K. Fishkin, B. Schiele, P. Nixon, and A. Quigley, ed., *PERVASIVE 2006: Proceedings of the 4th International Conference on Pervasive Computing*, vol. 3968 of *Lecture Notes in Computer Science*, pp. 366–382. Springer Berlin, Heidelberg, May 2006.

[9] K. M. Flegal, M. D. Carroll, C. L. Ogden, and C. L. Johnson. Prevalence and trends in obesity among us adults, 1999-2000. *JAMA*, 288(14):1723–1727, Oct 2002.

[10] D. Gallagher, M. Visser, D. Sepúlveda, R. N. Pierson, T. Harris, and S. B. Heymsfield. How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups? *Am J Epidemiol*, 143(3):228–239, Feb 1996.

[11] J. Gao, A. Hauptmann, A. Bharucha, and H. Wactlar. Dining activity analysis using a hidden markov model. In *ICPR 2004: Proceedings of the 17th International Conference on Pattern Recognition*, vol. 2, pp. 915–918, Aug. 2004. doi:10.1109/icpr.2004.1334408.

[12] A. Gonzalez, M. R. Kohn, and S. D. Clarke. Eating disorders in adolescents. *Aust Fam Physician*, 36(8):614–619, Aug 2007.

[13] M. R. Heath. The oral management of food: the bases of oral success and for understanding the sensations that drive us to eat. *Food Qual Prefer*, 13(7–8):453–461, October–December 2002. doi:10.1016/s0950-3293(02)00106-4.

[14] R. J. Hill and P. S. Davies. The validity of self-reported energy intake as determined using the doubly labelled water technique. *Br J Nutr*, 85(4):415–430, Apr 2001. doi: 10.1079/bjn2000281.

[15] R. Hubel, R. G. Laessle, S. Lehrke, and J. Jass. Laboratory measurement of cumulative food intake in humans: results on reliability. *Appetite*, 46(1):57–62, Jan 2006. doi: 10.1016/j.appet.2005.10.006.

[16] International Obesity TaskForce. Obesity prevalence data base. http://www.iotf.org/database/index.asp, June 2008. Last accessed: June 2008.

[17] E. Jequier and L. Tappy. Regulation of body weight in humans. *Physiol Rev*, 79(2): 451–480, Apr 1999.

[18] H. R. Kissileff, G. Klingsberg, and T. B. V. Itallie. Universal eating monitor for continuous recording of solid or liquid consumption in man. *Am J Physiol*, 238(1):R14–R22, Jan 1980.

[19] H. R. Kissileff, J. Thornton, and E. Becker. A quadratic equation adequately describes the cumulative food intake curve in man. *Appetite*, 3(3):255–272, Sep 1982.

[20] J. Mankoff, G. Hsieh, H. C. Hung, S. Lee, and E. Nitao. Using low-cost sensing to support nutritional awareness. In G. Goos, J. Hartmanis, and J. van Leeuwen, ed., *Ubicomp 2002: Proceedings of the 4th International Conference on Ubiquitous Computing*, vol. 2498 of *Lecture Notes in Computer Science*, pp. 371–376. Springer Berlin, Heidelberg, September-October 2002.

[21] Z. Mei, L. M. Grummer-Strawn, A. Pietrobelli, A. Goulding, M. I. Goran, and W. H. Dietz. Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents. *Am J Clin Nutr*, 75(6):978–985, Jun 2002.

[22] MyFoodPhone. World's first camera-phone & web-based-video nutrition service. Internet, Feb 2005. Accessed: August 2007.

[23] A. Oenema and J. Brug. Feedback strategies to raise awareness of personal dietary intake: results of a randomized controlled trial. *Prev Med*, 36(4):429–439, Apr 2003. doi:10.1016/s0091-7435(02)00043-9.

[24] C. L. Ogden, M. D. Carroll, L. R. Curtin, M. A. McDowell, C. J. Tabak, and K. M. Flegal. Prevalence of overweight and obesity in the united states, 1999-2004. *JAMA*, 295(13):1549–1555, Apr 2006. doi:10.1001/jama.295.13.1549.

[25] D. L. Paineau, F. Beaufils, A. Boulier, D.-A. Cassuto, J. Chwalow, P. Combris, C. Couet, B. Jouret, L. Lafay, M. Laville, S. Mahe, C. Ricour, M. Romon, C. Simon, M. Tauber, P. Valensi, V. Chapalain, O. Zourabichvili, and F. Bornet. Family dietary coaching to improve nutritional intakes and body weight control: a randomized controlled trial. *Archives of pediatrics & adolescent medicine*, 162(1):34–43, Jan 2008. doi:10.1001/archpediatrics.2007.2.

[26] D. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. In B. Rhodes and K. Mase, ed., *ISWC 2005: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*, pp. 44–51. IEEE Press, October 2005. doi:10.1109/iswc.2005.22.

[27] G. Plasqui, A. M. C. P. Joosen, A. D. Kester, A. H. C. Goris, and K. R. Westerterp. Measuring free-living energy expenditure and physical activity with triaxial accelerometry. *Obes Res*, 13(8):1363–1369, Aug 2005.

[28] B. Popkin. *Global trends in obesity*, chapter 1, pp. 1–13. CRC Press, Woodhead Publishing, 2005.

[29] A. P. Pérez, J. Y. Muñoz, V. B. Cortés, and P. de Pablos Velasco. Obesity and cardiovascular disease. *Public Health Nutr*, 10(10A):1156–1163, Oct 2007. doi:10.1017/s1368980007000651.

[30] D. A. Schoeller. Measurement of energy expenditure in free-living humans by using doubly labeled water. *J Nutr*, 118(11):1278–1289, Nov 1988.

[31] D. A. Schoeller. Limitations in the assessment of dietary energy intake by self-report. *Metabolism*, 44(2 Suppl 2):18–22, Feb 1995. doi:10.1016/0026-0495(95)90204-x.

[32] M. A. Sevick, B. Piraino, S. Sereika, T. Starrett, C. Bender, J. Bernardini, S. Stark, and L. E. Burke. A preliminary study of pda-based dietary self-monitoring in hemodialysis patients. *J Ren Nutr*, 15(3):304–311, Jul 2005.

[33] K. A. Siek, K. H. Connelly, Y. Rogers, P. Rohwer, D. Lambert, and J. L. Welch. When do we eat? an evaluation of food items input into an electronic food monitoring application. In E. Aarts, R. Kohno, P. Lukowicz, and J. C. Trainini, ed., *PHC 2006: Proceedings of the 1st International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–10. ICST, IEEE digital library, November 2006. doi:10.1109/pcthealth.2006.361684.

[34] A. J. Smeets and M. S. Westerterp-Plantenga. Acute effects on metabolism and appetite profile of one meal difference in the lower range of meal frequency. *Br J Nutr*, pp. 1–6, Dec 2007. doi:10.1017/s0007114507877646.

[35] S. A. Tanumihardjo, C. Anderson, M. Kaufer-Horwitz, L. Bode, N. J. Emenaker, A. M. Haqq, J. A. Satia, H. J. Silver, and D. D. Stadler. Poverty, obesity, and malnutrition: an international perspective recognizing the paradox. *J Am Diet Assoc*, 107(11):1966–1972, Nov 2007. doi:10.1016/j.jada.2007.08.007.

[36] T. L. Visscher and J. C. Seidell. The public health impact of obesity. *Annu Rev Public Health*, 22:355–375, 2001. doi:10.1146/annurev.publhealth.22.1.355.

[37] T. A. Wadden, M. L. Butryn, and C. Wilson. Lifestyle modification for the management of obesity. *Gastroenterology*, 132(6):2226–2238, May 2007. doi:10.1053/j.gastro.2007.03.051. Erratum in: Gastroenterology. 2007 Jul;133(1):371.

[38] J. Ward, P. Lukowicz, G. Tröster, and T. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE T Pattern Anal*, 28(10):1553–1567, Oct. 2006. doi:10.1109/tpami.2006.197.

[39] K. R. Westerterp and A. H. C. Goris. Validity of the assessment of dietary intake: problems of misreporting. *Curr Opin Clin Nutr Metab Care*, 5(5):489–493, Sep 2002.

[40] WHO. Physical status: the use and interpretation of anthropometry. WHO Technical Report Series 854, World Health Organization, Geneva, Switzerland, Geneva, Switzerland, 1995. Report of a WHO Expert Committee.

[41] WHO. *Obesity: Preventing and Managing the Global Epidemic*. Number 894 in Technical Report Series. World Health Organization, 2000.

[42] WHO. Global strategy on diet, physical activity and health (WHA57.17). In *Fiftyseventh World Health Assembly*. World Health Organization, May 2004.

[43] W. Willett. Food frequency methods. In W. Willett, ed., *Nutritional Epidemiology*, chapter 5, pp. 69–91. Oxford University Press, 1990.

[44] W. Willett. Foods and nutrients. In W. Willett, ed., *Nutritional Epidemiology*, chapter 5, pp. 20–33. Oxford University Press, 1990.

[45] W. Willett. Nature of variation in diet. In W. Willett, ed., *Nutritional Epidemiology*, chapter 5, pp. 34–51. Oxford University Press, 1990.

[46] W. Willett. Overview of nutritional epidemiology. In W. Willett, ed., *Nutritional Epidemiology*, chapter 5, pp. 3–19. Oxford University Press, 1990.

[47] R. R. Wing and S. Phelan. Long-term weight loss maintenance. *Am J Clin Nutr*, 82(1 Suppl):222S–225S, Jul 2005.

[48] J. C. Witschi. Short-term dietary recall and recording methods. In W. Willett, ed., *Nutritional Epidemiology*, vol. 4, pp. 52–68. Oxford University Press, 1990.

[49] B. A. Yon, R. K. Johnson, J. Harvey-Berino, and B. C. Gold. The use of a personal digital assistant for dietary self-monitoring does not improve the validity of self-reports of energy intake. *J Am Diet Assoc*, 106(8):1256–1259, Aug 2006. doi:10.1016/j.jada.2006.05.004.

[50] B. A. Yon, R. K. Johnson, J. Harvey-Berino, B. C. Gold, and A. B. Howard. Personal digital assistants are comparable to traditional diaries for dietary self-monitoring during a weight loss program. *J Behav Med*, 30(2):165–175, Apr 2007. doi:10.1007/s10865-006-9092-1.

2

Thesis summary

This chapter summarises the approach and most important achievements of the thesis. Specifically, on-body sensing solutions for diet monitoring are discussed, the performance of an activity event recognition procedure is summarised, and the results of different activity event fusion algorithms are presented.

Moreover, results of food type and amount estimation from the on-body sensing and recognition are summarised.

Conclusions, derived from the different achievements, are presented. The chapter closes with a discussion of limitations and an outlook, indicating open challenges and new research directions.

2.1. Summary of contributions

The most important results and novel achievements that advance the state-of-the-art in on-body diet sensing and recognition are presented below. The summary is structured according to the thesis aims introduced in Section 1.4 and illustrated in Figure 2.1. Detailed result descriptions and discussions can be found in the particular publication chapter referenced in this summary.

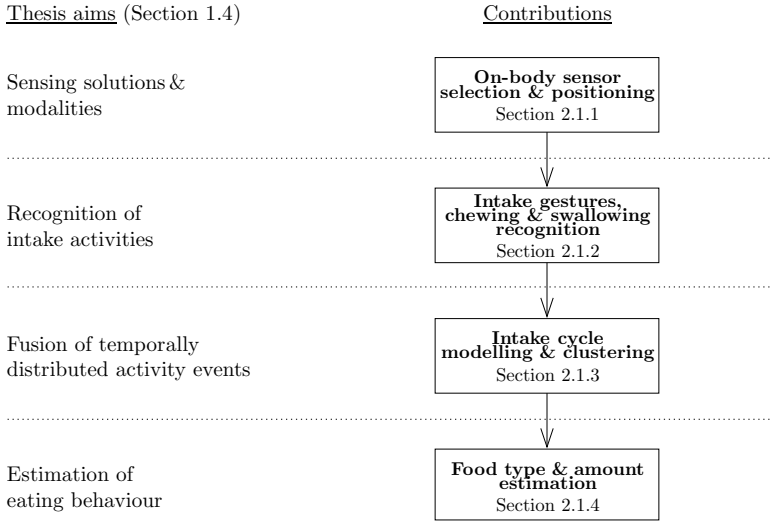


Figure 2.1. Outline of the contribution summary, presented in Section 2.1) according to the aims listed in Section 1.4. Arrows indicate result relations.

2.1.1. Sensing solutions and modalities

Regarding the first objective of this work (see Section 1.4 on page 6), the capabilities of body-worn sensing solutions for monitoring individual meal intake were reviewed. The analysis covered the physiology of eating as well as activities preparing food absorption. Table 3.1 on page 41 provides an overview of all considered on-body sensing solutions.

The timing of physiologic responses and activities is a crucial aspect for monitoring individual meal intake. It was assumed that responses following food intake with a long or variable delay greater than 10 min, are disturbed by other activities or subsequent food intake. Consequently, late stage digestion (gastric tract activity following the stomach) was excluded from the review.

- **Physiologic responses related to food intake.** The literature review showed that effects are variable in magnitude, duration and delay,

depending on various context aspects, such as mental state and physical activity (Section 3.5, page 38).

The varying magnitude, long response times (up to 3 h for cardiac activity) and variable response delays (15-30 min for heart rate) limit the feasibility of these information sources for ADM. Physiologic responses were not further considered in this work.

- **Activities preparatory to food absorption.** The review considered food preparation and ingestion (intake gestures), food breakdown in the mouth (chewing), bolus transport (swallowing, oesophageal movement) and gastric activity (stomach movement).

  Except intake gestures, all activities require an indirect measurement approach, due to comfort and privacy restrictions (Table 3.1, page 41). The review showed that all existing principles to assess oesophageal movement and gastric activity require controlled laboratory environments to maximise signal to body-noise ratio (Section 3.5, page 38).

Based on the review results, further discussion focuses on the following set of activities: intake gestures, chewing and swallowing. This set was selected, since it reflects the core activities of an intake cycle, as detailed in Section 2.1.3. For each activity, the selected sensor type, positioning and relevance for diet monitoring is summarised below.

- **Sensing of intake gestures.** Inertial sensors at lower and upper arms and the upper torso (Figure 5.9 on page 97) were investigated for intake gesture recording. Inertial measurement units (consisting of acceleration, gyroscope, magnetic field sensors) were initially used (Chapter 5). Later refinement showed that a subset of these sensors (lower arms and torso acceleration and gyroscopes) were sufficient to recognise four frequently used food intake gestures (Section 4.4, page 59).

  Food category is related to cutlery and, in turn, to the intake gestures used. For example, a soup is consumed with a spoon, rather than fork and knife. Inertial sensors can be integrated into clothing or accessories (Section 3.4.1, page 33).

- **Sensing of chewing.** Chewing was recorded from food breakdown sounds that propagate through mandible and skull. Section 6.1.4 on page 114 provides an introduction to the sensing approach. The evaluation showed that emitted sound pattern is related to food texture and can be used to identify chewed foods (Section 6.5, page 122).

  The chewing sound evaluation at various facial positions showed that the ear canal received +30 dB higher sound intensity among all positions

listed in Table 6.3 on page 120. Furthermore, sound intensity was +60 dB above the level of normal speaking (Section 6.2, page 120). The latter result depends on ear occlusion. Lower occlusion increases comfort, but lowers food recognition performance.

Two sensor prototypes with different occlusion were implemented and evaluated (Figure 3.3(a), page 37). Food classification rates increased up to 10% for a higher occlusion model (Section 3.4.2, page 35).

- **Sensing of swallowing.** Swallowing was assessed using surface Electromyography (EMG) at the hyoid (position close to the Adam's apple) and a stethoscope microphone at the lower neck (see Figure 9.2 on page 184 for exact positions). These sensors were investigated in a collar-prototype (Figure 3.4, page 38). Submental EMG (below chin) was investigated as additional source of information (Figure 9.2, page 184), however this position cannot be integrated in a collar. Furthermore, movement of the thyroid cartilage (Adam's apple) was analysed using a strain-sensitive fabric integrated in a collar (Figure 3.4, page 38).

  All investigated sensors were sensitive to head movement and voluntary neck contraction (Section 9.5, page 195). Hyoid EMG and sound were further analysed for the identification of swallowing, see Section 2.1.2 below.

2.1.2. Recognition of intake activities

To address the challenges of activity event recognition, introduced in Section 1.4, a recognition procedure was developed that accommodates the different sensing solutions considered in this work. In this effort, the following achievement were made.

- **Activity event recognition procedure.** The recognition procedure comprises (1) segmentation, (2) feature similarity search (FSS) and (3) event fusion was developed (Figure 4.1, page 54). As data-adaptive segmentation the Sliding-window and bottom-up (SWAB) algorithm was utilised for inertial sensors (intake gestures, Section 5, page 79) and EMG data (swallowing, Section 9.3, page 185). Equidistant segmentation was used with all sensing modalities, e. g. in Chapter 4 on page 47.

  The FSS algorithm was used to spot variable-length activity events, such as intake gestures. The search algorithm is illustrated in Section 4.3.1 on page 54.

  The event fusion step was used to combine several FSS detection instances in one recognition result. The fusion selected one event among all concurrently spotted events. To this end, the spotting result can be used with standard classification algorithms, such as hidden Markov

models (HMMs) (Section 5.4, page 87) or a linear discriminant analysis (LDA)-algorithm (Section 8.3, page 158). This event fusion is not intended to combine temporally distributed activity events, such as discussed in Section 2.1.3.

- **Temporal-spatial transformation of features.** This work demonstrates that the temporal event pattern provides important information for the detection and classification of activity events. This was exploited by the HMM-approach in Chapter 5.

  To assess this information during FSS detection, features were computed for evenly-sized sections within every activity event (Section 4.3.2, page 56). This approach represents a temporal-spatial transformation of event features. It allows to use the FSS detection without an additional HMM classification, such as in chewing event spotting (Section 4.5, page 62). However, this transformation multiplies the feature count by the number of event sections. In this work, three and four event sections were used in Chapter 8 and Chapter 4, respectively.

- **Competitive and supportive event fusion.** Using the classification as event fusion method is an inflexible closed-set concept that requires retraining once a class is added or removed. Moreover, it cannot be used for one-class problems, such as the spotting of swallowing events (Section 9.3, page 185).

  An alternate event fusion approach was proposed, using competitive and supportive event fusion (Section 4.3.1, page 56). For example, the swallowing event recognition was improved by combining independent sound and EMG-based spotting results. In this particular case, events were retained if both, sound and EMG-based spotting agreed (supportive event fusion, Section 4.6 on page 65). The event fusion reduced false positives (insertions) by -30% compared to the independent spotting results (Table 4.9, page 69).

- **Soft-alignment event performance assessment.** To account for imprecise event boundaries of recognition and ground truth a new activity event accounting technique based on a soft-alignment was introduced (Section 4.3.3, page 57). The soft-alignment was implemented using a jitter allowance for matching event start and end between recognition and ground truth (Eq. 4.2 on page 58).

  In this work, a jitter was allowed that corresponds to 50% of the event length. Consequently, if the boundary mismatch between recognition result and ground truth exceeded this jitter, the recognition result was counted as an insertion (false positive).

  Chapter 4 compares the soft-alignment technique to a sample-accurate counting. Differences of less than 10% between recall and accuracy

demonstrate good agreement of the assessments. However, the soft-alignment approach additionally provides the precision of a spotting algorithm.

The recognition procedure was evaluated in different configurations for intake gestures using inertial sensors (data rate: 50-100 Hz), chewing, using a microphone (44 kHz) and swallowing using EMG (0.5-2 kHz) and microphone (22-44 kHz). The following achievements were made for individual sensing solutions:

- **Recognition of intake gestures.** Using a equidistant segmentation, the FSS procedure achieved an average recall of 86% at 28% precision for four subjects (Table 4.3, page 62). The event fusion by comparing events (selecting the most probable event, COMP) boosted the result to 64% precision (+30% increase), while maintaining 80% recall. In a second evaluation using HMMs, precision gained +16% to 73%, while maintaining ∼80% recall (Table 5.7, page 102).

  In conclusion, the COMP approach is applicable and competitive, especially when a large number of classes are available (in this work only four classes were considered). In the case of a low-precision class however, such as the less distinctive "Handheld" (HH) gesture, the HMM approach achieved almost +20% increase in precision (to 59%), while COMP reached a +10% increase (to 38%) only.

- **Intake gesture recognition performance.** Intake gestures are affected by an accustomed eating style of every individual. The challenge to recognise these gestures was shown in comparison to object interaction gestures. For those object interaction gestures, the recognition procedure achieved a +10% higher recall compared to intake gestures (Table 5.7, page 102). This difference is explained by the increased variability in intake gestures.

  For intake gestures, the overall best result was achieved by using the recognition procedure with a SWAB segmentation and HMM classification (79% recall and 73% precision, Table 5.7, page 102). This recognition performance demonstrates that the gesture sensing and recognition approach is applicable for the intake cycle recognition, as summarised in Section 2.1.3 below.

- **Recognition of chewing.** Chewing events (corresponding to the mandible closing phase of chewing cycles) were spotted for two food-texture groups: wet- and dry crisp. For the chewing recognition, an event fusion based on logistic regression (LR) improved the event detection by

+18% in recall (to 93%) and +11% in precision (to 52%) (Table 4.6, page 67).

The low precision (52%) was attributed to an insufficient annotation of chewing cycles. As Figure 4.5 on page 65 illustrates, no ground truth information was available for the "cleanup"-phase after chewing sequences. Consequently, chewing events, retrieved in these sections were counted as errors. In a followup work, this issue was resolved (Section 8.2.3, page 157). In conclusion, chewing events of wet- and dry crisp texture were robustly detected and discriminated using the event spotting procedure.

- **Recognition of individual foods from chewing events.** The aforementioned recognition approach works for a texture-based grouping of foods. However, if similar-texture foods (such as lettuce, apple or potato chips) were discriminated, the FSS showed confusions between the foods. This is indicated by a low precision (∼35%) at a recall of 80% in Figure 8.3 on page 164. Utilising a classification-based fusion step improved precision by +5% to +10% (Figure 8.3, page 164).

- **Recognition of swallowing.** EMG and sound data were considered independently and in combination (feature-level fusion) for the spotting of swallowing events (using SWAB segmentation of EMG time-series, Section 9.3.2, page 186). For all combinations of EMG and sound, a high sensitivity was observed, yielding FSS recalls of 73%–84%. However, precision was very weak (15%–18%). While the feature-level fusion obtained a marginally higher precision (+1%), the event fusion removed more than 50% of the insertions, precision was 31% at 65% recall (Table 9.3, page 188).

- **Swallowing detection performance.** Compared to a detection based on EMG signal intensity, the spotting performance incurred only 50% of the insertions (Table 9.3, page 188). In conclusion, EMG and sound data are most relevant for the swallowing event detection when considered in combination using event fusion. Head and neck movements disturb the swallowing detection, resulting in insertion errors.

  Further work is needed to evaluate alternative sensors and recognition features (Section 4.7.4, page 71). The results presented in Chapter 9 were the first quantitative evaluation published on swallowing detection performance.

2.1.3. Fusion of temporally distributed activity events

In the effort to combine the activity event recognition from all sensing solutions (according to the thesis aims in Section 1.4), the following achievement were made.

- **Model of intake cycle.** A temporal model was developed to describe food intake cycles. The model represents a temporal composite of an intake gesture, multiple chewing events (chewing sequence), and swallowing events (intermediate and final swallows) (Eq. 10.4, page 208).

  The implementation of this concept requires that event recursions are resolved. These occur for chewing-swallowing event repetitions within an intake cycle.

  With this model, the plausibility of specific intake types (gesture type, chewed food, swallowing frequency) was determined (Figure 10.2, page 210). For example, eating lettuce (chewing events) is performed using fork and knife, rather than bare hands (intake gesture).

- **Intake cycle model evaluation.** The model was evaluated using probabilistic context-free grammar (PCFG) parsing (introduced in Section 10.1.3 on page 205). Individual grammars were derived for chewable foods (Eq. 10.6, page 210) and drinking (Eq. 10.7, page 211).

  An evaluation using annotated event data showed parsing recalls of >80% for 9 out of 11 intake types (10 foods and drinking). Precision was between 55%-100% for a non-recursive chewing-swallowing grammar (Figure 10.3, page 213). Using a refined grammar with recursion (Eq. 10.10, page 213) increased precision by up to +40% (Figure 10.4, page 214). This result indicates the relevance of recursion modelling for foods with intermediate swallows (foods with wet compartments as in lettuce and lasagna).

  Finally, food-texture grouping showed that the PCFG approach is feasible to detect solid foods as well as to identify drinking (86% recall and 95% precision, for 8 foods and drinking, Figure 10.5, page 215).

- **Estimation of chewing sequence phases.** Chewing sounds alter within a chewing sequence due to the progressing food destruction (Section 7.1, page 134). The existence of phases (temporal clusters of chewing events) was investigated using a chewing sequence model (Section 7.2.3, page 138).

  An analysis of four foods with different texture found a two phase result for all, with a shorter first phase (30%-40% of the sequence length, Section 7.4, page 143). The two-phase structure was confirmed by classification rates of ∼80% for potato chips and chocolate (Figure 7.5, page 146). This result was initially expected for the dry texture of potato chips only. In contrast, the strict temporal phasing was not confirmed for apple and lasagna (∼60% classification rate).

  In conclusion, foods that show a fast deterioration during oral breakdown (due to wetting with saliva or melting) adhere to a two-phase

structure. This result can be used to derive recognition models adapted for specific food groups and sequence phases.

2.1.4. Estimation of eating behaviour

The final part of the thesis contributions addresses eating behaviour information derived from on-body sensing and recognition solutions. Specifically, food type and amount estimation was considered, according to the thesis aims in Section 1.4.

- **Food category and type extraction from chewing.** Acoustic emissions during food breakdown reflect the material texture (Section 6.1.4, page 114). The classification of chewing events from 19 foods and three individuals, resulted in an average classification rate of 83%. Figure 3.3(b) on page 37 visualises the classifier confusion. While these foods were chosen to represent a large variety of textures, they also included similar ones, such as in lettuce, apple and carrots. This classification result demonstrates the discrimination capabilities of acoustic chewing patterns.

  In contrast, using the spotting procedure, FSS precision was 30%-40% at a recall >80% in three foods, see Section 2.1.2 above. Grouping foods according to their texture simplifies the recognition task (as shown in Table 4.6 on page 67). The continuous recognition of individual foods was further improved by chewing sequence information (majority vote for intake cycles), see below.

- **Food identification from intake cycles.** Based on a fixed-size sliding-window approach, chewing recognition rates ranged between 66% and 86% in four foods (Table 6.4, page 125). In this approach a sound energy-threshold was used to identify chewing events. Majority voting for all chewing events in a chewing sequence led to performance gains of up to +20% (Table 6.5, page 125).

  The performance gain is even more profound, if a chewing event annotation is available to train the classifier. Based on the three-food recognition cited earlier (Chapter 8, page 153), a classification and chewing sequence vote resulted in >90% correct identified sequences (+50% increase in precision, to ∼70%, Figure 8.4, page 165). In conclusion, sequence information is vital for an accurate identification of individual foods.

  The classification and sequence vote were demonstrated to work with up to four foods. To increase the number of foods, further information of the intake cycle was used. This included the intake gesture type and the chewing-swallowing interaction, as summarised in Section 2.1.3 above.

- **Food weight estimation from chewing.** Food weight of single habitual bites was estimated from the chewing event microstructure. Variables

derived from chewing sequences, such as the total number of chewing events, showed high correlations (up to 0.96) with bite weight. Figure 8.6 on page 167 shows the variable relevance for three foods and eight individuals.

Bite weight was predicted with an average error of 19% for apples, 28% for potato chips and 31% for lettuce (Table 8.2, page 169). The error obtained for apples is in the range of natural fruit weight variation, hence it is comparable to the amount quantification in simplified self-reports (without weighting).

Degradation due to fruit storage, addition of toppings (e. g. for lettuce) increase uncertainty on the correct weight (Section 8.6, page 169). Nevertheless, these results are encouraging to investigate further foods and combine the prediction with the swallowing bolus volume classification as summarised below.

- **Food volume classification.** Swallowing events were classified according to bolus volume in a study using fixed bolus sizes (Table 9.4, page 190 summarises the considered food items). The classification rate for two bolus volumes, large volume (15 ml water) and small volume (5 ml water, spoonful of yogurt and 2 cm^3 bread), was ∼70% for five participants (Table 9.6, page 194). Stethoscope sound provided the best-discriminating features.

2.2. Conclusion

Diet monitoring is relevant for clinical interventions on eating behaviour and prevention programs on weight coaching alike. Self-assessments, however, obtain a low respondent compliance due to the high effort required in maintaining them. These assessments are not appropriate when following a normal everyday life.

An essential change in the monitoring paradigm was proposed in this work: freeing the individual from manual logging of food intake. Specifically, this work evaluated new techniques for on-body dietary monitoring. Based on the summary presented in Section 2.1 above, the following conclusions were made:

- On-body sensing and recognition solutions provide vital information for diet monitoring: Activities related to food intake can be monitored, namely, intake gestures (using inertial sensors at lower arms and torso), chewing (using an ear-worn microphone to record food breakdown sounds) and swallowing (using hyoid EMG and a stethoscope microphone).
- The evaluations showed that intake gestures and chewing events are robust sources of information. For intake gestures, a recognition performance of 79% recall and 73% precision was obtained. For chewing events, recall was 93% at 52% precision. Swallowing event detection requires further investigations (65% recall at 31% precision).
- A recognition procedure for activity event spotting and event fusion was introduced and evaluated, using various sensing modalities (inertial sensors, EMG, sound). The recognition procedure is applicable for activity event spotting and identification.
- Food categories and chewing events, aligned to food-texture groups, can be recognised from chewing sounds (93% recall, 52% precision for two groups: wet- and dry-crisp texture). Event fusion methods improve the recognition result (precision of 70% with recalls above 90% for three individual foods). The classification of chewing cycles demonstrates the discrimination capabilities of acoustic chewing patterns (classification of 19 foods resulted in an accuracy of 83%).
- Bite weight was estimated from the chewing event recognition with an average error of less than 20% for apples. Foods with low bite weight ($<4\,\mathrm{g}$), such as potato chips and lettuce resulted in 30%-35% prediction error.
- Both, food type and amount estimation depend on the segmentation of intake cycles in continuous data. The intake cycle is a vital step to combine the activity event information from three sensing solutions. Robust results were achieved for the recognition (86% recall, 95% precision for eight solid foods and drinking).

2.3. Limitations and relevance

Based on the sensing solution results obtained, it was concluded that an estimation of intake timing is a feasible recognition task. However, it was not evaluated in this work. Furthermore, the estimation of energy intake was not investigated. In the practice of self-reports, energy intake is estimated from food product information. However, these information details (exact food product, brand and ingredients) were not automatically recognised in this work. Nevertheless, an average energy level could be derived from the food type and amount estimation presented in this work. Further work in this direction is needed.

This work has focused on the evaluation of sensing and recognition methods using a small number of foods (up to 19) only. The recognition approach, using food-texture grouping as well as the partitioning of chewing sequences were initial attempts to expand and generalise the food set. These approaches require further investigations and should be applied in larger food sets with various textures.

Recognition performance was evaluated in presence of noise and similar foods. Therefore, the results demonstrate that initial systems can be realised instantly for a fixed set of pre-trained foods. Moreover, an early practically applied system may be allowed to ask the user if recognition confidence is low. It can offer a choice of the most likely foods or food categories to support the monitoring.

Eating behaviour is an accustomed habit that differs strongly between individuals. Consequently, personalised pattern models were required for all recognition solutions in this work. Ongoing and future research on diet monitoring needs to address this issue.

2.4. Outlook

This work has opened new and promising research perspectives that may converge to applicable solutions in the near future. To raise awareness on diet monitoring among researchers working in pervasive healthcare and exchange research results, an international symposium series was initiated in 2007, called "e-Nutrition" (http://www.e-nutrition.org).

Further research should address the following challenges:

- A combination of chewing and swallowing recognition will resolve ambiguities in detection of both activities, since chewing and swallowing are tightly coupled. Moreover, this combination has potential for the food amount estimation, based on the results presented in this work.
- Selecting appropriate features is a challenging task even in classification problems. The number of available features typically exceeds the required amount of modelling data by far. This problem is exacerbated by expensive datasets, e. g. due to the annotation requirements for chewing. In this work, promising results were achieved using a feature selection procedure adapted to activity event spotting (Chapter 8). Further investigations on feature selection are needed for all sensing solutions.
- In order to validate diet monitoring solutions, systems should be evaluated in typical use scenarios, as soon as technically feasible. The evaluated solutions for intake gestures and chewing have reached this maturity. Further work on swallowing recognition is needed.
- The combination of on-body and environmental sensors offers vast potential for resolving shortcomings that both approaches have independently. While RFID technology is promising to identify foods, knowing the location will provide information needed to reduce the set of potential foods, e. g. from the menu in a restaurant. Even plausibility checks are useful, such as eating on an gym ergometer is unlikely, while drinking is. These concepts will decrease the recognition complexity. Moreover, information on food preferences, e. g. from food frequency questionnaires, are a sensible approach to reduce the set of likely foods and intake types.

3

Automatic dietary monitoring: on-body sensing domains

Oliver Amft and Gerhard Tröster

submitted to IEEE Pervasive Computing, June 2008.

Abstract

Automatic dietary monitoring aims to recognise eating behaviour from sensors. This information is required to adapt and personalise feedback of weight and diet coaching programs. On-body sensors can be used for continuous monitoring of eating behaviour.

3.1. Introduction

The balance between energy in consumed food and energy expenditure is a key to success for good long-term health. However this balance is challenging to maintain as alerted by the pandemic of overweight and obesity. Worldwide more than one billion adults are overweight and 400 million are obese (WHO statistics 2005, http://www.who.int/topics/obesity/en/). By 2015 WHO predicts an increase to more than 700 million obese patients.

Weight and diet management programs have been established to support weight changes. The programs coach individuals to improve eating behaviour by daily or weekly status feedback, meal suggestions and behaviour recommendations. However, only 20% of the individuals that achieved at least 10% reduction in body weight, are able to maintain the new weight for one year [21]. From these outcomes researchers have concluded that support durations of two to five years are needed to raise success of coaching programs. Practicability of current programs is their main limitation. Participants have to complete detailed self-reports on eating behaviour, while maintaining their lifestyle and eating behaviour modification on a day-to-day basis. Besides a personal profile, self-reports are the unique source of information to adapt and personalise feedback and recommendations for coaching programs participants. Unfortunately, self-reports have a high bias and are hard to maintain.

Automatic dietary monitoring (ADM) aims to replace manual reporting of eating behaviour with a sensor-based estimation. In this article we discuss requirements and options for on-body sensing of eating behaviour. We demonstrate that indeed, on-body sensor information can resemble some information of self-reports. These initial solutions towards ADM are research prototypes and consequently not yet comfortable enough for long-term (months and years) continuous use. However, they highlight crucial benefits of on-body sensing and the ADM concept for future eating behaviour coaching.

Besides energy, food provides essential nutrients for the organism. Eating disorders, such as binge eating, underline psychological influence on eating. As a consequence, strict everyday energy balance is not the primary optimisation goal in food choice. Self-reports capture these aspects in a set of items to answer.

Nevertheless, self-reports and similar manual assessments of eating behaviour suffer from a number of shortcomings. These include the respondents motivation to complete questionnaires, awareness for food intake, snacks in particular, as well as memorising, perception and literate capabilities [22]. Reporting errors range between 50% under- and overestimation [17].

3.2. Towards ADM - What will it help?

Researchers have proposed a broad range of alternate self-reporting solutions and, more recently, attempted to automatically recognise eating behaviour from ubiquitous sensors (see box: diet monitoring approaches). We envision that a sensor-based automatic monitoring will release individuals from a stringent manual reporting and provide more robust eating behaviour information. Hence, ADM will simplify long-term coaching programs on eating behaviour that are urgently needed, and infeasible using current, manual monitoring techniques.

To replace manual logging, ADM systems shall supply information on eating behaviour, as self-reports conceptually intend. This information - the dimensions of eating behaviour - include:

- intake timing,
- food type or category,
- food amount, and
- energy content (calories)

of every consumed food piece. Moreover, ADM systems shall be applicable for long-term use regarding operational requirements, robustness and user comfort.

3.2.1. Challenges for ADM

The challenge for self-reports and ADM solutions is to capture the diversity of consumed foods and the variability in personal eating behaviour. For example, energy intake is most accurately determined if the calories of consumed food products are reported. However, even with direct calorie reporting, energy estimation requires additional information, including amount of consumed food and whether certain changes had been made (e.g. addition of a lettuce dressing). Furthermore, calorie reporting is often infeasible for self-prepared meals.

Personal preferences regarding choice of food or food category and meal schedule exist. ADM solutions can integrate these preferences as prior information for eating behaviour estimation. Nevertheless, actual eating behaviour is influenced by varying environmental and psychological aspects, including constraints in food availability, social interaction during meals, and emotions.

A particular challenge for ADM solutions is to robustly recognise eating behaviour from sensor data. No single sensor, independent of its location and recorded physiological or activity information, can capture all dimensions of eating behaviour. This challenge is reflected in restrictions of initial ADM approaches. Typically, these solutions emphasise particular dimensions of eating behaviour, such as recording consumed food amount using a weight scale,

while restricting location to the weighting-enabled table. Moreover, solutions that rely on environment-embedded sensors only, raise the challenge in assigning measurements robustly to one person. While these works represent relevant advancements towards ADM, we concluded that a multimodal sensing approach will support the monitoring of several eating behaviour dimensions.

3.2.2. Benefits of on-body sensing for ADM

Monitoring eating behaviour continuously and independent of a particular location is a vital property of an ADM system, since modern lifestyles imply many location-changes, for work and leisure purposes. Consequently, food is consumed in various locations and in transit. Solutions that depend on a particular environment, such as a home location, will miss a snack "in between" or an entire business lunch. Such partly coverage limits the effect of behaviour coaching severely and could lead to misleading recommendations. Hence coaching requires a continuous monitoring that covers all daily situations.

On-body sensors can provide continuous monitoring of eating behaviour, independent from dedicated sensor-enabled environments. In contrast to environment-embedded sensors, on-body sensors allow a direct association of recorded information to the wearer.

3.3. Diet monitoring approaches

Classic dietary monitoring techniques require manually recording of eating behaviour. Among these assessments, respondent self-reports are intended to capture every food intake as required by weight and diet management programs. However, low adherence and accuracy restricts the report validity, and consequently the feasibility of coaching programs that use self-reports [6].

Multiple attempts were made to simplify tedious and error-prone logging. Studies confirmed that electronic devices, as replacement for paper-based self-reports could not reduce reporting errors, e.g. [24].

We highlight here some alternate *manual* methods for capturing eating behaviour information. Jennifer Mankoff and her colleagues scanned shopping receipts to simplify diet monitoring [14]. MyFoodPhone Nutrition, Inc. (http://www.myfoodphone.com) introduced commercial service to assess food intake from mobile phone pictures. Katie Siek and her colleagues used bar codes and voice recordings to replace self-report questionnaires [18].

For all manual dietary monitoring, participants of a coaching programs are asked to record their eating behaviour. In contrast, *automatic* dietary monitoring aims to estimate eating behaviour without the participant in the loop.

Approaches towards automatic dietary monitoring can be categorised by their sensing approach into environment-embedded, on-body and implantable solutions. A few pioneering solutions have been developed using environment-embedded sensors. Keng-hao Chang and his colleagues developed a dining ta-

ble that detected the weight of foods and identified food bowls from radio-frequency identification tags (RFID) [7]. Jiang Gao and colleagues recognised arm movements to the mouth from surveillance video [9]. In a general evaluation of RFID for home monitoring, Donald Patterson and colleagues estimated morning activities, including breakfast consumption timing [16].

Implantable solutions, such as in-oral sensing [19], could provide more precise information on the eating process. However, this solution is technical challenging and alters oral sensation. Hence, it appears infeasible for long-term diet monitoring.

3.4. Evaluation of on-body sensing solutions

We analysed on-body sensing approaches and modalities to evaluate the benefits for ADM. The analysis covered both, activities related to eating and physiological responses to food consumption (see Figure 3.1 for an overview).

To assess the relevance for ADM, we evaluated sensing solutions regarding eating behaviour information and wearer comfort. For the first evaluation, we analysed what particular dimensions a solution can estimate as well as their limitations.

As summarised before, the estimation of energy intake requires at least food category and amount information, combined with a more complex inference. Hence energy intake was not considered in the evaluation. Table 3.1 and 3.2 summarise our evaluation and review results on dimensions of eating behaviour, particular limitations and comfort for all sensing solutions.

From all sensing solutions we selected three activity-based solutions: intake gestures, chewing and swallowing. These activities represent a temporal description of food intake and permit the recognition of intake cycles. In our analysis of these solution, we evaluated estimation performances regarding food category and amount in user studies. Here we used a Naïve Bayes classifier preceded by linear discriminant feature extraction, to obtain person-adapted performances. To ensure robustness of results, we deployed a five-fold cross-validation.

3.4.1. Intake gestures

Movements of the upper body (arms and trunk) are required for most forms of intake. They can be separated into a coarse preparation of food or beverage items, such as unpacking, cooking and plate loading, and actual food intake phase. Food intake includes movements to fine-cut and maneuvering prepared piece to the mouth. In the intake phase, tools such as fork and knife are used. We focused our recognition approach on these intentional arm movements. Inspired by the observation that gestures reflect intake types (eating or drinking) and food category (from tools used), intake gestures provide timing and food category information.

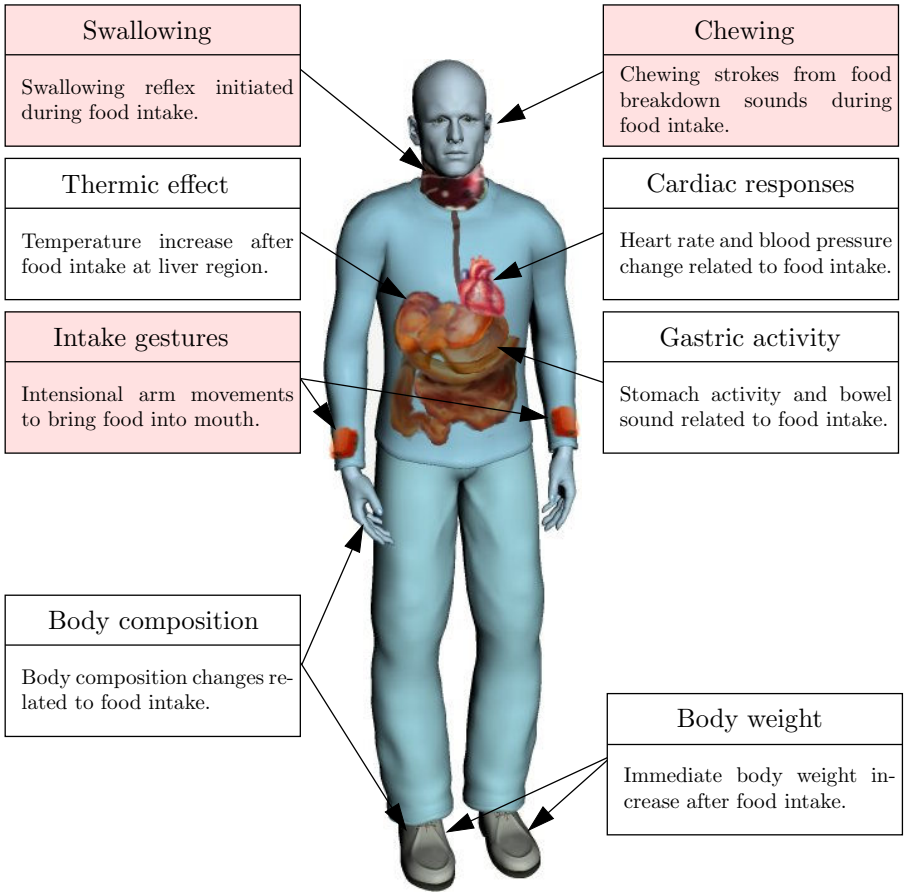


Figure 3.1. Major on-body sensing solutions for food intake. We selected intake gestures, chewing and swallowing to estimate food intake cycles.

Intake gestures can be recorded using inertial sensors at wrists and upper back. We derived a comfortable recording setup by integrating commercial motion sensors (http://www.xsens.com) in a jacket (see Figure 3.2(a)). The sensing units contain three-dimensional acceleration, gyroscope and magnetometers.

To evaluate the discrimination performance of different gestures, we conducted a study with four students eating foods from four different movement categories [12]. The categories included, eating lasagna with fork and knife, drinking from a glass, eating a soup with a spoon, and eating bread using one hand only. The students ate all foods in random orders, without particular movement instructions. During recording breaks, they performed further

activities (reading newspaper, using phone) to promote natural movement variability. In total, 1020 intake gestures were recorded in 4.68 hours. Using the classification procedure, we obtained an accuracy of 94%. Figure 3.2(b) shows the result for all gesture categories. Only temporal features from arm acceleration sensors were used for this recognition. We observed that the temporal structure of intake gestures can be modelled by computing features in four sections of each gesture instance. Without these features we achieved similar classification results, but required all modalities of the motion sensors and hidden Markov models [12].

While the motion sensor jacket was a useful research prototype, we plan to replace it with less complex sensors. The classification using only acceleration shows that sensors can be reduced. However, already in the current study wearers reported that the jacket was comfortable for sitting activities.

3.4.2. Chewing

Chewing strokes (jaw opening and closing) can be monitored from masseter and temporalis muscle activation using surface Electromyography (EMG). Since muscles are located in exposed facial regions, privacy cannot be retained with this technique.

Nevertheless, we found a feasible solution: chewing generates sound emissions during food breakdown that conduct through mandible, skull and body tissue. Using an ear-attached microphone, we recorded these chewing sounds. From their acoustic profile during chewing we classified foods [3] and analysed different microphones and ear-device cases. Figure 3.3(a) shows a device, where a miniature microphone was embedded into a standard headphone case. In another construction, we used an ear-pad case. With this setup we studied how users perceived the ear occlusion. Smaller pads reduced occlusion and increased user comfort, however it reduced the signal to noise ratio too. Users found the headphone device convenient, especially when they were used to wearing similar models with music players.

We studied the scalability of food classification using various foods. We asked three male students with natural dentation to eat 19 standard foods as they were used to. In several sessions we recorded chewing using a low occlusion ear-pad device. In this setup, the wearer could understand office-room conversation in 2 m distance. Totally, we obtained ∼12000 chewing strokes in 5 hours of data. For the classification of all foods, we obtained a high accuracy of 80%. For the headphone case, we observed an accuracy drop by 5% to 10%, depending on environmental noise. As features, spectral energy bands, cepstral and linear predictive coefficients were used, detailed further in [5]. We selected these features based on robust results obtained with earlier recordings.

Figure 3.3(b) shows a colour-coded classifier confusion. This representation provides a quick assessment of the classifier performance for all foods. Confusions (non-white colour besides the main-diagonal) indicate acoustic groups

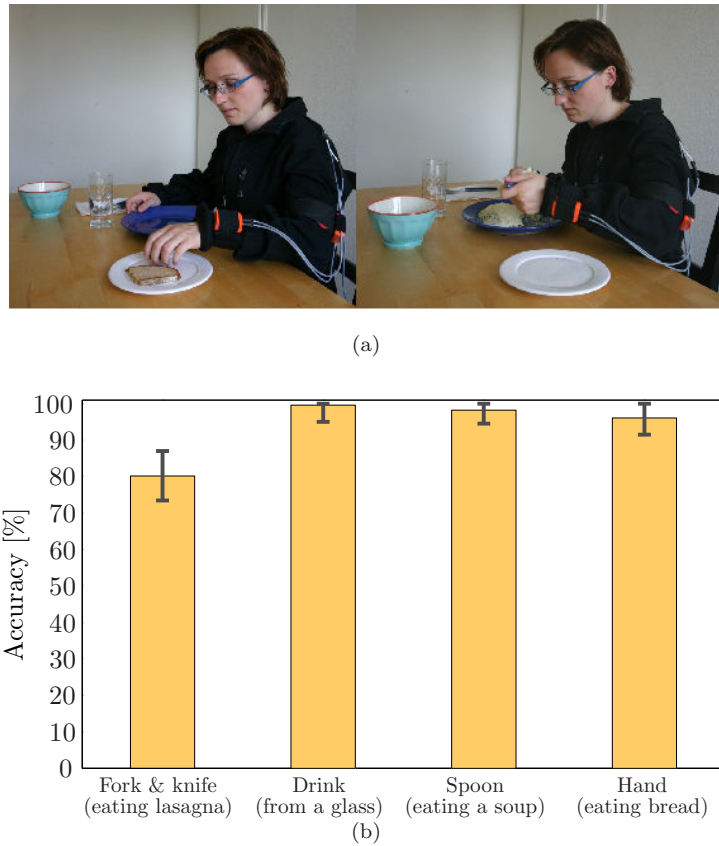
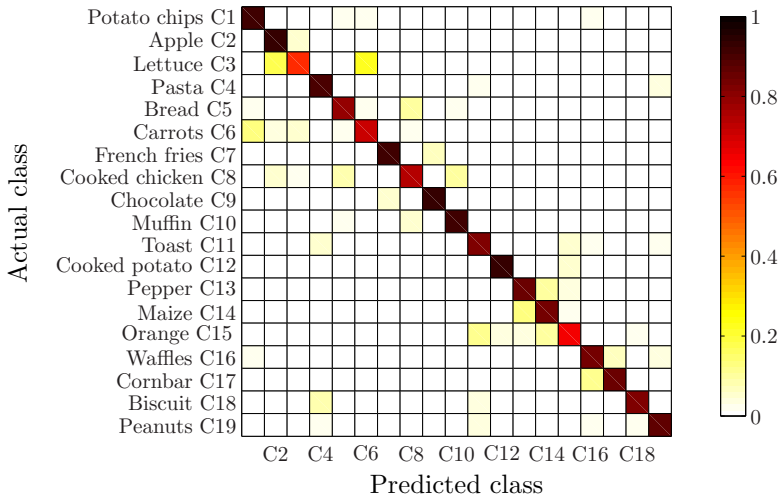


Figure 3.2. Intake gestures: (a) User wearing the motion sensor jacket during eating. (b) Classification rates for different intake gestures, including inter-person min-max values.

among foods. For example, lettuce is partly confounded with carrots and apples, indicating that sound patterns are primarily controlled by food texture.

Food texture was our main selection criteria in this evaluation. The set includes similar textures, e.g. lettuce, apples, and covers a broad variety of materials and preparation styles, e.g. cooked meat. While this result demonstrates texture-based discrimination capabilities, we deploy chewing sound recognition for nutritional-relevant food groups in the food pyramid. For example, fruits and vegetables can be grouped, based on a similar "wet-crisp" texture and recognised in continuous sound data [5].

(a)



(b)

Figure 3.3. Chewing sounds: (a) Miniature microphone integrated in headphone case. (b) Colour-coded classifier confusion for chewing of 19 foods. This classification confirms individual sound patterns in foods.

3.4.3. Swallowing

Swallowing often occurs unconsciously during a day, with increased frequencies during food intake [13]. After food was converted into a bolus by chewing, tongue movements initiate a reflex of throat muscles that propel a bolus through the throat into the oesophagus.

Most swallowing studies analyse abnormal swallowing in laboratory set-

tings. Since tongue and oesophageal movements are challenging to monitor with on-body sensors, we focused on the swallowing reflex using sensors at the throat. We developed a set of collars to investigate different sensing modalities.

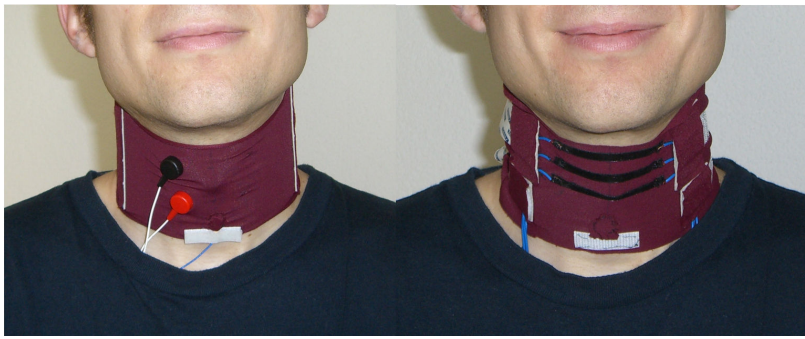


Figure 3.4. Swallowing: Collar prototypes with integrated surface EMG and microphone (left) and carbon-loaded rubber elongation sensors (right).

In one collar system, we monitored textile elongation to detect skin movement during swallowing (see Figure 3.4). Elongations occur for male subjects mainly, since females have a less prominent Adam's apple. Moreover, the strain sensing collar required accurate positioning. Signals were impaired by movements of neck and collar itself.

In a second solution, we combined surface EMG and a stethoscope-like microphone, to monitor both, throat muscle contraction in deep tissue layers and swallowing sounds (see Figure 3.4). While EMG is impaired by other throat muscle activations, sound pattern is influenced by food viscosity. We combined both modalities to determine swallowed food amount.

We used the sensors with five students eating foods and drinking water as they naturally do. From several sessions we analysed totally 4.85 hours of data and 868 swallows [4]. We discriminated low swallowing volume (5 ml water, spoonful yoghurt, $2\,cm^3$ bread pieces) vs. large volume (15 ml water) with an accuracy of 73%. Similar to chewing sound classification, swallowing volume discrimination required a spectral feature set, described in [4].

As expected, users found both collars uncomfortable for long-term monitoring. Our current work aims to replace the collar prototypes with more convenient systems, such as a collar-shirt.

3.5. Further on-body sensing options

We analysed whether further sensing solutions could provide eating behaviour information. Our goal was to review activities and physiological responses closely related to food intake and summarises available knowledge.

3.5.1. Gastric activity

Swallowed food arrives at the stomach after ∼15 minutes. It is subsequently decomposed by stomach muscle contractions. Further digestion in the gastrointestinal tract incurs time delays in the range of hours with respect to the originating intake and thus is far less deterministic.

On-body sensing options are rare for late stages of digestion. The electric and magnetic fields of stomach muscles were captured by researchers using laboratory setups, such as Electrogastrography (EGG) [1]. However EGG has not reached broad clinical acceptance. Furthermore, abdominal sounds from food movement in intestines can be assessed by stethoscope. While bowel sounds are typically loudest after fasting, a relation to intake was recently confirmed for laboratory settings [23]. All measurements are perturbed by heart and respiration activity as well as body movement.

3.5.2. Thermic effect of food intake

The thermic effect of food intake (TEF) is a thermogenesis in response to intake above resting metabolic rate. Although TEF is the smallest component in human energy expenditure, researchers studied its relation to intake restraint and obesity.

Optimal TEF assessment requires a respiratory chamber to measure changes in resting metabolic rate before and after intake. TEF starts immediately after food reached the stomach and peaks after ∼60 minutes. For unrestrained eating in normal weight individuals, skin temperature above the liver increased between 0.8 and 1.5 K [20]. TEF depends on regularity of intake and is lower for irregular intake [8].

3.5.3. Body weight

Food intake is associated with immediate gain in body weight. If weight is monitored, intake timing and food amount can be determined. Typical meals are in the range from 50 g, to 500 g or more, e.g. for multiple course menus. Snack sizes are in the range of a few grams (5 g and more) but could contribute an important share in intake, such as snacks from high-calorie foods or sweets.

In contrast to classical body weight measurements once a week, intake-related weight changes require a continuous weighting. Shoes would serve ideally for this purpose. Compared to a scale, the challenges for shoe-based weighting are related to a low mechanical profile, high torsion flexibility and low system weight. Weight must be measured from foot force distribution in (even short) moments, when the user is standing.

Classic load cells do not fulfil the mechanical constraints. Pressure sensing arrays struggle at resolution requirements. Capacitive in-shoe gait measurement systems, have an error of 2.7% [11], corresponding to 1890 g for a 70 kg person. We studied arrays of Force Sensitive Resistors (FSRs) and observed

even larger errors due to signal noise and shoe torsion. In conclusion, a wearable measurement of body weight remains unsolved.

3.5.4. Cardiac responses

After meal intake, blood is redistributed to the stomach and lower gastrointestine tract. Studies reported an increase in heart rate 30 minutes after intake [15].

Blood pressure is known to depend on food composition, especially on salt and sugar. Classic blood pressure measurements require cuff-based solutions that are inconvenient for everyday use. Novel cuff-less approaches, e.g. based on pulse arrival time, are part of ongoing research.

The responses depend on a variety of aspects, including physical activity, body posture, fasting time and time of day.

3.5.5. Body composition

Single food intake modifies body composition immediately. Clinically, body impedance is measured between hand and foot. In a laboratory setting composition altered 30 minutes after intake [10]. The effect depends on gender and food type. Further investigations are needed to study the validity of composition assessments. Movement artefacts make the effect impractical for ADM systems.

Table 3.1. Evaluation summary of on-body sensing solutions for ADM.

Sensing solution	Dimensions of eating behaviour	Notes on sensing modalities and comfort for everyday use
Intake gestures	**Timing**: cont. recognition of four gesture types, R: 79%, P: 73% [12][a]. **Food type**: movement related to food category, recognition of four types, C: 94%[a]. **Food amount**: N/A. **Limit**: Recognition errors occur for arbitrary arm movements to the head and non-typical long gestures.	**Modalities**: inertial sensors at lower arms and upper back. **Comfort**: conveniently integration in smart clothing or accessories, e.g. watch or bracelet.
Chewing	**Timing**: cont. recognition for two food categories, R: 93%, P: 52% [5][a]. **Food type**: chewing strokes, recognition of 19 foods, ear-pad device, C: 80%[a]. **Food amount**: N/A. **Limit**: Environment noise and low ear occlusion perturbed recognition.	**Modalities**: ear-pad microphone [3], similar to ear-attached hearing aid devices. **Comfort**: related to ear occlusion, convenient for headphone case.
Swallowing	**Timing**: cont. recognition of four bolus types, R: 65%, P: 31% [4][a]. **Food type**: N/A. **Food amount**: bolus volume, recognition of low vs. large volume: C 73% [4][a]. **Limit**: Measurements impaired by head and neck movements, chewing and speaking.	**Modalities**: surface Electromyography (EMG), stethoscope microphone or similar acoustic transducer [4], skin movement (this publication), throat impedance or capacitive sensing. **Comfort**: large sensor-collars are uncomfortable, improvement with collar-shirt expected.

[a]R: Recall, P: Precision, C: Classification accuracy.

Table 3.2. Evaluation summary of on-body sensing solutions for ADM (continued).

Sensing solution	Dimensions of eating behaviour	Notes on sensing modalities and comfort for everyday use
Gastric activity	**Timing**: stomach activity increases ∼15 min after intake [1], dependencies on duration unclear. **Food type & amount**: relations unclear. **Limit**: Approaches require strict laboratory settings, not feasible for non-stationary monitoring.	**Modalities**: Electrogastrography (EGG) [1], impedance gastrography, bowel sounds [23]. **Comfort**: electrodes/sensors need tight attachment to chest or belly.
Thermic effect	**Timing**: temperature increase of 0.8 to 1.5 K ∼60 min after intake [20], dependencies on duration unclear. **Food type & amount**: relations unclear. **Limit**: Temperature depends on regularity of food intake [8]. Relationship is altered by unrestricted physical activity and environmental temperature.	**Modalities**: temperature sensor. **Comfort**: requires attachment to skin in proximity of the liver.
Body weight	**Timing**: body weight increase immediately, duration of intake not assessable. **Food type**: N/A. **Food amount**: weight monitoring at resolution of <50 g for meals and <5 g for snacks. **Limit**: Shoe-based weight measurement requires still-standing and is impaired by uneven floor surfaces.	**Modalities**: shoe-embedded weight or force sensor array (unsolved). However, current in-shoe force sensors do not provide appropriate resolution [11]. **Comfort**: related to shoe torsion flexibility and weight.
Cardiac responses	**Timing**: heart rate: increases ∼30 min after intake [15], increase duration up to 3 h (laboratory). **Food type & amount**: heart rate: relations unclear; blood pressure: influenced by salt and sugar. **Limit**: Relationships altered by physical activity, fasting time, time of day. Measurements perturbed by physical activity.	**Modalities**: Electrocardiogram (ECG) chest strap or close-fitting shirt (heart rate), cuff-based or cuff-less monitor (blood pressure). Research on cuff-less approaches is ongoing. **Comfort**: cuff-based monitor impractical for long-term use.
Body composition	**Timing**: body impedance altered ∼30 min after intake [10] in clinical settings, duration unknown. **Food type & amount**: relations unclear. **Limit**: Measurements perturbed by body movement.	**Modalities**: body impedance using electrodes. **Comfort**: hand to foot electrodes potentially inconvenient.

3.6. Intake cycle modelling

Intake gestures, chewing and swallowing represent a temporal description of food intake. We selected these solutions to construct a hierarchical recognition procedure to identify intake cycles as shown in Figure 3.5(a). In our approach, an intake cycle stretches from an intake gesture (taking a bite of food) until swallowing of this bite. We deployed individual detectors to recognise activity events from each sensing solution.

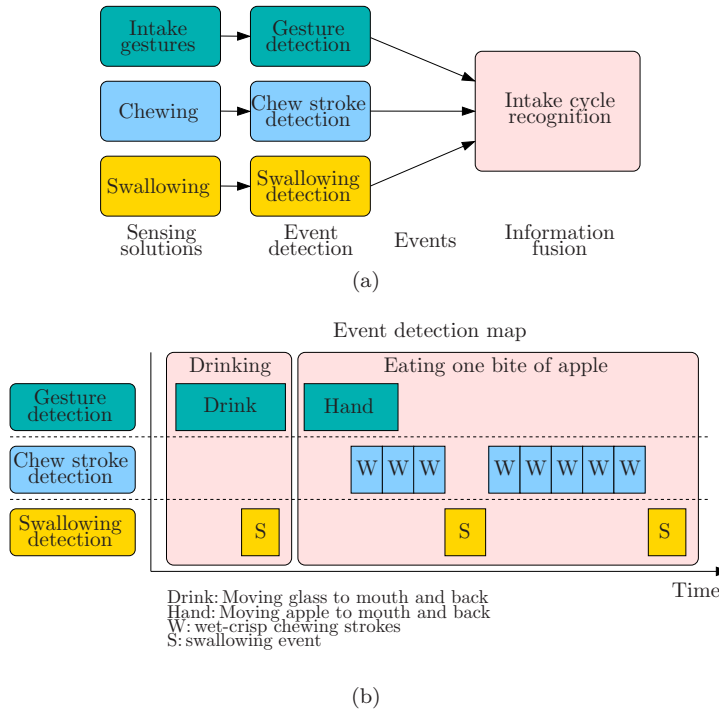


Figure 3.5. Intake cycle recognition approach. (a) Hierarchical recognition procedure, e.g. for food category estimation. (b) Intake event sequences for drinking and eating one bite of apple.

Figure 3.5(b) illustrates two event sequences, representing intake cycles for drinking and eating. To recognise intake cycles from activity events, we implemented a probabilistic context-free grammar (PCFG) parser [2]. The parser estimates the fit of event sequences to an intake grammar. We derived grammars for particular food categories, such as drinking and eating fruits. With PCFGs, recursive event structures can be modelled, such as the recursion of chewing and swallowing events shown in Figure 3.5(b) for eating apple.

The approach provides a number of benefits for estimating of eating behaviour:

- The temporal fusion of individual food category estimations from intake gestures and chewing permits more diverse food categories.
- Estimation errors of individual sensing solutions can be complemented by the fusion.
- At event level, the hierarchical recognition allows simplified synchronisation of sensing solutions with different sampling rates.

ADM aims to replace manual diet monitoring that is currently in practice for weight and diet coaching. Hence, eating behaviour information that is obtained using manual monitoring provide requirements and benchmark for ADM solutions.

In our evaluations, we observed that recognising intake activities from on-body sensors provides information on intake timing, food category and amount. Moreover, by using on-body sensors, information is obtained continuously, independent from particular locations. Nevertheless, most on-body sensing solutions have limitations regarding sensor artefacts and wearer comfort.

By combining selected solutions in a hierarchical recognition, we could compensate estimation errors. Still, this approach refines estimations for food categories only. In comparison to self-reports that include an exact food type reporting, this is a limitation of on-body sensing solutions. Similar restrictions apply for food amount and hence, estimation of energy intake. However, if practical issues and bias of self-reports are considered, even a categorical information indicates ADM benefits. We expect that initially deployed systems will track a small number of food categories, such as fruits and vegetables, related to particular nutritional recommendations. In our studies, we achieved high recognition performances for identifying these categories.

Among all selected solutions, primarily the swallowing solutions lacks in comfort. In our on-going research we aim to replace the current collar prototypes with more convenient solutions. Moreover, we plan to combine on-body and environmental sensing solutions, to leverage the advantages of both approaches.

Bibliography

[1] T. L. Abell and J. R. Malagelada. Electrogastrography. *Dig Dis Sci*, 33(8):982–992, August 1988. doi:10.1007/bf01535995.

[2] O. Amft, M. Kusserow, and G. Tröster. Probabilistic parsing of dietary activity events. In S. Leonhardt, T. Falck, and P. Mähönen, ed., *BSN 2007: Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks*, vol. 13, pp. 242–247. IFMBE Proceedings, Springer, March 2007. doi:10.1007/978-3-540-70994-7_41.

[3] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of chewing sounds for dietary monitoring. In M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, ed., *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing*, vol. 3660 of *Lecture Notes in Computer Science*, pp. 56–72. Springer Berlin, Heidelberg, September 2005. doi:10.1007/11551201_4.

[4] O. Amft and G. Tröster. Methods for detection and classification of normal swallowing from muscle activation and sound. In E. Aarts, R. Kohno, P. Lukowicz, and J. C. Trainini, ed., *PHC 2006: Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–10. ICST, IEEE digital library, November 2006. doi:10.1109/pcthealth.2006.361624.

[5] O. Amft and G. Tröster. Recognition of dietary activity events using on-body sensors. *Artif Intell Med*, 42(2):121–136, February 2008. doi:10.1016/j.artmed.2007.11.007.

[6] L. E. Burke, S. M. Sereika, E. Music, M. Warziski, M. A. Styn, and A. Stone. Using instrumented paper diaries to document self-monitoring patterns in weight loss. *Contemp Clin Trials*, 29(2):182–193, Mar 2008. doi:10.1016/j.cct.2007.07.004.

[7] K.-H. Chang, S.-Y. Liu, H.-H. Chu, J. Y. Hsu, C. Chen, T.-Y. Lin, C.-Y. Chen, and P. Huang. The diet-aware dining table: Observing dietary behaviors over a tabletop surface. In K. Fishkin, B. Schiele, P. Nixon, and A. Quigley, ed., *PERVASIVE 2006: Proceedings of the 4th International Conference on Pervasive Computing*, vol. 3968 of *Lecture Notes in Computer Science*, pp. 366–382. Springer Berlin, Heidelberg, May 2006.

[8] H. R. Farshchi, M. A. Taylor, and I. A. Macdonald. Decreased thermic effect of food after an irregular compared with a regular meal pattern in healthy lean women. *Int J Obes Relat Metab Disord*, 28(5):653–660, May 2004. doi:10.1038/sj.ijo.0802616.

[9] J. Gao, A. Hauptmann, A. Bharucha, and H. Wactlar. Dining activity analysis using a hidden markov model. In *ICPR 2004: Proceedings of the 17th International Conference on Pattern Recognition*, vol. 2, pp. 915–918, Aug. 2004. doi:10.1109/icpr.2004.1334408.

[10] E. Gualdi-Russo and S. Toselli. Influence of various factors on the measurement of multifrequency bioimpedance. *Homo*, 53(1):1–16, August 2002. doi:10.1078/0018-442x-00035.

[11] H. Hsiao, J. Guan, and M. Weatherly. Accuracy and precision of two in-shoe pressure measurement systems. *Ergonomics*, 45(8):537–555, Jun 2002. doi:10.1080/00140130210136963.

[12] H. Junker, O. Amft, P. Lukowicz, and G. Tröster. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recogn*, 41(6):2010–2024, June 2008. doi:10.1016/j.patcog.2007.11.016.

[13] C. S. Lear, J. B. Flanagan, and C. F. Moorrees. The Frequency Of Deglutition In Man. *Arch Oral Biol*, 10:83–100, 1965.

[14] J. Mankoff, G. Hsieh, H. C. Hung, S. Lee, and E. Nitao. Using low-cost sensing to support nutritional awareness. In G. Goos, J. Hartmanis, and J. van Leeuwen, ed., *Ubicomp 2002: Proceedings of the 4th International Conference on Ubiquitous Computing*, vol. 2498 of *Lecture Notes in Computer Science*, pp. 371–376. Springer Berlin, Heidelberg, September-October 2002.

[15] D. R. Parker, K. Carlisle, F. J. Cowan, R. J. Corrall, and A. E. Read. Postprandial mesenteric blood flow in humans: relationship to endogenous gastrointestinal hormone secretion and energy content of food. *European Journal of Gastroenterology & Hepatology*, 7(5):435–440, May 1995.

[16] D. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. In B. Rhodes and K. Mase, ed., *ISWC 2005: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*, pp. 44–51. IEEE Press, October 2005. doi:10.1109/iswc.2005.22.

[17] D. A. Schoeller. Limitations in the assessment of dietary energy intake by self-report. *Metabolism*, 44(2 Suppl 2):18–22, Feb 1995. doi:10.1016/0026-0495(95)90204-x.

[18] K. A. Siek, K. H. Connelly, Y. Rogers, P. Rohwer, D. Lambert, and J. L. Welch. When do we eat? an evaluation of food items input into an electronic food monitoring application. In E. Aarts, R. Kohno, P. Lukowicz, and J. C. Trainini, ed., *PHC 2006: Proceedings of the 1st International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–10. ICST, IEEE digital library, November 2006. doi:10.1109/pcthealth.2006.361684.

[19] E. Stellar and E. E. Shrager. Chews and swallows and the microstructure of eating. *Am J Clin Nutr*, 42(5 Suppl):973–982, Nov 1985.

[20] M. S. Westerterp-Plantenga, L. Wouters, and F. ten Hoor. Deceleration in cumulative food intake curves, changes in body temperature and diet-induced thermogenesis. *Physiol Behav*, 48(6):831–836, Dec 1990. doi:10.1016/0031-9384(90)90235-v.

[21] R. R. Wing and S. Phelan. Long-term weight loss maintenance. *Am J Clin Nutr*, 82(1 Suppl):222S–225S, Jul 2005.

[22] J. C. Witschi. Short-term dietary recall and recording methods. In W. Willett, ed., *Nutritional Epidemiology*, vol. 4, pp. 52–68. Oxford University Press, 1990.

[23] K. Yamaguchi, T. Yamaguchi, T. Odaka, and H. Saisho. Evaluation of gastrointestinal motility by computerized analysis of abdominal auscultation findings. *J Gastroenterol Hepatol*, 21(3):510–514, Mar 2006. doi:10.1111/j.1440-1746.2005.03997.x.

[24] B. A. Yon, R. K. Johnson, J. Harvey-Berino, and B. C. Gold. The use of a personal digital assistant for dietary self-monitoring does not improve the validity of self-reports of energy intake. *J Am Diet Assoc*, 106(8):1256–1259, Aug 2006. doi:10.1016/j.jada.2006.05.004.

4

Recognition of dietary activity events using on-body sensors

Oliver Amft and Gerhard Tröster

Artificial Intelligence in Medicine, 42(2), 121–136, February 2008.
DOI: 10.1016/j.artmed.2007.11.007

Abstract

Objective: *An imbalanced diet elevates health risks for many chronic diseases including obesity. Dietary monitoring could contribute vital information to lifestyle coaching and diet management, however current monitoring solutions are not feasible for a long-term implementation. Towards Automatic Dietary Monitoring, this work targets the continuous recognition of dietary activities using on-body sensors.*

Methods: *An on-body sensing approach was chosen, based on three core activities during intake: arm movements, chewing and swallowing. In three independent evaluation studies the continuous recognition of activity events was investigated and the precision-*

recall performance analysed. An event recognition procedure was deployed, that addresses multiple challenges of continuous activity recognition, including the dynamic adaptability for variable-length activities and flexible deployment by supporting one to many independent classes. The approach uses a sensitive activity event search followed by a selective refinement of the detection using different information fusion schemes. The method is simple and modular in design and implementation.

Results: *The recognition procedure was successfully adapted to the investigated dietary activities. Four intake gesture categories from arm movements and two food groups from chewing cycle sounds were detected and identified with a recall of 80% to 90% and a precision of 50% to 64%. The detection of individual swallows resulted in 68% recall and 20% precision. Sample-accurate recognition rates were 79% for movements, 86% for chewing and 70% for swallowing.*

Conclusions: *Body movements and chewing sounds can be accurately identified using on-body sensors, demonstrating the feasibility of on-body dietary monitoring. Further investigations are needed to improve the swallowing spotting performance.*

4.1. Introduction

Daily dieting behaviour strongly influences the risk for developing disease conditions. The most prevalent disease associated to an imbalanced diet is obesity. Current estimations account for over one billion of overweight and 400 million obese patients worldwide. This still increasing trend was attributed to the rapid changes in society and behavioural patterns in the last decades [42]. However, obesity is not a unique diet-related disease that decreases healthy life-years in many populations. Rather, it increases the risk for related diseases, including diabetes mellitus, different types of cancer and cardio-vascular diseases. Often the diseases confound or overlay each other, preventing accurate accounting.

Several key risk factors have been identified, that are controlled by dieting behaviour. These include the timing of food intake and integration into daily schedule. For example, intermediate snacking was found to add a major part to the daily energy intake [34]. Another critical aspect is the food selection. High-energy food can be replaced by lower energy densities, such as fruits and vegetables. This improves the diet quality and lowers body weight [31].

Minimising individual risk factors is a preventive approach to systematically fight the origin of diet-related diseases. It is the most promising solution for improving quality of life in the future. Since nutrition is an inherent part of daily activities, the adoption of a healthy diet requires individual lifestyle changes. These changes need to be implemented and maintained over periods of months and years. For this purpose, a convenient long-term monitoring

of dietary behaviour could become a vital tool to assess eating disorders and support diet modifications through feedback and coaching.

4.1.1. Dietary behaviour monitoring

No single-sensor solution exist that could capture the process of food intake and is simple to implement for diet management. Currently, dietary activities are studied manually by entering the information into food intake questionnaires. Mobile devices and Internet appliances are used to support the information entry, *e.g.* by taking pictures of the food [28] and estimating calories from entered data [7]. Further approaches to simplify data entry include the scanning of shopping receipts [27] as well as bar codes or recording voice logs [33].

These manual acquisition methods require a considerable effort of study participants, primarily to remember entering the information into the questionnaire, and study managers, to verify and analyse the data. Typically, this method is prone to errors such as imprecise timing due to back-filling, missing food item details, *e.g.* when using voice recordings [33] and low user compliance, especially for paper-based diaries [36].

Many dietary parameters such as the rate of intake (in grams/sec.) or the number of chews for a food piece are rarely assessed because adequate sensing facilities are only available in laboratory settings. However, these parameters are related to palatability, satiety and speed of eating [41]. Behavioural investigations have utilised weighting tables in controlled settings to measure the amount and rate of food intake during the consumption of individual meals [22]. An oral implant sensor was developed to acquire information about these parameters [35]. However these techniques certainly influence the user's behaviour and are not feasible for long-term monitoring.

All non-invasive dietary monitoring techniques suffer from estimation errors regarding the exact amount and calories of every consumed food item. However, a rough estimation for relevant parameters such as ratio of fluid and solid foods, food category and timing information, such as eating schedule and meal intake durations over the day, will provide a solid basis for behavioural coaching. We believe that much of this information can be extracted from on-body sensors.

4.1.2. Paper contributions and outline

In this work, we evaluate on-body sensing methods to automatically monitor dietary intake behaviour. In particular, three core aspects of dietary activity (*sensing domains*) were investigated by on-body sensors:

1. Characteristic arm and trunk movements associated with the intake of foods, using inertial sensors.

2. Chewing of foods, monitored by recording the food breakdown sound with an ear microphone.

3. Swallowing activity, acquired by a sensor-collar containing surface Electromyography (EMG) electrodes and a stethoscope microphone.

We derive pattern models for specific activity events using the sensor data of each domain and analyse the event recognition performance. For example, individual chews are considered as events in the domain chewing. In particular, the paper makes the following contributions:

1. We present a flexible event spotting method that can be applied either to an individual sensing modality or a combination of several. The approach obtains its adaptivity from a variable-length feature pattern search. Its selective power originates from competitive and supportive fusion of event spottings with largely independent sources of errors. We summarise the domain-specific adaptations of the procedure. The pattern description is achieved by using time and frequency-domain features that model the temporal characteristics of an event. Using this approach, more complex algorithms, like hidden Markov models (HMMs) were avoided.

2. We analyse the recognition of individual arm movements as well as chewing and swallowing activities from the intake of different food items. For each domain, we describe the activity sensing approach, the domain-specific recognition constraints and the conducted case studies to obtain naturalistic evaluation data. Since our work targets a combined detection and classification of the activity events, we present quantitative results for both, indicating a good performance and the feasibility of the sensing approaches for Automatic Dietary Monitoring.

The evaluations are performed on data from three different studies. To analyse the recognition performance under realistic conditions, the data sets included other common activities, *e.g.* conversations and arbitrary movements.

4.2. Dietary activity domains and related work

Activity monitoring and recognition has attracted researchers from many backgrounds, including machine vision and more recently pervasive and wearable computing. An exhaustive review of the literature is beyond the scope of this work. Instead, we focus on systems for behaviour and Automatic Dietary Monitoring as well as research on the three sensing domains considered in this work.

Approaches towards Automatic Dietary Monitoring typically build on intelligent infrastructures. Chang et al. [10] developed a monitoring table to detect activities in a dining scenario. The table is partitioned into several sensing sections equipped with radio-frequency-identification (RFID) readers to identify food containers and weight sensors to track food transport between containers and personal plates. The precision of the system is bound to the spatial resolution of table sensing sections and requires static assignment of food containers

to these sections. The concept of load sensing on a table surface for user activity detection was introduced earlier by Schmidt et al. [32]. In their approach coarse object movements were estimated from a single sensing section.

Beigl et al. [8] equipped household objects with sensing capabilities. In the presented example, a cup was chosen to identify activities carried out with it.

For dietary monitoring applications, RFID technology has great potential as a combined wearable and environmental sensing modality. Patterson et al. [30] attached tags to 60 household objects. The detection was restricted to morning activities, recorded by an RFID reader worn at the user's hand. The activities included, using the bathroom, preparing breakfast foods and eating breakfast.

The infrastructure sensing approaches provide valuable information on various user activities were sensors can be easily attached or hidden. However the approaches generally suffer from the user identification problem: while one user may prepare the foods, several others can consume them. Wearable sensors can bridge this gap and associate the user directly to the activities. Moreover, since worn at the body, the sensors can reveal more detailed information that otherwise would require laboratory setups.

4.2.1. Movement recognition

Movements and gestures related to dietary intake can be roughly discriminated into a preparation phase of the food or beverage items, such as unpacking, opening, cooking and plate or cup filling, and the actual feeding. The feeding movements target the fine-cutting, loading, and manoeuvring of the prepared piece to the mouth. In the feeding phase specific tools, such as fork and knife can be used.

Our focus is to recognise intentional arm and upper body movements during the feeding phase. These movements are a result of handling the tool in the hand(s) and the food material properties viscosity and size. These properties relate directly to the food category. For example a soup is usually feed with a spoon while a glass, cup, or bottle is used for drinking. Hence all relevant movement events can be characterised as directed gestures of the left or right arm, supported by the upper body.

A large base of existing works addressed the problem of classification on well-defined sequences or previously isolated gestures, *e.g.* for Kung Fu moves [9] or in a worker assembly scenario [29]. Works that targeted the continuous recognition used explicit segmentation steps or implicit segmentation capabilities of algorithms, such as HMMs. Lee and Kim [24] used HMMs and introduced a threshold model to eliminate detection noise. The threshold model is constructed from all trained gesture models. Explicit segmentation was used by Ward et al. [39] in an assembly task. Recognition was achieved by fusing classifier outputs. Lee and Yangsheng [23] used acceleration thresholds in combination with HMMs. In previous works of the authors on intake gesture

recognition, HMMs were used together with an explicit data-adaptive segmentation [2].

While HMMs are helpful to model the temporal structure of movements, they were avoided in this work to minimise the complexity of the search procedure for both training and actual search.

4.2.2. Chewing recognition

Chewing targets simultaneous food breakdown and lubrication to form a food bolus that can be swallowed. A chewing sequence starts after the food piece is transferred to the mouth. The food breakdown is composed of arbitrary tongue movements and cyclic opening and closing of the jaw (*chewing cycle*). During the material breakdown sounds are emitted that are partially audible by air-conduction in the near vicinity, but effectively transmitted by bone-conduction from teeth and jawbone to the skull and the ear canal.

The emitted sounds are related to the food material texture. Interaction of chewing with the acoustic sensation and perception of food items has been investigated to study food preferences. Typically, studio recording setups were used to analyse air-conducted chewing sounds [38] and laboratory installations to assess the deformation sounds with a destruction instrument [12]. The loudness of a food item during chewing depends mainly on its inner structure, the arrangement of cells, impurities and existing cracks [1]. Wet cellular materials, such as apples and lettuce, are termed *wet-crisp* since the cell structures contain fluids, whereas *dry-crisp* products, such as potato chips have air inclusions [18].

The food deformation in a chewing cycle is understood as a gradually decomposition of the material structure, observed as a decline of the sound level [17]. Initial attempts were made by DeBelie et al. [14] to discriminate two classes of crispness in apples by analysing principal components in the sound spectrum of the initial bite. In a followup work DeBelie et al. [15] classified the sound emissions from the initial bite of different dry-crisp snacks. Both works addressed the isolated classification. In our previous work the microphone positioning and classification of four different foods was investigated [4]. The ear canal provided the best signal (chewing) to noise (user speaking) ratio. This sensor positioning can be comfortable and socially acceptable for continuous monitoring, comparable to mobile headsets or hearing aids.

In this work, following our recognition approach, the identification of individual chewing cycles from food breaking sounds was targeted. The food category is subsequently classified from the sound pattern of the cycle.

4.2.3. Swallowing recognition

Swallowing is a frequent activity during food intake. It is mostly performed unconsciously and when initiated, controlled by a pattern of muscle activa-

tions [19]. The swallowing act is often partitioned into (1) oral preparation phase (food in the mouth), (2) pharyngeal phase (food bolus in the throat) and (3) oesophageal phase (food propulsion towards the stomach) [16]. After transforming the food to a sallowable bolus in the oral phase, the swallowing reflex is initiated by the tongue, starting the pharyngeal phase. In this phase a sequence of muscle activations is used to transport the bolus and protect the respiratory tract.

A number of clinical assessment methods have been developed to analyse the complex interaction of swallowing, phonation and respiration at the pharynx and diagnose abnormal swallowing in the pharyngeal phase. The assessment methods can be broadly grouped as invasive methods, that require a strict laboratory or clinic setting and a variety of non-invasive sensing methods. In the latter category, the following main approaches were taken: sensing muscle activations by surface EMG, *e.g.* [20], listening to the throat sounds using a stethoscope [26] as well as stethoscope-like acoustic transducers or sealed microphones [11].

A large share of research works targeted the basic understanding of the swallowing process, only few addressed the continuous monitoring. Danbolt et al. [13] used sensors to detect hyoid movement at the throat. It was found that the sensor incurs heavy measurement artifacts from neck and tongue movements as well as from speaking. Limdi et al. [25] tracked muscle contraction intensity based on surface EMG to inform the user of elevated swallowing rates. Sukthankar and Reddy et al. [37] used surface EMG and vibration sensors and targeted applications in dysphagia rehabilitation. Both latter works did not present a performance evaluation for their approaches to the continuous recognition problem. In our previous work [5], swallowing was analysed from surface EMG and sound for the isolated classification of swallowed bolus types, *e.g.* solid or fluid. Moreover, an initial investigation towards the continuous detection was made. The approach is taken forward in the present evaluation by extending the swallowing study and evaluating the performance of different fusion methods.

4.3. Recognition and evaluation methods

The envisioned system shall be continuously worn during daily routine. In all sensing domains relevant activity events occur only sporadically, often embedded into a large set of other, non-relevant activities (*NULL class*). For example, stethoscope-like sound recordings intended to record swallowing sounds at the throat, inherently pick up speaking, or even environmental noises.

A method that targets the spotting of relevant activity events should be effective in retrieving correct events while omitting NULL class data. However, the sensing domains considered in this work have very few constraints, resulting in a highly variable NULL class. As a consequence of this diversity, it is

not feasible to derive a model for NULL (garbage model) without integrating assumptions about these random activities. Moreover, training of the relevant event model(s) should be critically reviewed for its dependency on NULL.

Another challenge is the variable length of the activities, leading to duration variances in the relevant events. Consider for example a intake gesture using fork and knife where the food must be cut into appropriate sized pieces before manoeuvring it to the mouth. This indicates that a simple, fixed sliding window search would not be able to identify the gestures accurately.

Our approach to detecting and classifying dietary activities is based on three main steps: (1) an explicit segmentation of signals to define search bounds, (2) a sensitive event detection using a feature similarity search algorithm with an adaptive, dynamically defined window size, and (3) a selective fusion of detection results exploiting independent sources of error to filter out false positives and obtain an event classification in the same step. Figure 4.1 outlines the components of our event detection and classification method.

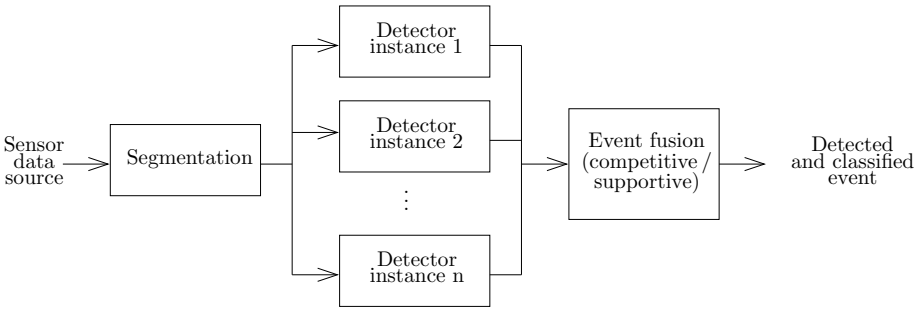


Figure 4.1. Event detection and classification procedure used in the work. The detector instances (1 to n) can be trained to spot activity event patterns of specific classes or individual modalities. The event fusion can combine events of different type (competitive) or modalities for one type (supportive). Both concepts are presented in this work.

4.3.1. Event recognition procedure

In the first step a segmentation is obtained that specifies the bounds for the following search. Various data-adaptive methods or a fixed distance can be used for this purpose. In this work, we used the latter approach with a domain-specific distance setting.

Event detection using feature similarity search

The event detection step utilises the segmentation points to search for potential activity event sections using a similarity-based algorithm. The search is

performed by comparing features of a data section under investigation to a previously trained pattern.

The following search principle is illustrated in Figure 4.2. For a given segmentation point, the history of sensor data is analysed between a lower and upper search bound. These bounds are determined in the training step from the overlapping of manually annotated events and the segmentation points. For each search section the similarity of a feature set to a pre-trained set is quantified by computing the Euclidean distance (D_{Event}) between them. A distance threshold (D_{Thres}), also obtained during the training, is used to remove unlikely sections. The similarity search works as a detector that returns a list of event sections associated with a distance to the training pattern.

One benefit of this algorithm is that it can operate as a single pattern detector, when applied to retrieve one relevant type from continuous sensor data only. Using the feature similarity search, multiple detector instances can be combined to independently spot different classes. This permits an independent feature set for each class. Furthermore, as we will show for the detection of swallowing, instances trained from independent sensing modalities can be used to detect one event type in parallel.

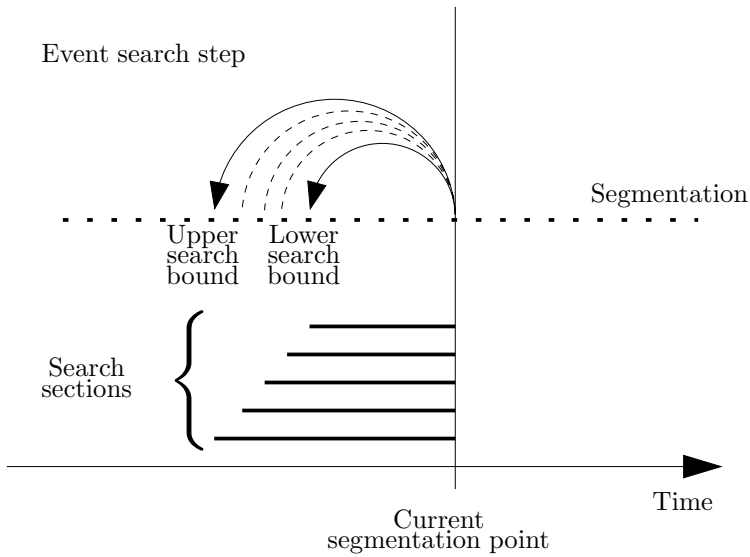


Figure 4.2. Schematic of the activity event search step. The segmentation is indicated by the dotted line. The search is performed by computing feature sets from the sensor data (not shown) between lower and upper search bounds. The search sections are evaluated by comparing their feature sets to a pre-trained pattern. (Please refer to the text for more details.)

Competitive and supportive event fusion

By selecting an appropriate distance threshold (D_{Thres}), the similarity search is configured to spot most of the activities in the sensor data. Consequently it can incur false positives. In the fusion step different class- or modality-specific event detectors are combined to reduce these errors. This improvement originates from the independent sources of error of each detector and modality.

For multiple detectors a competitive fusion strategy was used to select the final events. A supportive strategy was deployed to combine the modality-specific detection of one activity type, since here the detectors could reinforce each other.

In this work we evaluated different fusion methods: (1) comparison of the events, keeping the event with the highest confidence (COMP), (2) agreement of the detectors (AGREE) and (3) re-weighting of the detection by logistic regression (LR). The methods are commonly used to combine classifier outputs [21, 39]. In this work, COMP corresponds to the competition strategy and AGREE implements a supportive approach. LR can be used for both strategies.

To select the most probable from concurrently reported events, the competitive fusion compares a confidence associated to each event. This confidence was derived from the similarity search distances (D_{Event}) by normalisation using the distance threshold (D_{Thres}) in each detector instance (Equation 4.1).

$$Confidence = \frac{D_{Thres} - D_{Event}}{D_{Thres}} \tag{4.1}$$

A sliding buffer of candidate events is used and continuously updated as new events are entering from the detector instances. For each entering event the collision (temporal overlapping of the event section with events already in the buffer) is resolved according to the selected fusion strategy. The events are released from the buffer after a timeout as final result of the procedure.

4.3.2. Feature computation

The temporal structure of many complex activities is a key element for their pattern modelling and subsequent machine recognition. For example, movements are frequently modelled with HMMs and time-continuous features to capture this effect.

In this work, we integrated the temporal structure of the activity events in individual single-value features. The features were computed for predefined sections of an event. We spitted the event in two or four slices. This solution provided an acceptable trade-off between temporal description and total number of features. The solution permits a combination of sliced features and features for the entire event. Moreover, this approach can simplify both modelling and event search, compared to time-continuous features. We used it with the recognition approach presented above. The similarity search is then performed using the features to describe each event and search every section.

4.3.3. Evaluation procedure

Experimental concept

The analysis of each sensing domain was based on experimental data, individually acquired for each domain. Figure 4.3 indicates the sensor attachment at the body for all domains. For the recording of movements a commercial motion acquisition system based on inertial sensors was used. Customised systems were utilised for the chewing (ear microphone) and swallowing (sensor collar) recordings. Table 4.1 provides a detailed description of the sensors used. In each study the activities were manually annotated by an observer. The study procedures are further detailed in the evaluation sections for each sensing domain.

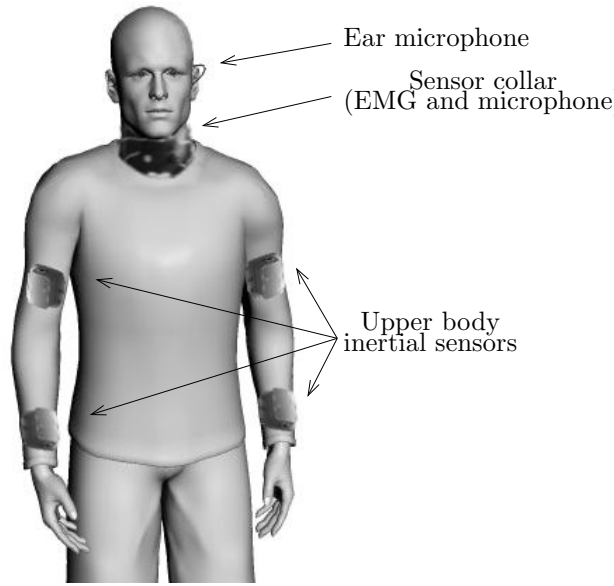


Figure 4.3. Schematic sensor positioning at the body. (See Table 4.1 for a detailed description.)

Soft alignment procedure

In order to account an event as recognised, the detection procedure must return a valid begin and end of an activity section and its identity (for multi-class detections). The section boundaries were compared to begin and end of the annotated events. However the boundaries do not match exactly since the

Table 4.1. List of sensors systems used in the dietary activity studies.

Sensor type	Sensor description	Sensing domain
Inertial sensors	Sensor modules containing acceleration sensors, gyroscopes (rate of turn) and compass sensors (magnetic field), each in 3 dimensions. The modules were attached to the user's arms. Manufacturer: XSens, model: MTi.	Movement activity
Ear microphone	Electret miniature condenser microphone. The microphone was embedded into an ear pad foam and worn at the ear canal. Manufacturer: Knowles Acoustics, model: TM-24546.	Chewing activity
Stethoscope microphone	Electret condenser microphone. The microphone was attached with medical tape or worn in a collar below the hyoid. Manufacturer: Sony, model: ECM-C115.	Swallowing activity
Electromyogram (EMG)	Electromyogram electrodes and acquisition system. Electrodes were directly attached or worn in a collar at the infra-hyoid throat position. Manufacturer: MindMedia, model: Nexus-10.	

manual annotation was not accurate on the granularity of each sample and the segmentation algorithm can introduce a small alignment error in the detection.

For the feasibility in the envisioned dietary monitoring application the exact alignment is not a critical aspect, if the event is associated to the true activity at all. Hence, we applied a soft alignment matching, following the concept of a boundary jitter. Equation 4.2 describes the accounting of correct events.

$$Recognised = \begin{cases} \text{true}, & \text{if } j \leq max\left(\frac{|A_{Begin}-E_{Begin}|}{A_{End}-A_{Begin}}, \frac{|A_{End}-E_{End}|}{A_{End}-A_{Begin}}\right) \\ \text{false}, & \text{otherwise} \end{cases} \tag{4.2}$$

The parameters A_{Begin} and A_{End} correspond to start and stop sample of the manual annotation and likewise, E_{Begin} and E_{End} to the retrieved event. The jitter parameter j can be set, depending on the acceptable jitter for an application. The jitter $j = 0$ corresponds to an exact matching of the boundaries and $j = 1$ would allow a jitter in size of the event duration. Moreover,

this accounting procedure assures that large events, covering more than the annotation section, will be rejected as well, if their begin and end do not conform to Equation 4.2. Multiple counts of matches and misses were especially avoided.

For the evaluation in this work a jitter of $j = 0.5$ was chosen. We believe that this is an adequate accuracy for applications in dietary monitoring.

Performance measurement

To account for variations in the acquired data sets, a four-fold cross-validation procedure was used to determine training and testing set for the performance analysis. For training, three of four data parts were used. Evaluation was performed on the left-out data part. This procedure was repeated until all four parts were used for testing once. The partition boundaries were adapted to avoid intersecting the manually annotated event sections. The choice of four partitions reflects an empirical trade-off between processing effort, the need for enough training observations in all combinations of the partitions and the intended averaging effect for the final results. An additional performance gain could be achieved by higher iteration counts, potentially using more events for training.

To analyse the recognition performance, we used the metrics *Precision* and *Recall*, commonly used for information retrieval assessments. These metrics are derived as follows:

$$Recall = \frac{\text{Recognised events}}{\text{Relevant events}}, \qquad Precision = \frac{\text{Recognised events}}{\text{Retrieved events}} \tag{4.3}$$

Relevant events corresponds to the manually annotated number of actually occurred event instances. *Retrieved events* represents the number of events returned by the event recognition procedure. Finally, *Recognised events* refers to the correctly returned number of events. Both metrics have a value range of $[0, 1]$. A recall value of one indicates a perfect accuracy of a method (all relevant events are recognised), while a precision value of one indicates that the method does not return false positives (insertion errors).

4.4. Movement recognition

4.4.1. Study description

To evaluate our recognition approach for movements, a case series was recorded, utilising commercially available inertial sensors. Table 4.1 specifies the sensors used. The inertial sensors were attached onto a jacket at the lower and upper arm as well as the upper back. Figure 4.3 illustrates the sensor positions.

The movements of the arms and upper body was recorded with a sampling rate of 100 Hz from four right-handed volunteers (1 female, 3 male, aged between 25 to 35 years). The participants were seated in front of a table carrying the food items and tools. They were instructed to eat and drink as they would normally do.

Intake sessions were recorded from each participant on separate days. Four intake activities were recorded for each session: (1) eating meat lasagne with fork and knife (cutlery, CL), (2) fetching a glass and drinking from it (DK), (3) eating a soup with a spoon (SP), and (4) eating slices of bread with one hand only (HD). All meals were served at adequate temperature for normal eating/drinking. Table 4.2 summarises the acquired data which was inspected and annotated.

In order to enrich diversity of the data set and avoid long periods without movements, the participants were asked to conduct a set of other, non-relevant movements and gestures. Besides arbitrary movements of the participants the following additional arm gestures have been recorded and annotated to quantify the data set noise: scratching head (96 times), touching chin (92 times), reading and turning pages of newspaper (99 times), using tissue (89 times), glancing at the watch (92 times) and answering a simulated mobile phone call (90 times), all total numbers of the data set.

Table 4.2. Movement study: Statistics of acquired and annotated intake gestures.

Number of participants	4
Annotated gestures	1020
Relevant event share	97.44 min (34.7%)
Total length of data set	4.68 hours

4.4.2. Evaluation results

The event recognition procedure was adapted to the movement domain in the following way:

1. A time constant of 0.5 s was used for segmentation.
2. For each of the four gesture categories an event detector instance was trained. Using the Euler angles of the lower arms, features such as mean, variance and signal sum in four sliced sections and for the complete gesture were computed. By visually inspecting test recordings we found that the upper arm and the back sensors could not support the recognition without constructing a more complex body model. Hence, they were excluded from the analysis.
3. The event fusion using the competitive strategy was subsequently applied to the detector instance results and the event category with the highest

confidence was selected as final result. Due to variable lengths of gestures in our data set, the candidate buffer was configured to release events only after 30 s.

Figure 4.4 shows precision-recall (PR) graphs for a user-specific evaluation of the movement event fusion using the COMP method. The curves were created by evaluating the performance at various confidence thresholds for every class and for every participant (A-D). Best performance is found towards the top-right corner (high precision, high recall).

Both graphs indicate a good performance for the movement event recognition. The best result was achieved for the category DK, while HD performed less well. Since the latter gesture is very simple it was often confused with other movements towards the head. In contrast, DK is more complex (fetching, drinking). The second graph shows that all participants performed similarly well.

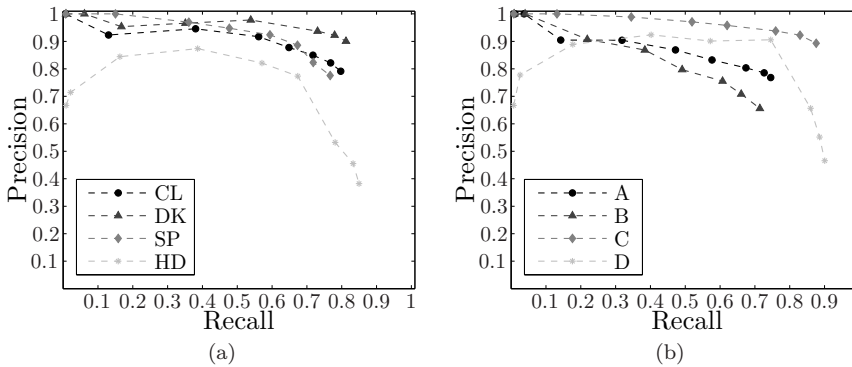


Figure 4.4. Movement study: User-specific PR analysis (confidence threshold sweep) of the event fusion results using the COMP method. Best performance is found towards the top-right corner (high precision, high recall). (a) Analysis for every category (CL=cutlery, DK=drink, SP=spoon, HD=hand only). (b) Analysis for every study participant (A-D).

Table 4.3 summarises the results obtained from the event detection and the event fusion. For the SP gestures, we observed that participants bend themselves over the bowl, to avoid spilling and to minimise the movements. This affected the detection performance, since only lower arm features were used in the evaluation.

Table 4.4 shows a confusion matrix of the event recognition, obtained by comparing the recognition results to the annotation for each sensor data sample. Complementary to the soft alignment counting scheme used for the results

in Table 4.3, this representation shows the sample-accurate result. For all categories and NULL a recognition rate of 75% to 82% was achieved. This rate was computed as class-relative accuracy ($\frac{correct_C}{relevant_C}$).

Table 4.3. Movement study: Summary for the user-specific performance for the event detection and the fusion method COMP.

Metric	**Event detection**				**Event fusion (COMP)**				
	CL	DK	SP	HD	CL	DK	SP	HD	**Total**
relevant	276	245	266	233	276	245	266	233	**1020**
retrieved	347	247	284	717	278	221	263	518	**1280**
recognised	223	210	208	201	220	199	204	198	**821**
deletions	53	35	58	32	56	46	62	35	**199**
insertions	124	37	76	516	58	22	59	320	**459**
recall	0.81	0.86	0.78	0.86	0.80	0.81	0.77	0.85	**0.80**
precision	0.64	0.85	0.73	0.28	0.79	0.90	0.78	0.38	**0.64**

Table 4.4. Movement study: Confusion matrix of the final user-specific evaluation result using COMP fusion (duration in seconds and ratios).

		Predicted category				
		NULL	CL	DK	SP	HD
Actual category	NULL	8869 (81%)	613 (6%)	233 (2%)	305 (3%)	982 (9%)
	CL	452 (17%)	2130 (82%)	0 (0%)	0 (0%)	8 (0%)
	DK	302 (20%)	1 (0%)	1182 (78%)	0 (0%)	34 (2%)
	SP	237 (22%)	19 (2%)	0 (0%)	807 (75%)	10 (1%)
	HD	103 (16%)	20 (3%)	0 (0%)	0 (0%)	541 (81%)

4.5. Chewing recognition

4.5.1. Study description

For the evaluation of chewing sounds we used an ear microphone as indicated in Figure 4.3. The miniature microphone was build into a standard type ear

pad and kept at the ear canal by an ear hook, as it is used for mobile phone headsets. In a single case study the chewing sounds from different foods were recorded at 16 bit, 44 kHz from a male individual with natural dentition (aged 29 years).

The participant was seated conveniently on a chair close to a table carrying the foods. He could still hear normal-level conversation in the room and was allowed to move and speak during the recording sessions. The room was controlled for a constant noise level of an office environment (the recording in a sound studio was avoided). Recordings were made in individual sessions on separate days. The participant took bites from the foods as he wished. All of the foods belonged to his normal diet. The food products included for the recognition analysis were:

1. Dry-crisp food: potato chips, approx. 3 cm in diameter
2. Wet-crisp foods: (1) mixed lettuce, containing endive, sugar loaf, frisée, raddichio, chicory, arugula, and (2) raw carrots.
3. Soft foods: (1) cooked chicken meat and (2) pasta.

The foods evaluated in this work, contained many chewing cycles. Manual annotation of every chewing cycle was performed in a post-recording step by reviewing the waveforms and listening to the sounds. This procedure is accurate in identifying every chewing cycle until the food bolus is swallowed, however it makes the recordings very expensive.

The recordings included chewing sounds from further food products (bread and chocolate), as well as environmental conversation and speaking. Table 4.2 summarises the acquired data which was inspected and annotated.

Table 4.5. Chewing study: Statistics of acquired and annotated chewing sounds.

Number of participants	1
Annotated chewing cycles	1947
Relevant event share	10.50 min (21.7%)
Total length of data set	0.81 hours

4.5.2. Evaluation results

The event recognition procedure was adapted to the chewing domain in the following way:

1. A time constant of 125 ms was used for segmentation. This choice was made based on the average duration of a chewing sound (as annotated) of 350 ms or less, depending on the food type.

2. Initially, for each of the three food categories a feature similarity instance was trained. Using the microphone data, spectral features such as band energy, auto-correlation and cepstral coefficients in four sliced sections were computed. We observed during the evaluation, that the detector for soft foods worked poorly, resulting in many insertion errors. This behaviour was attributed to the low signal to noise ratio. We omitted this model in the further evaluation to demonstrate the good performance of the dry and wet food detectors.

3. The event fusion using the competitive strategy was subsequently applied to the detected chewing cycles and the category with the highest confidence was selected as final result. We analysed the COMP and LR methods for the fusion.

The low-amplitude chewing sounds from the soft foods (meat and pasta) created a special problem for the detector. While a high recall was achieved, the detection was very sensitive to other sounds (as seen in the low precision in Table 4.6). COMP and LR fusion of the three detectors did not solve this problem, because the number of soft-food insertions was too high.

For every intake cycle all chews were annotated until the food bolus was swallowed and the normal mouth cleaning phase began. In this phase, chews were hard to observe in the sound waveform. However the algorithm was still able to detect them. Figure 4.5 visualises an example waveform including a chewing sequence of potato chips, the cleanup and a conversation phase. For this food the chewing cycles can be seen very well in the sound waveform. The vertical bars indicate the annotation. In the lower plot, the detected chewing events are shown as horizontal bars. As the diagrams shows, additional events were reported for the cleanup phase. We exemplarily verified that these chews were correctly retrieved.

Since the actually existing chews in the cleanup phase could not be automatically verified, they were counted as insertion errors. The impact can be seen in the PR performance analysis in Figure 4.6 and the summary in Table 4.6. For both food categories the COMP and LR fusion methods return good results. We concluded from the quantitative summary in Table 4.6 that LR removes slightly more insertion errors and has less deletions.

Table 4.7 shows the confusion matrix derived by applying the LR method. Using the same procedure as presented for the movement confusion analysis, class-relative recognition rates of 85% to 87% were achieved. This indicates a very good performance. Especially, a low confusion rate of the dry and wet categories was observed.

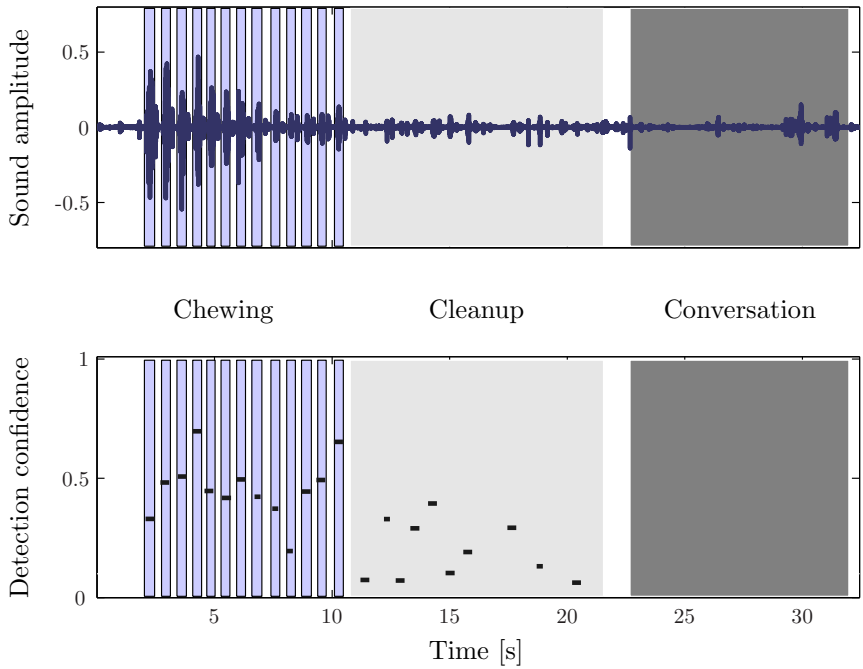


Figure 4.5. Chewing study: Example waveform of a chewing sequence of potato chips, cleanup and conversation phases, indicated by the shaded areas. Upper plot: sound waveform. Lower plot: chewing cycle detection result. (The detector correctly identified chewing cycles in the cleanup phase, that were not annotated. Please see the related text for more details.)

4.6. Swallowing recognition

4.6.1. Study description

Swallowing was analysed from surface EMG electrodes and a microphone sensor. The sensor positioning was equal for all participants. For some participants the sensors were embedded in a collar. The collar helped to quickly attach the sensors to the correct throat region. The location of the EMG was constantly verified, however the collar supported the stable positioning at the infra-hyoid position very well. The microphone was situated at the lower part of the throat, below the larynx. EMG was recorded at 24 bit, 2 kHz and bandpass filtered. Sound data was recorded at 16 bit, 22 kHz. Figure 4.3 and Table 4.1 summarise positioning and setup of the sensors and the collar.

Six volunteers (4 male, 2 female, aged 20 to 30 years) without known swallowing abnormalities were instructed to eat and drink different food items:

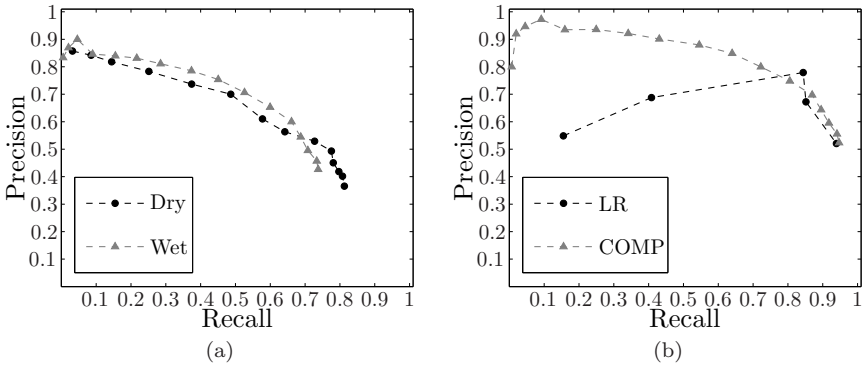


Figure 4.6. Chewing study: User-specific PR analysis (confidence threshold sweep) of the event fusion stage. Best performance is found towards the top-right corner (high precision, high recall). (a) Analysis for the two food categories ("dry" and "wet"). (b) Analysis for the two competitive fusion methods (COMP and LR).

5 and 15 ml of water, spoonfuls of yoghurt and pieces of bread (approx. 2 cm^3). The individuals were seated conveniently on a chair in front of a table carrying the foods. They were allowed to move, chew and speak normally during the recording sessions. The room was controlled for a normal and constant noise level of an office environment. To account for physiologic variations, two intake sessions were recorded on different days. The participants were asked to swallow the food items in one piece after chewing and manipulating the bolus as usual. None of the participants expressed a dislike for any of the included foods nor problems to swallow the selected bolus sizes. Table 4.8 summarises the acquired data that was inspected and annotated.

4.6.2. Evaluation results

The event recognition procedure was adapted to the chewing domain in the following way:

1. A time constant of 250 ms was used for segmentation.
2. Feature similarity instances were trained using the EMG and microphone data individually. The foods were initially grouped regarding their expected bolus size into small (5 ml water, spoonfuls of yoghurt and pieces of bread) and large (15 ml water). This approach was dropped, since no clear discrimination of the two categories was found. In the following, we targeted the detection without further classification. We concluded from early tests that the EMG is disturbed by different muscle activations, independent from swallowing. The investigated hyoid muscle is covered by

Table 4.6. Chewing study: Summary for the user-specific performance for the event recognition (three categories) and the fusion methods (COMP and LR). The fusion results were derived using the food categories "Dry" and "Wet" only.

Metric	Event detection			Event fusion					
				COMP			LR		
	Dry	Wet	Soft	Dry	Wet	**Total**	Dry	Wet	**Total**
relevant	187	979	781	187	979	**1166**	187	979	**1166**
retrieved	1327	2098	3483	416	1693	**2109**	416	1687	**2103**
recognised	186	909	460	152	722	**874**	184	900	**1084**
deletions	1	70	321	35	257	**292**	3	79	**82**
insertions	1141	1189	3023	264	971	**1235**	232	787	**1019**
recall	0.99	0.93	0.59	0.81	0.74	**0.75**	0.98	0.92	**0.93**
precision	0.14	0.43	0.13	0.37	0.43	**0.41**	0.44	0.53	**0.52**

Table 4.7. Chewing study: Confusion matrix of the final user-specific evaluation result using LR fusion (duration in seconds and ratios).

		Predicted category		
		NULL	Dry	Wet
Actual category	NULL	2791 (86%)	100 (3%)	344 (11%)
	Dry	12 (13%)	76 (87%)	0 (0%)
	Wet	57 (15%)	3 (1%)	332 (85%)

several layers of other muscle tissue. We concentrated on a simple activity detection using time domain features such as sum, maximum and peaks of the signal. For the sound data, spectral features such as band energy, auto-correlation coefficients and signal energy were used. An initial test of sliced features did not lead to an improvement in recognition.

3. The event fusion using a supportive strategy was subsequently applied to the detected swallowing events from EMG and sound data. We analysed the performance of AGREE and LR methods.

For the AGREE fusion all participants reached a high recall, indicating that the detection procedure was able to retrieve many events. Figure 4.7 presents the corresponding PR analysis. The evaluation revealed two groups: for participants (C and D) the detection performance was higher than for the others. However, these participants did neither belong to the same gender, nor were they recorded with the collar. We observed that many other participants exhib-

Table 4.8. Swallowing study: Statistics of acquired and annotated swallowing activity.

Number of participants	6
Annotated swallows	1265
Relevant event share	44.58 min (9.3%)
Total length of data set	7.93 hours

ited either a high EMG response or sound, for C and D both sensors provided a consistent event pattern. Consequently, both EMG and sound-based detection more often returned a correct result for them, whereas for the remaining participants no reduction of the insertion errors was achieved. Further investigation of this issue is required.

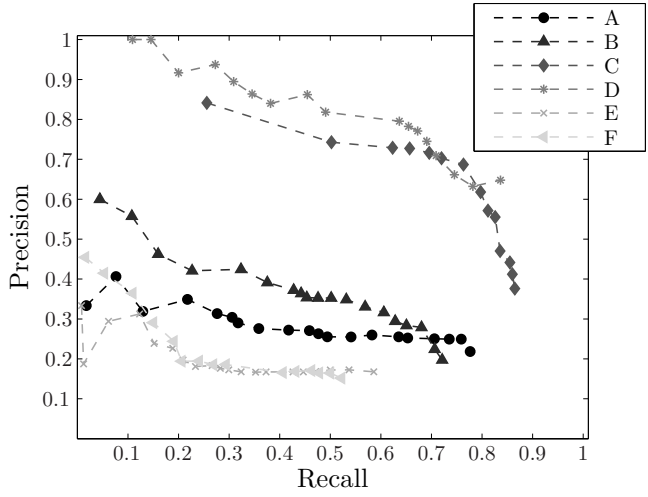


Figure 4.7. Swallowing study: PR analysis (confidence threshold sweep) for each study participant (A-F) using the agreement fusion (AGREE). Best performance is found towards the top-right corner (high precision, high recall).

On average for all participants, the AGREE fusion method improved the precision. LR did not improve the individual spotting results. Table 4.9 summarises the results obtained from the event detection instances and the fusion methods.

The sample-accurate detection result was determined from the AGREE fusion result. The swallowing recognition rate was 64%, for the NULL class 75% were obtained. This indicates that the detection provides a sensible result.

Table 4.9. Swallowing study: Summary for the user-specific performance for the event detection using muscle activity (EMG), audio (SND), and the fusion methods (LR and AGREE).

Metric	Event detection		Event fusion	
			LR	**AGREE**
	EMG	SND	EMG+SND	EMG+SND
relevant	1265	1265	1265	1265
retrieved	6046	8093	8085	4345
recognised	955	834	824	861
deletions	310	431	441	404
insertions	5091	7259	7261	3484
recall	0.75	0.66	0.65	0.68
precision	0.16	0.10	0.10	0.20

4.7. Discussion

4.7.1. Methodology

The continuous recognition of dietary activity events from sensor data patterns was evaluated in this work. Spotting activity events in continuous sensor data is a vital prerequisite for the deployment of activity detection in general. While the targeted activities can be described by a domain expert, the embedding data (NULL class) cannot be modelled due to the degrees of freedom in the human activities and the cost for large training data sets. Consequently, assumptions about the embedding should be minimised to achieve an acceptable performance generalisation. We believe that the current work is a step towards resolving this challenge, although the presented method is not completely free from assumptions. The most critical aspects in this respect include the selection of features and event detection thresholds.

A combination of individual single-value features for activity event slices were used for the detection. With this approach the temporal structure of the activities was transformed into a spatial representation. This is a useful concept to model activities for the continuous search. In an earlier work, we applied this principle to the recognition of gaming gestures only [6]. For each domain, features were selected from visual inspection of the sensor waveforms and from previous experience. We expect that the recognition performance could be improved by a thorough feature search and selection strategy. This will also help to identify sensors that can be omitted or adjusted in its placement.

We introduced the scheme of competitive and supportive event fusion to construct a selective refinement step for spotted events. By design of the recognition system, the choice of the fusion strategy is made. The supportive strat-

egy was applied for spottings from independent sensors, describing the *same event type*. Using competitive fusion, we selected the most appropriate event from *different event type* spottings. Both strategies could be combined to more complex selection schemes. In related works, they have been used to combine classifier outputs mostly [39].

An advantage of our method is its ability to work on single event detection classes with individual feature sets. For the detection of one event type, typically a supportive fusion strategy can still be used, by deploying different sensors. An application for detecting single event types in dietary monitoring was shown in the swallowing evaluation. Further applications are the detection of drinking gestures to assess fluid consumption or using a single food model to assess one category of foods in dietary intake.

In order to describe the complexity of the event detection as a search problem, we listed the embedding size of the data sets. This size was expressed as ratio of total annotated event duration over the total length of the data set. For the data sets in this work, the ratio was 34.7% for the movement, 21.7% for chewing and 9.3% for the swallowing study. The ratio indicates the severity of the search: the smaller the ratio, the more difficult it is to achieve a good recognition results due to the large and potentially diverse embedding data. However, we believe that the high embedding size in the swallowing study is not the unique reason for its weak precision. Section 4.7.4 discusses the swallowing study in detail.

We introduced a soft alignment measure to account for the variability in alignment between annotation and event detection. A boundary jitter normalised by the annotated length of the event was defined as threshold, below which the event is counted as recognised. The larger the jitter, the more mismatch in alignment is allowed and an event reporting that may otherwise be accounted as insertion/deletion will be accepted as correct. In its extreme, the counting of correct events could be made by simply checking if an overlap with the annotation exist at all. For the targeted applications in dietary monitoring an exact match is less critical as long as the activity is captured at all. Therefore, we selected a jitter value that is neither too optimistic (by permitting large alignment errors) nor pessimistic (being overly strict in the boundary match). The comparison with sample-accurate confusion matrices confirms that the soft alignment is a sensible solution for event spotting performance analyses. For a more detailed analysis of detection errors, the Error Distribution Diagrams [40] could be used.

4.7.2. Movement recognition

Different gesture types were defined, that occur frequently in European and American diets, to evaluate the recognition of food intake movements. The results indicate that all types could be recognised from lower arm motion, most of them with good accuracy. To improve the recognition of certain gestures,

information from inertial sensors at the subject's back could be added. The proposed event fusion method is a valuable addition to the feature similarity search for movement detection. In a related work of the authors, a two-stage approach based on a similarity search and HMMs was used [2]. While the HMMs proved valuable for refining the detection result in the second stage, they add a high complexity in both, initial design and parameter estimation. In comparison, the performance achieved with the event fusion approach in the current work could match the recall, but performs approx. 10% lower in precision than the HMMs on the same data set. Further refinement of features and segmentation could close this gap. Moreover, we presented a rigorous evaluation framework using cross-validation in this work, that was not previously available.

4.7.3. Chewing recognition

For the recognition of chewing sounds, novel achievements on a chew-accurate detection were presented. Using the recognition procedure, individual chewing cycles were identified in two food categories with good performance. This result was achieved by considering the chew as a non-stationary event and grouping the foods with similar textures. In comparison to our earlier investigation ([4]), the current recognition rates are approx. 15% higher and a majority vote over multiple chewing cycles could be avoided. However, for low-amplitude chewing sounds, found in soft foods such as cooked pasta or meat, a low detection performance persists with the current approach. This effect was attributed to the low signal to noise ratio of these sounds. Moreover, the chewing sequence is not consistent over the entire intake cycle as assumed in the current approach [3]. This is observed as a variability in the detection confidences and hinders fusion methods such as LR to achieve a higher performance. Consequently, food models should include the sequence information more carefully.

4.7.4. Swallowing recognition

The automatic detection of swallowing using EMG and sound information was evaluated. We found that swallows can be retrieved from continuous data at high recall rates using both sensing sources. By observing the final detection, we found that the method is disturbed by neck movements and coughing. In comparison to our previous work ([5]), we presented results from additional fusion methods (AGREE, LR) and an extended study. The AGREE fusion was able to remove a large share of insertion errors. The current results confirm the previous findings: while the detection works to some extend in controlled environments, it retrieved many false positives in our evaluation. These errors could not be completely removed by the currently applied fusion techniques.

The collar worked well to standardise and maintain the sensor positioning. No differences in the spotting results were observed for the collar-based swal-

lowing data. For a subgroup of two participants an improved performance was achieved. The difference could not be explained by the available information. A larger study with more participants could reveal whether the subgroups persist. Further investigations are required to analyse options for food bolus categorisation and to increase the algorithm precision.

4.8. Conclusion

We presented novel approaches to monitor dietary activities from body-worn sensors. Three sensing domains were analysed, that are directly linked to the sequence of dietary activities: intake movements, chewing, and swallowing. We presented evaluation results from studies in each domain using an event recognition procedure, that supports the detection and identification of specific activities in continuous sensor data.

The recognition of natural movements, such as for dietary intake, is a challenging task, since it is strongly related to personal habits. The detection procedure in combination with the simple comparison fusion yielded good recognition results for different intake types. This is a valuable result for the intended application, since the intake movements help to categorise the consumed foods. Moreover, the movement recognition could be used independently. For example, the detection of drinking movements can be used to monitor fluid consumption and avoid dehydration.

Chewing is a very important part in the intake process. In this work a successful continuous recognition of two food types was achieved. This is a vital result for a detailed analysis of food chewing. Based on the presented approach, additional models can be derived that reflect the mechanical properties of foods. Besides the identification of consumed foods, the chewing recognition permits the assessment of dietary parameters, such as chews per food and chewing speed. Both parameters can be used as indications for too fast, or stress eating.

Swallowing concludes the intake cycle. The swallowing frequency depends on the food category, where foods containing fluid compartments require elevated swallowing rates. The current detection method, using sound and muscle activity at the throat, still incurs many insertion errors. However, it does provides an indication for swallowing events. We plan to use this information in combination with the previous sensing domains. Further works will address different fusion strategies and additional sensors.

The three domains provide a comprehensive picture of dietary activities and a broad amount of information, that is vital for a long-term dietary coaching and health management. This includes the food type as well as intake timing and the overall meal schedule.

We have shown in this work, how our recognition procedure to spot sporadic activity events can be slightly adapted to fulfil the requirements of very differ-

ent sensor modalities and activities. We believe that the procedure is a helpful tool for Automatic Dietary Monitoring and similar applications in continuous activity recognition.

Bibliography

[1] W. AlChakra, K. Allaf, and A. Jemai. Characterization of brittle food products: Application of the acoustical emission method. *J Tex Stud*, 27(3):327–348, July 1996. doi:10.1111/j.1745-4603.1996.tb00078.x.

[2] O. Amft, H. Junker, and G. Tröster. Detection of eating and drinking arm gestures using inertial body-worn sensors. In B. Rhodes and K. Mase, ed., *ISWC 2005: IEEE Proceedings of the Ninth International Symposium on Wearable Computers.*, pp. 160–163. IEEE Press, October 2005. doi:10.1109/iswc.2005.17.

[3] O. Amft, M. Kusserow, and G. Tröster. Automatic identification of temporal sequences in chewing sounds. In T. Hu, I. Mandoiu, and Z. Obradovic, ed., *BIBM 2007: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pp. 194–201, San Jose, CA, USA, November 2007. IEEE Press. doi:10.1109/bibm.2007.18.

[4] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of chewing sounds for dietary monitoring. In M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, ed., *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing*, vol. 3660 of *Lecture Notes in Computer Science*, pp. 56–72. Springer Berlin, Heidelberg, September 2005. doi:10.1007/11551201_4.

[5] O. Amft and G. Tröster. Methods for detection and classification of normal swallowing from muscle activation and sound. In E. Aarts, R. Kohno, P. Lukowicz, and J. C. Trainini, ed., *PHC 2006: Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–10. ICST, IEEE digital library, November 2006. doi:10.1109/pcthealth.2006.361624.

[6] D. Bannach, O. Amft, K. S. Kunze, E. A. Heinz, G. Tröster, and P. Lukowicz. Waving real hand gestures recorded by wearable motion sensors to a virtual car and driver in a mixed-reality parking game. In A. Blair, S.-B. Cho, and S. M. Lucas, ed., *CIG 2007: Proceedings of the 2nd IEEE Symposium on Computational Intelligence and Games*, pp. 32–39. IEEE Press, April 2007. doi:10.1109/cig.2007.368076.

[7] J. Beidler, A. Insogna, N. Cappobianco, Y. Bi, and M. Borja. The PNA project. *J Comput Small Coll*, 16(4):276–284, May 2001.

[8] M. Beigl, H.-W. Gellersen, and A. Schmidt. MediaCups: Experience with design and use of computer-augmented everyday artefacts. *Computer Networks*, 35(4):401–409, March 2001. ISSN 1389-1286. doi:10.1016/s1389-1286(00)00180-8. Special Issue on Pervasive Computing.

[9] S. Chambers, S. Venkatesh, G. West, and H. Bui. Hierarchical recognition of intentional human gestures for sports video annotation. In R. Kasturi, D. Laurendeau, and C. Suen, ed., *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 2, pp. 1082–1085. IEEE Press, August 2002. doi:10.1109/icpr.2002.1048493.

[10] K.-H. Chang, S.-Y. Liu, H.-H. Chu, J. Y. Hsu, C. Chen, T.-Y. Lin, C.-Y. Chen, and P. Huang. The diet-aware dining table: Observing dietary behaviors over a tabletop surface. In K. Fishkin, B. Schiele, P. Nixon, and A. Quigley, ed., *PERVASIVE 2006: Proceedings of the 4th International Conference on Pervasive Computing*, vol. 3968 of *Lecture Notes in Computer Science*, pp. 366–382. Springer Berlin, Heidelberg, May 2006.

[11] J. A. Y. Cichero and B. E. Murdoch. Detection of swallowing sounds: methodology revisited. *Dysphagia*, 17(1):40–49, 2002.

[12] C. Dacremont, B. Colas, and F. Sauvageot. Contribution of air- and bone-conduction to the creation of sounds perceived during sensory evaluation of foods. *J Tex Stud*, 22 (4):443–456, January 1991. doi:10.1111/j.1745-4603.1991.tb00503.x.

[13] C. Danbolt, P. Hult, L. T. Grahn, and P. Ask. Validation and characterization of the computerized laryngeal analyzer (CLA) technique. *Dysphagia*, 14(4):191–195, 1999.

[14] N. DeBelie, V. De Smedt, and D. B. J. Principal component analysis of chewing sounds to detect differences in apple crispness. *Postharvest Biol Technol*, 18:109–119, 2000.

[15] N. DeBelie, M. Sivertsvik, and J. DeBaerdemaeker. Differences in chewing sounds of dry-crisp snacks by multivariate data analysis. *J Sound Vib*, 266(3):625–643, September 2003.

[16] D. M. Denk, H. Swoboda, and E. Steiner. Physiology of the larynx. *Radiologe*, 38(2): 63–70, Feb 1998. In German.

[17] B. Drake. Food crushing sounds. an introductory study. *J Food Sci*, 28(2):233–241, March 1963. doi:10.1111/j.1365-2621.1963.tb00190.x.

[18] J. Edmister and Z. Vickers. Instrumental acoustical measures of crispness in foods. *J Tex Stud*, 16(2):153–167, 1985.

[19] C. Ertekin and I. Aydogdu. Neurophysiology of swallowing. *Clin Neurophysiol*, 114 (12):2226–2244, Dec 2003.

[20] V. Gupta, N. P. Reddy, and E. P. Canilang. Surface EMG measurements at the throat during dry and wet swallowing. *Dysphagia*, 11(3):173–179, 1996.

[21] T. K. Ho, J. Hull, and S. Srihari. Decision combination in multiple classifier systems. *IEEE T Pattern Anal*, 16(1):66–75, Jan. 1994. doi:10.1109/34.273716.

[22] H. R. Kissileff, G. Klingsberg, and T. B. V. Itallie. Universal eating monitor for continuous recording of solid or liquid consumption in man. *Am J Physiol*, 238(1):R14–R22, Jan 1980.

[23] C. Lee and X. Yangsheng. Online, interactive learning of gestures for human/robot interfaces. In N. Caplan and C. G. Lee, ed., *ICRA 1996: Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 4 of *IEEE Robotics and Automation Society*, pp. 2982–2987. IEEE Press, April 1996.

[24] H.-K. Lee and J. H. Kim. An HMM-based threshold model approach for gesture recognition. *IEEE T Pattern Anal*, 21(10):961–973, October 1999.

[25] A. Limdi, M. McCutcheon, E. Taub, W. Whitehead, and I. Cook, E.W. Design of a microcontroller-based device for deglutition detection and biofeedback. In *EMBS 1989: Proceedings of the Annual International Conference of the IEEE Engineering in Engineering in Medicine and Biology Society.*, vol. 5, pp. 1393–1394. IEEE Press, November 1989. doi:10.1109/iembs.1989.96257.

[26] W. J. Logan, J. F. Kavanagh, and A. W. Wornall. Sonic correlates of human deglutition. *J Appl Physiol*, 23(2):279–284, Aug 1967.

[27] J. Mankoff, G. Hsieh, H. C. Hung, S. Lee, and E. Nitao. Using low-cost sensing to support nutritional awareness. In G. Goos, J. Hartmanis, and J. van Leeuwen, ed., *Ubicomp 2002: Proceedings of the 4th International Conference on Ubiquitous Computing*, vol. 2498 of *Lecture Notes in Computer Science*, pp. 371–376. Springer Berlin, Heidelberg, September-October 2002.

[28] MyFoodPhone. World's first camera-phone & web-based-video nutrition service. Internet, Feb 2005. Accessed: August 2007.

[29] G. Ogris, T. Stiefmeier, H. Junker, P. Lukowicz, and G. Troster. Using ultrasonic hand tracking to augment motion analysis based recognition of manipulative gestures. In B. Rhodes and K. Mase, ed., *ISWC 2005: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*, pp. 152–159. IEEE Press, October 2005. doi: 10.1109/iswc.2005.54.

[30] D. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. In B. Rhodes and K. Mase, ed., *ISWC 2005: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*, pp. 44–51. IEEE Press, October 2005. doi:10.1109/iswc.2005.22.

[31] B. J. Rolls, A. Drewnowski, and J. H. Ledikwe. Changing the energy density of the diet as a strategy for weight management. *J Am Diet Assoc*, 105(5 Suppl 1):S98–103, May 2005. doi:10.1016/j.jada.2005.02.033.

[32] A. Schmidt, M. Strohbach, K. van Laerhoven, A. Friday, and H.-W. Gellersen. Context acquisition based on load sensing. In G. Goos, J. Hartmanis, and J. van Leeuwen, ed., *Ubicomp 2002: Proceedings of the 4th international conference on Ubiquitous Computing*, vol. 2498 of *Lecture Notes in Computer Science*, pp. 333–350. Springer Berlin, Heidelberg, September-October 2002.

[33] K. A. Siek, K. H. Connelly, Y. Rogers, P. Rohwer, D. Lambert, and J. L. Welch. When do we eat? an evaluation of food items input into an electronic food monitoring application. In E. Aarts, R. Kohno, P. Lukowicz, and J. C. Trainini, ed., *PHC 2006: Proceedings of the 1st International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–10. ICST, IEEE digital library, November 2006. doi:10.1109/pcthealth.2006.361684.

[34] A. Sjöberg, L. Hallberg, D. Höglund, and L. Hulthen. Meal pattern, food choice, nutrient intake and lifestyle factors in the Göteborg Adolescence Study. *Eur J Clin Nutr*, 57 (12):1569–1578, Dec 2003. doi:10.1038/sj.ejcn.1601726.

[35] E. Stellar and E. E. Shrager. Chews and swallows and the microstructure of eating. *Am J Clin Nutr*, 42(5 Suppl):973–982, Nov 1985.

[36] A. A. Stone, S. Shiffman, J. E. Schwartz, J. E. Broderick, and M. R. Hufford. Patient non-compliance with paper diaries. *BMJ*, 324(7347):1193–1194, May 2002.

[37] S. M. Sukthankar, N. P. Reddy, E. P. Canilang, L. Stephenson, and R. Thomas. Design and development of portable biofeedback systems for use in oral dysphagia rehabilitation. *Med Eng Phys*, 16(5):430–435, Sep 1994.

[38] Z. M. Vickers. The relationships of pitch, loudness and eating technique to judgments of the crispness and crunchiness of food sounds. *J Tex Stud*, 16(1):85–95, 1985.

[39] J. Ward, P. Lukowicz, G. Tröster, and T. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE T Pattern Anal*, 28(10): 1553–1567, Oct. 2006. doi:10.1109/tpami.2006.197.

[40] J. A. Ward, P. Lukowicz, and G. Tröster. Evaluating performance in continuous context recognition using event-driven error characterisation. In M. Hazas, J. Krumm, and T. Strang, ed., *LoCA 2006: Proceedings of the Second International Workshop on Location- and Context-Awareness*, vol. 3987 of *Lecture Notes in Computer Science*, pp. 239–255. Springer, Berlin/Heidelberg, May 2006. doi:10.1007/11752967_16.

[41] M. S. Westerterp-Plantenga. Eating behavior in humans, characterized by cumulative food intake curves–a review. *Neurosci Biobehav Rev*, 24(2):239–248, Mar 2000. doi: 10.1016/s0149-7634(99)00077-9.

[42] WHO. Global strategy on diet, physical activity and health (WHA57.17). In *Fiftyseventh World Health Assembly*. World Health Organization, May 2004.

5

Gesture spotting to detect user activities

Holger Junker, Oliver Amft, Paul Lukowicz and Gerhard Tröster

Full publication title: Gesture spotting with body-worn inertial sensors to detect user activities.

Pattern Recognition, 41(6), 2010–2024, June 2008.
DOI: 10.1016/j.patcog.2007.11.016

Abstract

We present a method for spotting sporadically occurring gestures in a continuous data stream from body-worn inertial sensors. Our method is based on a natural partitioning of continuous sensor signals and uses a two-stage approach for the spotting task. In a first stage, signal sections likely to contain specific motion events are preselected using a simple similarity search. Those preselected sections are then further classified in a second stage, exploiting the recognition capabilities of Hidden Markov models. Based on two case studies, we discuss implementation details of our approach and show that it is a feasible strategy for the spotting of various types of motion events.

5.1. Introduction

Monitoring and classification of human activity using simple body-worn sensors is emerging as an important research area in machine learning. Activity monitoring itself is motivated by a variety of mobile and ubiquitous computing applications, such as personalisation of the user interface, behavioural monitoring in medicine, medication assessment, assistive systems for the elderly and cognitively disabled or intelligent information delivery and recording systems for industrial assembly and maintenance.

The choice of simple sensors, such as accelerometers instead of computer vision stems from the limited computational resources of mobile and ubiquitous systems and the very diversified, dynamic environment in which such systems need to operate. The later often implies varying light conditions, changing backgrounds and a large clutter. This makes extracting relevant information from visual signals difficult and computationally intensive. Body-mounted motion sensors on the other hand, are influenced by user activity only. The problem with activity recognition using such sensors lies less in the extraction of relevant features than in the fact that the information is often ambiguous and incomplete. Thus, once a vision system has managed to track, for example the user's arm, relatively exact trajectories could be obtained for activity recognition. In contrast, arm worn accelerometers react to a combination of earth gravity and arm speed changes. Gyroscopes describe rotational motions of the arm. However none of the above provides exact trajectory information.

Despite the disadvantages listed above, body-worn motion sensors have been successfully used for a variety of tasks (see related work). One area where little progress has been made so far, is the spotting of sporadically occurring activities in a continuous data stream. This is known to be difficult, even if complete trajectory information is available from a vision system. It is even more difficult in a wearable sensors based environment.

This paper describes a novel method for tackling this problem based on appropriately adapted machine learning techniques. Focusing on activities associated with distinct arm gestures, the performance of the proposed method is evaluated in two elaborate case studies.

5.1.1. Paper Scope and Background

Depending on the specific application, very different types of activity recognition are needed. As an example, consider a system designed to monitor the overall physical activity level of a person. The idea behind such systems is to provide general information about the effect of certain behavioural recommendations or to estimate energy expenditure without having the patient admitted to stationary care or a laboratory for observation. A wearable system deploying appropriate body-worn sensors can be used to collect this data. Obviously, the type of information, that such systems need to deliver is not about single,

specific actions, but more about the overall level of activity. Often, the activity level can be assessed by averaging parameters, such as mean acceleration of specific body parts. In a way, this is a very simple form of activity recognition.

On the other side of the spectrum are applications, where reliable recognition on a more fine-grained level is needed. Such applications may include e.g. the monitoring of specific tasks and/or movements in a rehabilitation scenario, the spotting of specific gestures for novel, more natural human computer interface or the classification of dietary intake gestures for an automated diet monitoring system. Such recognition tasks are particularly difficult, because the relevant activities occur sporadically in between a large variety of other activities. For example, in between the actual activity a user might fetch tools, drink, chat with anther person or just scratch the head. As a consequence, the task at hand can be described as *activity spotting*. It is widely recognised as a particularly complex domain of activity recognition and is still an open problem.

The work described in this paper is part of a larger effort of our groups, directed at this problem, e.g. [31, 34, 35]. It focuses on activities that are associated with a characteristic arm gestures. For such activities, the paper presents a novel gesture spotting method based on arm-worn motion sensors. The method uses a natural partitioning of human motions. In order to achieve a balance between precision and recall with reasonable computational effort, the task is partitioned into a fast highly sensitive stage to pick up potentially interesting signal segments and a more complex, highly selective second stage to narrow down the selection and get rid of false positives.

Our method is primarily intended as part of a large activity spotting system that uses additional information such as location, modes of locomotion, e.g. sitting standing, walking [14], supplementary location sensors [26] or information on objects involved in the activity [27]. Nonetheless, we present experiments on activities from two different everyday life domains indicating, that even on its own our method achieves reasonable performance.

5.1.2. Related work

In contrast to isolated motion recognition that has been shown in various areas, the spotting task is much more challenging. The difficulty of spotting specific human motion events stems from a number of sources. These include, among others, co-articulation, where consecutive gestures influence each other [13], as well as intra- and inter-person variability. Another challenge, the system has to deal with, is the fact that the motion events to be spotted may only occur sporadically, in a continuous data stream, while at the same time being embedded into other, partly arbitrary movements (called *zero class*). These movements however are inherently difficult to model, due to their complexity and unpredictability. As a consequence, conventional recognition schemes for continuous classification, such as Hidden Markov Models (HMMs) are not di-

rectly applicable for our recognition task, since they rely on appropriate zero class models. Consequently, we cannot take advantage of the implicit data segmentation capabilities, that HMMs provide. Moreover, we have to deal with the fact, that motion events are typically very short. This means that for any explicit segmentation-based recognition, exact localisation of event boundaries is important.

The recognition of gestures has been studied extensively over years and many approaches have been proposed to tackle the diverse problems. In general, these approaches can be broadly categorised in either of the two following categories: Gesture recognition, requiring external infrastructure and gesture recognition, focusing on wearable instrumentation.

The first category is dominated by vision-based motion recognition, using a single or multiple cameras. While an exhaustive review of literature is beyond the scope of this work, we exemplary indicate related works. Starner [32] proposed an approach for American Sign Language recognition, Campbell and Bobick [10] developed a system for recognising classical ballet steps, Yamato et al. [37] worked on the recognition of different tennis strokes, Brand et al. [7] targeted T'ai Chi movements, Lee and Kim [20] dealt with typical gestures for interacting with a computer and Rao and Shah [29] aimed at manipulative gestures. Further literature on vision-based motion capture and recognition can be found in [24, 30, 36].

More recently the use of wearable instrumentation for gesture recognition has gained much attention mainly due to the success in sensor miniaturisation. Various approaches dealing with the recognition of activities or events have been presented. Chambers et al. [11] targeted Kung Fu moves and Benbasat [5] focused on the recognition of "atomic" gestures. Kern et al. [18] looked at activities, such as keyboard typing, writing on a white-board and shaking hands. Cakmakci et al. [9] tried to identify when a person was looking at the watch. Bao [4] aimed at typical household activities including vacuuming, folding laundry, watching TV or brushing teeth. Lukowicz et al. [23] concentrated on workshop activities including sawing, hammering, drilling, and filing. Brashear et al. [8] dealt with gestures for American Sign Language and Lementec and Bajcsy [21] worked on the recognition of gestures used to instruct pilots after landing.

Although many motion recognition approaches exist, few are dealing with the spotting task itself. Deng and Tsui [12] proposed a method for spotting gestures in continuous data. Their approach makes use of an HMM-based accumulation score, that supports endpoint detection of a particular gesture in a continuous data stream. Based on a potential endpoint their algorithm searches for a corresponding start point using the viterbi algorithm. While this approach seems promising, it has been evaluated solely for the recognition of two-dimensional trajectories (Arabic numbers). Lee and Kim [20] developed a method deploying HMMs directly, to spot gestures in a continuous stream of sensor data. They introduced the concept of a threshold model that calculates

the likelihood threshold of an input pattern and provides a confirmation mechanism for the provisionally matched gesture patterns. The threshold model is a weak model for all trained gestures and is constructed from all existing gesture models. Lukowicz et al. [23] demonstrated continuous, on-line motion recognition by partitioning the incoming data using an intensity analysis based on the signals of two microphones exploiting the fact that the movements to be recognised are accompanied with a particular sound.

While the first two approaches made use of the implicit segmentation capabilities of HMMs, the third approach used an explicit segmentation step to facilitate spotting. We believe that explicit gesture segmentation can be very helpful and efficient to facilitate the spotting task. Lee and Yangsheng [19] developed a system for online gesture recognition using HMMs. They were among the first researchers to use segmentation as a pre-processing step to gesture recognition and were able to recognise 14 different gestures online. While they proposed acceleration thresholds for segmentation, they used a simple velocity-based segmentation relying on the fact that there must be short pauses between two consecutive gestures. They successfully demonstrated good recognition performance for the trained gestures, however they did not deal with the rejection of non-relevant movements. Kahol et al. [16] proposed a gesture segmentation algorithm which employs a hierarchical layered structure to represent the human anatomy. The algorithm used low-level motion parameters to characterise motion in the various layers of this hierarchy and was able to predict segmentation boundaries based on profiles, generated from segmentation results. The segmentation, in turn, was provided by observers, who manually segmented training data. In a recent work, Kahol et al. [15] used the concept to fully document every motion in dance activities using a Vicon camera system. Wang et al. [33] presented an approach for automatically segmenting sequences of natural activities into atomic sections and clustering them. The segmentation was based on finding the local minimum of velocity and local maximum of change in direction. The minimum below and the maximum above the certain threshold were selected as segment points. The limitation of their approach is that it can only segment and label continuous human gestures, but not spot them. Liang and Ouhyoung [22] used a temporal segmentation based on the discontinuity of the movements according to four gesture parameters and HMMs to perform real-time continuous gesture recognition of sign language. Their approach allows the recognition of gestures that were defined in vocabularies only, thus rejection of non-gesture patterns is not considered. Morguet [25] proposed a two-step approach to the continuous recognition of gestures in video sequences. In a first step, a simple segmentation algorithm was used to identify start and end points of potentially meaningful segments. This segmentation process used a threshold on a specific motion parameter in conjunction with simple rules to obtain valid segments. These segments were then classified in isolation. However, this approach cannot reject non-gesture patterns that are falsely retrieved in the first stage.

5.1.3. Paper Contributions and Organisation

As stated in the introduction, the work presented in this paper is part of a large effort towards reliable spotting of complex activities using simple on-body sensors. It focuses on the recognition of gestures, that build the basis for the inference of more abstract activities. The primary aim is to support complex activity spotting systems rather then to develop an activity spotting system based solely on arm gestures. Nonetheless, we show how, for suitable domains, good performance can be achieved without any additional information.

Within this scope the paper makes the following contributions:

1. It presents a novel, two-stage gesture spotting method based on body-worn motion sensors. The method is specifically designed towards the needs and constraints of activity recognition in wearable and pervasive systems. This includes a large null class, lack of appropriate models for the null class, large variability in the way gestures are performed and a variable gesture length. It also refrains from excessively computationally intensive operations such as correlations over large data sets or complex searches. Instead, it uses a natural partitioning of human motions, combined with a simple parametrisation scheme as a computationally cheap preselection stage that identifies potentially interesting data sections. These sections are then reevaluated using HMMs to reduce the number of classification errors. This combination of a cheap, highly sensitive initial stage with a highly selective second stage is what makes our approach unique and well suited to the intended domain.

2. The paper describes the verification of the proposed method on two scenarios that together comprise of nearly a thousand relevant gestures. The first one, interaction with different everyday objects, is part of a wide range of wearable systems applications. The second one, food intake, is a highly specialised application motivated by the needs of a large industry dominated health monitoring project. In both cases studies we arrive at recall values between 80% and 90% and a precision of over 70%. The significance of these case studies and results is twofold. First, they confirm the soundness of our approach. Second they are a strong indication for the feasibility of reliable activity spotting using wearable sensors, in particular, since the approach presented in this paper is meant to be used as part of a large system that uses other information to further improve the results.

As indicated in the related work section, two-stage activity spotting approaches have been tried before. However, to our knowledge, the specific approach described in this paper, with its focus on the peculiarities of activity spotting using simple sensors and wearable systems is novel. Taking into account the results achieved in our case studies it represents a significant contribution to the field.

Paper Organisation In Section 2 of the paper, we introduce our two-stage spotting approach. In Section 3, we describe the case studies used to validate our approach. In Section 4, we focus on implementation details of the spotting algorithm and in Section 5, we detail the experimental setup to acquire sensor data for the case studies. In Section 6, we finally present our evaluation results. In Sections 7 and 8, we discuss the results and highlight future work, respectively.

5.2. Spotting Approach

Our two-stage spotting approach consists of a preselection stage (1st stage) and a classification stage (2nd stage) as shown in Figure 5.1. The task of the preselection stage is to localise and preselect sections in the continuous signal stream, likely to contain relevant motion events. These candidate sections are then passed on to the classification stage and are classified in isolation using appropriate classifiers.

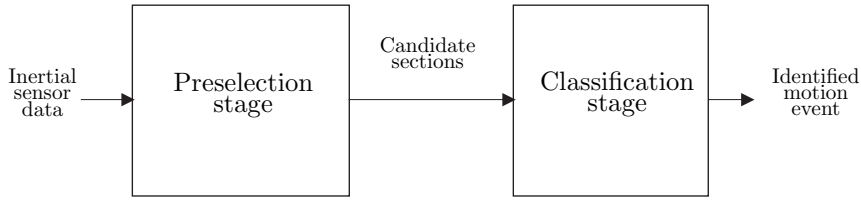


Figure 5.1. Sensor data flow through the two-stage recognition framework.

The preselection of sections in a continuous signal stream can be considered as a search problem. In a naive approach, the search may be performed on all possible sections in the data stream. The major problem of this exhaustive approach is its computational effort. To reliably capture human motions with inertial sensors, the sensors are usually sampled with up to 100 Hz. Considering that a relevant motion event may take several seconds, the above mentioned search strategy would require to check a large number of sections.

Obviously, one solution to reduce the complexity is to apply a coarse search, where not all but only certain sections in the continuous signal stream are considered for the search. One way to implement such a coarse search is to partition the signal stream into segments which are significantly longer than a single sampling interval and to consider the segment boundaries as possible start/end points of the sections to be searched. However, an artificial partitioning is likely to miss the exact boundaries of the relevant motion events contained in the data stream. This makes the recognition more complicated, since sections may contain only parts of the relevant motion event as well as other motions.

We propose to use a natural partitioning of the data into 'motion segments'. Inspired by the taxonomy of Bobick [6] these motion segments are described as non-overlapping, atomic units of human movement, characterised by their spatio-temporal trajectories. Assuming that a motion event can be subdivided into a sequence of motion segments, we can obtain a natural, non-ambiguous partitioning of the overall motion with the start and end of the motion events corresponding to the start/end of a specific motion segment. Thus, the search can be constrained to those sections, whose boundaries coincide with the boundaries of the motion segments. Table 5.1 summarises the terminology used in this work.

For the partitioning task, motion parameter(s) were used that represent the motion event closely. The number and types of motion parameters to be used is specific to the motion event to be recognised. For arm-related motion events, they include e.g. relative orientation information, such as joint angles between the lower and the upper arm, absolute orientation information of the arm segments to an earth-fixed reference frame, or simply the raw signals from the sensors attached to the arm segments.

While the preselection stage identifies potential candidate sections, the classification stage is used to eliminate those sections that have been falsely retrieved in the preselection stage. This is achieved by individually classifying the candidate sections using HMMs and comparing the classification result to the result of the preselection stage.

The main motivation behind this two-stage approach is to reduce the complexity of the spotting task, by constraining the search space within the continuous data stream and by applying a simple similarity analysis to preselect potential sections. The subsequent classification stage is used to make the recognition more robust and retain only relevant sections.

5.3. Case Studies

In order to discuss the implementation of our approach, we considered the spotting of typical, everyday-life gestures in a continuous data stream from body-worn inertial sensors. Specifically, we investigated two different case studies:

Case study 1 deals with the spotting of diverse object interaction gestures, reflecting common activities of daily living. The detection of such gestures is considered as key component in a context recognition system, to monitor complex human activities. Furthermore, such gestures may facilitate more natural human-computer interfaces.

Case study 2 focuses on dietary intake gestures. The spotting of body motions related to human intake are expected to become one sensing domain of an automated dietary monitoring system [1]. Although the automatic determination of exact type and amount of all foods is rather visionary, we believe that an

Table 5.1. Applied terminology of human motion in this work.

Term	Description
Motion segment	Represents atomic, non-overlapping unit of human motion that can be characterised by their spatio-temporal trajectory.
Motion event	Span a sequence of motion segments. A gesture can be considered as a particular class of motion events, mainly involving movements of the arms and trunk.
Activity	Describes a situation, that may consist of various motion events. Thus, it refers to higher-level context.
Signal segment	A slice of sensor data that corresponds to a motion segment.
Candidate section	A slice of sensor data that may contain a gesture.

assistive system based on different sensors is conceivable. Hence, the gestures included in this study refer to frequently used human feeding motions. Detecting such gestures reveals information about the timing of nutrition events, e.g. breakfast or lunch and on the category of the food item, e.g. a soup is fed with a spoon, a glass, cup or bottle is usually used for drinking.

Figure 5.2 and Figure 5.3 illustrate the gestures that we aimed to recognise (relevant gestures) in each case study (see Table 5.2 for a brief description). All relevant gestures are characterised by distinctive movements of the left or right arm. While in case study 1 only movements from the right arm and trunk were used to detect the gestures, case study 2 uses information from both arms as well as from the trunk.

5.4. Spotting Implementation

The implementation of our two-stage spotting approach is detailed in this section. The first stage preselects candidate sections and the second stage refines the preselection (see Figure 5.4).

5.4.1. Preselection Stage

This section details the segmentation scheme used for the initial partitioning of the continuous signal stream into motion segments, the search strategy and similarity measure applied to identify potential sections and finally, the selection of candidate sections.

Table 5.2. Description of the relevant gestures in case study 1 and 2. Unless otherwise noted, all gestures were conducted with the right arm pointing downwards at start/end in case study 1 and with both arms at rest on table at start/end in case study 2.

	Gesture	Description
Case Study 1	Light button (LB)	Press light button to turn lights on.
	Handshake (HS)	Greet person by shaking hands.
	Phone up (PU)	Pick up receiver. Start position: arm resting on leg, end position: hold receiver to ear.
	Phone down (PD)	Put down telephone receiver: End position: arm resting on leg.
	Door (DR)	Turn door knob and open door of cabinet.
	Coin (CN)	Take out purse from right back pocket of trousers - open purse with right hand - take coin and insert it into slot of vending machine - close purse with right hand - put purse back into pocket.
Case Study 2	Cutlery (CL)	Meal intake of Lasagne using fork and knife. Fork tap, loading and manoeuvring to mouth and back with left hand.
	Drink (DK)	Pick up cup from table - take a sip - put cup back on table.
	Spoon (SP)	Meal intake of cereals or soup using a spoon. Spoon loading and manoeuvring to mouth and back.
	Hand (HD)	Meal intake of bread slice or chocolate bar using the hand only: Moving the left hand to mouth and back.

(a) Handshake

(b) Light button

(c) Door

(d) Phone

(e) Coin

Figure 5.2. Visualisation of the relevant gestures (acted) as performed in case study 1.

(a) Cutlery

(b) Drink

(c) Spoon

(d) Hand

Figure 5.3. Visualisation of the relevant gestures (acted) as performed in case study 2.

Motion segment partitioning

The task of the segmentation algorithm was to partition a motion parameter into non-overlapping, meaningful segments. This task can be considered as

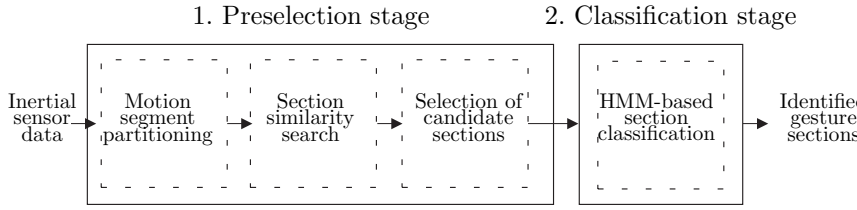


Figure 5.4. Detailed structure of the two-stage recognition framework.

a time-series segmentation problem, which has been extensively studied in many application domains. An excellent review of time series segmentation approaches was provided by Keogh et al. [17].

As motion parameter, we used the pitch and the roll of the lower arm, which are the angle of the lower arm segment to the horizontal plane and the rotation angle with the rotation axis along the limb of the lower arm (see Figure 5.5).

These angles have been chosen mainly for the following reasons: Many movements of the entire arm typically involve movements of the lower arm as well. Furthermore, the signals of the lower arm orientation (and in particular pitch and roll) correlated well with our visual perception of the gestures. Despite good initial results by using the pitch in case study 1, the roll was additionally investigated in case study 2. For certain gestures the segmentation based on the roll matched the gesture boundaries better. This can be explained by the typical feeding motion (moving the hand with a tool to the mouth), involved in the gestures of case study 2.

Although relative orientation information between the lower arm and the upper arm segment, such as joint angles would generally be well suited for the partitioning of the signal streams, we found that the estimation of those angles using inertial sensors attached to the arm segments can be prone to large errors. The two major sources of errors were inaccurate orientation estimation of the involved sensors (mainly due to magnetic disturbances) and the loose attachment of the sensors to the arm segments. Attachment issues make the sensors susceptible to displacement while moving the arm. Conversely the pitch and roll of the lower arm could be derived very robustly. The estimation of these angles from raw sensor data was less prone to magnetic disturbances than other orientation angles, specifically the orientation in the horizontal plane.

For the segmentation task, we used the Sliding-window and Bottom-Up algorithm (SWAB) introduced by Keogh et al. [17]. Based on the evaluation of typical test data, we found the algorithm to be well suited for our application. SWAB combines the advantages of a precise bottom-up segmentation scheme with those of a sliding window algorithm. This allows the algorithm to be used on-line while keeping a global view on the data.

The algorithm kept a small buffer of the signal data. A bottom-up segmentation was applied to the data in the buffer. From the resulting signal

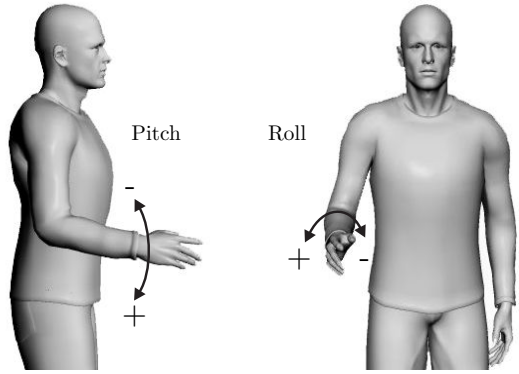


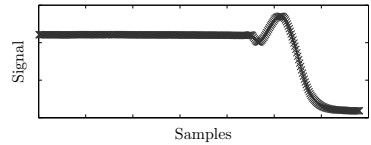
Figure 5.5. Orientation angle 'pitch' and 'roll' of the lower arm segment.

segmentation, the segment with the oldest data was extracted from the buffer and new data was added using the sliding window approach. This procedure was repeated as long as new data was available, potentially forever.

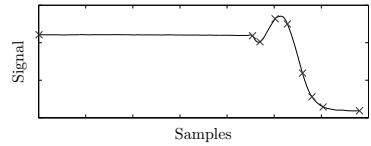
The bottom-up partitioning of each buffer of length n started from the arbitrary segmentation of the signal into n/2 segments. Next, the cost of merging each pair of adjacent segments was calculated and the lowest cost pair in the buffer was iteratively merged. As the algorithm iterates, more signal segments were merged until all adjacent segments in the buffer exceeded a cost threshold when merged. Figure 5.6(a)-5.6(b) depicts the segmentation process of the buffered signal for different segmentation steps (iterations). At iteration 0 the fine-grained initial partitioning can be seen. The final state is depicted in Figure 5.6(b). The sliding-window algorithm of SWAB reported the left-most segment from the bottom-up buffer and added new data accordingly. The procedure was restarted with this new data in the bottom-up buffer.

The cost metric for merging two segments was based on the error of approximating the signal with its linear regression (residuals) in the bounds defined by the merged segment. This method can be explained as follows: When the pair of segments differ strongly in its signal shape, the approximation of the merged segments incurs large residuals. Hence it is less likely that the segments belong to the same motion segment. We used the squared sum of the residuals in the bounds of the merged segment as cost function.

To ensure that the algorithm provided a good approximation of the signal, a small cost threshold was required, typically leading to a large number of segments for any of the relevant gestures. These segments did not correspond well to the small number of visually perceived sub-movements of the gesture. As a solution to this problem, we merged adjacent segments, as created by the SWAB-algorithm, if their linear regressions had similar slopes. As result of this extension we obtained motion segment boundaries. Figure 5.7 depicts



(a) Initialisation (iteration 0)



(b) Termination (iteration 332)

Figure 5.6. Segmentation of an example signal stored in the bottom-up buffer at different algorithm iterations. The 'cross'-symbols indicate segment boundaries.

an example of the segmentation steps, based on the 'DK' gesture that uses the pitch angle as segmentation signal.

For each gesture an individual segmentation parameter could be chosen. Person-specific training was used to accommodate for the dominant body side. In the investigated case studies, the body side was fixed. Table 5.3 summarises the final choices made in our implementation.

Table 5.3. Motion parameter selection for the SWAB algorithm.

	Gesture	**SWAB motion parameter**	**Body side used in studies**
Study 1	Light button (LB), Handshake (HS), Phone up (PU), Phone down (PD), Door (DR), Coin (CN)	$pitch_{LA}(t)$ [a]	right
Study 2	Cutlery (CL)	$roll_{LA}(t)$ [a]	left
	Drink (DK), Spoon (SP)	$pitch_{LA}(t)$	right
	Hand (HD)	$pitch_{LA}(t)$	left

[a] Pitch, roll, and yaw are Euler angles representing rotations of an object in 3-dimensional Euclidean space. The orientation of pitch and roll angles are described in Section 5.4.1 and Fig. 5.5. The yaw angle corresponds to absolute orientation in the horizontal plane.

The mean number of segmentation points per gesture for the data sets of both case studies are shown in Table 5.4. The ratio of segmentation points to the total recorded samples indicates the achieved reduction in search effort.

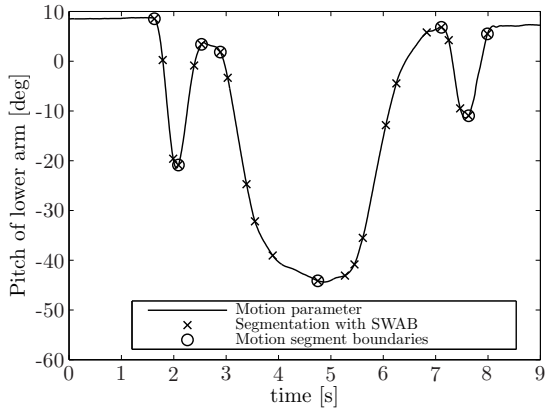


Figure 5.7. Segmentation of the 'DK'-gesture using the pitch of the lower arm as segmentation signal. The cross symbols ('x') correspond to segmentation boundaries obtained from the SWAB-algorithm. The circles ('o') highlight the remaining segmentation points (motion segment boundaries) based on the proposed extension of the segmentation algorithm.

Table 5.4. SWAB segmentation results. The SWAB segmentation points correspond to the total number of segmentation points for the entire data sets. The ratio of segmentation points by total recorded samples indicates the reduction in search effort achieved by the preselection stage.

Segmentation category	Case Study 1	Case Study 2
Mean number of SWAB segmentation points per gesture	15506	13020
Ratio of segmentation points per gesture by total recorded samples	2.2%	0.77%

Section similarity search

A coarse search based on the motion segment boundaries was used to find sections that contain relevant gestures. The search was performed by considering each motion segment endpoint as potential end of a gesture. For each endpoint, potential start points were derived from preceding motion segment boundaries. The search was performed for each gesture separately. To confine the search space, we introduced two constraints on the sections to be searched. These constraints were adapted to the gesture by training data:

1. For the actual length T of the section we constrained $T_{min} \leq T \leq T_{max}$, where T_{min} and T_{max} denote the minimum and maximum length of the section to be considered.

2. For the number of motion segments n_{MS} in the actual section, we selected $N_{MS,min} \leq n_{MS} \leq N_{MS,max}$ where $N_{MS,min}$ and $N_{MS,max}$ corresponds to the minimum and maximum number of motion segments to be contained in the section, respectively.

As search criterion, we used the normalised Euclidean distance[1] given in Eq. 5.1, where f_{PS} denotes the N_F-dimensional feature vector of the preselection stage, derived from the section under consideration.

We used simple single-value features, such as minimum and maximum signal values of the lower and upper arm pitch and roll, sum of signal samples, the duration of the gesture and the number of motion segments in the section under consideration. In case study 2, the minimal distance between the hand and estimated head position was additionally used.

The parameters μ_{ik} and σ_{ik} represent the mean and the standard deviation of the i-th element of the feature vector of gesture G_k. These were computed from training data.

$$d(\mathbf{f}_{PS}; G_k) = \sqrt{\sum_{i=1}^{N_F} \left(\frac{f_{PS_i} - \mu_{ik}}{\sigma_{ik}} \right)^2}, \qquad \mathbf{f_{PS}} = [f_{PS_1}, .., f_{PS_{N_F}}] \tag{5.1}$$

The normalised Euclidean distance provided a measure of how similar the motion pattern given in the section were to a specific gesture. During the evaluation of all possible start points for one endpoint, only the section with the minimal distance were retained.

If the distance $d(\mathbf{f}_{PS}, G_k)$ was smaller than a gesture-specific threshold value $d_{min}(G_k)$, the section under investigation was considered to contain gesture G_k. If the condition was satisfied for more than one gesture, the section was considered to contain either one of the corresponding gestures. Depending on the application such collisions need to be checked and handled.

Selection of candidate sections

Figure 5.8 schematically shows the collision of two sections obtained by the section search procedure. These overlapping candidate sections were resolved by selecting sections with the smallest similarity values for every occurring collision. In this way non-overlapping candidate sections were obtained for a particular gesture.

[1]The normalised Euclidean distance corresponds to the Mahalanobis distance where the covariance matrix is a diagonal matrix.

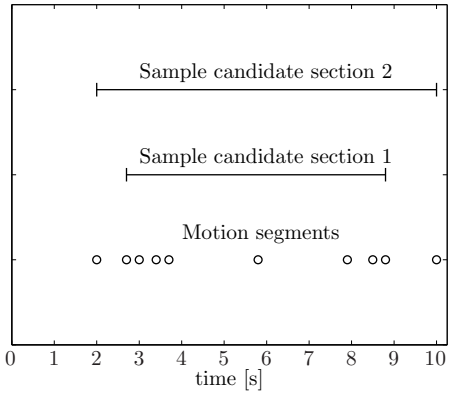


Figure 5.8. Overlapping candidate sections.

5.4.2. Classification Stage

In the classification stage, we used Hidden Markov Models (HMMs), which have long been used in speech recognition, due to their ability to cope with temporal and spatial variations of input patterns [28].

For our evaluation, we considered left-right models with eight continuous features. The features used for the classification differ from the features used in the preselection stage. While in the preselection stage, data sections were characterised by single-valued features, such as the minimum and maximum signal value and the duration of the section, the HMMs were fed with time-series features derived from the candidate sections. Moreover, a separate definition of the feature set was useful to address the classification goal.

The following features were used for the HMM-based classification stage:

- Pitch and roll angles from the lower arm sensors.
- Pitch and roll angles from the upper arm sensors.
- Derivative of the acceleration signal from the lower arm sensor, with the measurement orientation along the pitch angle measurement.
- The cumulative sum of the acceleration from the lower arm (orientation as before).
- Derivative of the rate of turn signal from the lower arm sensor, with the measurement orientation along the roll angle measurement.
- The cumulative sum of the rate of turn from the lower arm (orientation as before).

We found that all gestures could be modelled using single Gaussian models. Our gesture models consisted of 4 to 10 states. The choice of the states for each gesture model reflects a trade-off between the complexity of the gesture on the one hand, and available training data, which is necessary to estimate the model parameters properly, on the other. Although some gestures may require more states, we achieved good recognition results with our models, as shown below.

5.5. Experiments

For the experimental evaluation of our approach, we recorded a variety of different data sets using a commercially available measurement system[2] with five inertial sensors placed on the body (see Figure 5.9). Sensors were attached to the wrists, upper arms and on the upper torso.

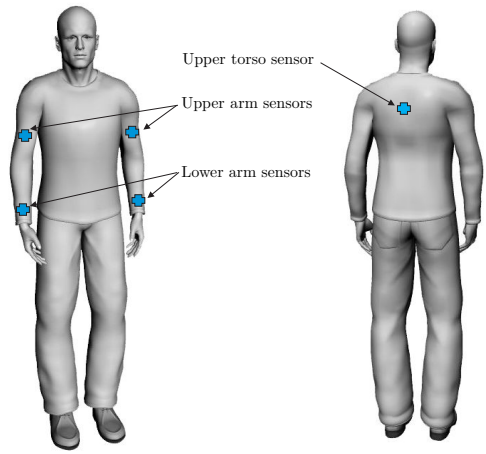


Figure 5.9. Sensor placement for gesture recording.

Using this setup, we independently recorded continuous data sets from one female and three male right-handed subjects, aged 25 to 35 years in both case studies. In case study 2 food intake was recorded in two sessions on different days. The subject data sets (S1.1-S1.4 for case study 1 and S2.1-S2.4 for case study 2) were used for testing of our spotting approach. Additional person-specific data was used for training purposes. The purpose of the studies was explained to the subjects. However, the subjects were asked to perform the movements as natural as possible while wearing the sensors.

In order to obtain data sets with a realistic zero class, we did not set constraints to the movements of the subjects, except that we asked the subject to perform the relevant gestures according to the descriptions given in Table 5.2.

[2]http://www.xsens.com

Moreover, to enrich the diversity of movements and to avoid wide intervals constituting no motion, we defined eight additional gestures to be carried out during the recording which were similar to those gestures we intended to spot. In total 2 hours of motion data were recorded for case study 1, and 4.7 hours for case study 2, with only 25.4% and 34.7% of the data sets containing relevant gestures for case study 1 and 2, respectively (see Table 5.5).

Table 5.5. Statistics of the recorded data sets.

Feature	**Case Study 1**	**Case Study 2**
Total duration of all data sets	7185 Sec (2.00 Hours)	16848 Sec (4.68 Hours)
Share of relevant gestures in data sets	25.4% (1826 Sec)	34.7% (5846 Sec)

5.6. Results

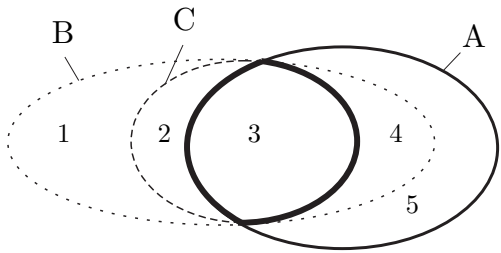
For the evaluation of our approach, the evaluation metrics *Precision* and *Recall* were used. These metrics were derived as follows:

$$Recall = \frac{\text{Recognised Gestures}}{\text{Relevant Gestures}} \quad Precision = \frac{\text{Recognised Gestures}}{\text{Retrieved Gestures}}$$

Relevant gestures are those gestures that have been conducted by the subject, while retrieved gestures represent the sections that have been reported in either preselection stage or classification stage. A recognised gesture is a relevant gesture that has been retrieved. Furthermore, we derived the number of insertions (sections that have been retrieved but do not contain a relevant gesture), and the number of deletions (relevant gestures that have not been reported). Figure 5.10 illustrates the different evaluation metrics schematically. Set A corresponds to the relevant gestures, set B to the retrieved gestures after the preselection stage (PS) and set C to gestures retained after the classification stage (CS). The depicted subsets (1 to 5) reflect the metrics used in this paper.

5.6.1. Preselection Stage

For the spotting of sections likely to contain motion events, appropriate threshold values $d_{min}(G_k)$ were identified for each gesture G_k, by evaluating the performance of the preselection stage on the training data. In general, we observed that the larger the threshold value, the more relevant gestures were retrieved. However, at the same time, the total number of falsely retrieved gestures increased. The precision-recall curves given in Figure 5.11 for the gesture



Relevant:	3 + 4 + 5 (Set A)
Recognised:	3
Retrieved (PS):	1 + 2 + 3 + 4 (Set B)
Insertions (PS):	1 + 2
Deletions (PS):	5
Retrieved (CS):	2 + 3 (Set C)
Insertions (CS):	2
Deletions (CS):	4 + 5

Figure 5.10. Visualisation of the applied evaluation metrics for the preselection (PS) and classification (CS) stages.

'HS' from case study 1 and Figure 5.12 for the gesture 'SP' from case study 2 illustrate this trade-off for the test data sets (S1.1 to S2.4) respectively. Moreover, the individual curves in Figures indicate the variation of the detection performance among the subjects.

The vertical lines towards the maximum recall in Figure 5.12 can be seen as limitation of the similarity search. For these gestures, some instances were not successfully detected due to variation between training and testing gestures.

Based on such precision-recall curves derived from training data, appropriate threshold values can be chosen considering application-specific requirements. For further evaluation of our approach, we set the thresholds for the individual gestures such, that at least 90% of the relevant gestures (gestures that have been conducted) were retrieved in case study 1 and 70% of the relevant in case study 2. This corresponds to a recall value of 0.90 and 0.70 respectively.

Table 5.6 finally summarises the results of the preselection stage for both case studies. For an overall recall value larger than 0.90, we obtained an overall precision value of 0.47 in case study 1. In case study 2, with a recall larger than 0.70, precision dropped to 0.57. The low precision indicated many falsely retrieved sections, that did not contain a relevant gesture (insertions). As can be seen, the spotting of simple gestures such as 'HS' and 'HD' tend to cause more insertions (smaller precision values) than the others.

5.6.2. Classification Stage

We used HMMs to refine the spotting results from the preselection stage.

Model Training and Initial Testing

To accommodate for varying quality in the training process, that is due to random initialisation of certain HMM parameters, we trained 10 instantiations of each model and kept the one with the highest score.

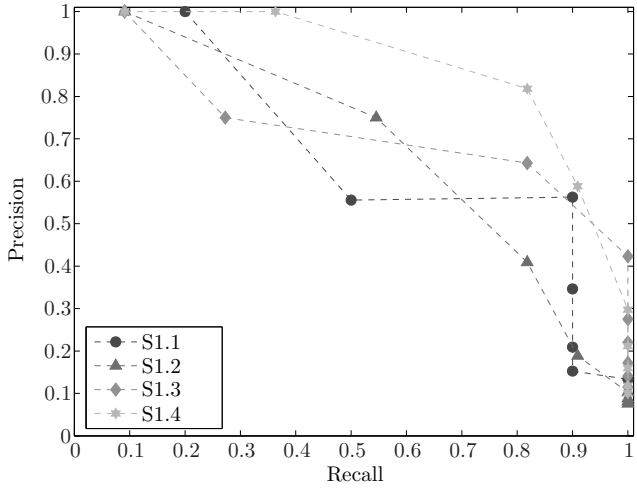


Figure 5.11. Precision-recall curves for the 'HS'-gesture from case study 1, based on the evaluation of data sets from four different test subjects (S1.1-S1.4).

For initial model validation, isolated recognition was performed on the test data based on manually added labelling information. From 258 gestures in case study 1, 254 were classified correctly, leading to a recognition rate of 98.4%. For case study 2, a recognition rate of 97.4% was reached from 784 gestures. The results indicate that the models represented the gestures well and were able to recognise the different gestures in the test set accurately.

Classification of Candidate Sections

The trained models were used to classify the candidate sections that have been retrieved in the preselection stage. Only those sections were retained, for which the recognition of preselection and classification stages agreed. Table 5.7 presents the final results of this stage for both case studies and all subjects.

The classification stage correctly recognised most of the relevant gestures that have been retrieved in the preselection stage (the average recall value was slightly reduced from 0.96 to 0.93 for case study 1 and from 0.80 to 0.79 for case study 2). The classification stage discarded many sections that have been falsely retrieved, leading to much higher precision values, especially in case of the 'HD', 'HS' and 'LB' gestures. Finally, Figure 5.13 depicts the summarised spotting results for all gestures of the case studies 1 and 2.

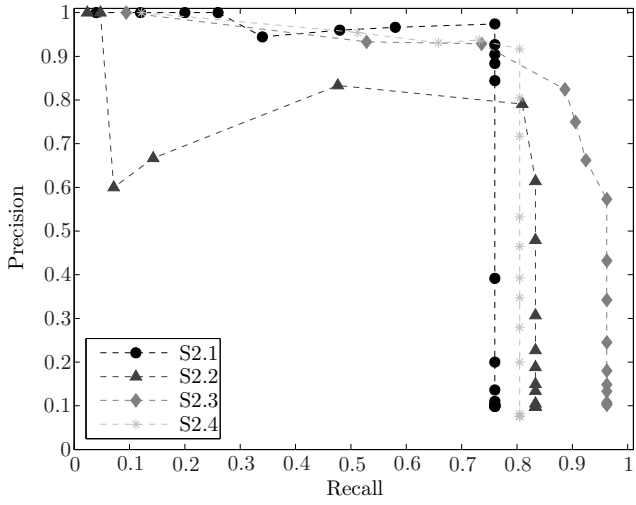


Figure 5.12. Precision-recall curves for the 'SP'-gesture from case study 2, based on the evaluation of data sets from four different test subjects (S2.1-S2.4).

5.6.3. Extensions of the classification stage

Several options exist in which our spotting approach can be extended. One possibility was to include a zero-class model in the classification stage. The modelling of the zero-class is a challenging and yet unsolved problem. We evaluated the use of two different zero-class models as extension of the classification stage. These extensions propose no viable elements of our spotting approach, but rather indicate directions of further research. The preliminary results of this investigation are shown in this section.

In case study 1 we evaluated the performance of a zero-class model that is extracted from all relevant gesture models, following the approach presented by Lee and Kim [20]. This modified classification stage yields a total recall performance of 0.81 (without threshold model: 0.93) and a total precision of 0.82 (without threshold model: 0.74). In direct comparison to the classification without the threshold model, a further increase of the precision was achieved, however at the cost of decreased recall. Figure 5.13 shows the results graphically.

In case study 2, we evaluated the spotting performance using a zero-class model that was constructed on the basis of additional gestures that were carried out by the subjects. An equal number of the gestures was used to build one additional HMM. This garbage model was included in the classification stage. The modified classification stage yielded a total recall performance

Table 5.6. Evaluation results of preselection stage.

	Case Study 1							*Case Study 2*				
	HS	**CN**	**DR**	**LB**	**PU**	**PD**	**Total**	**CL**	**DK**	**SP**	**HD**	**Total**
Relevant[a]	43	43	43	43	43	43	258	196	165	186	153	700
Retrieved[a]	159	64	90	97	63	65	473	278	199	196	310	983
Recognised[a]	41	41	41	40	42	43	248	146	138	154	125	563
Insertions[a]	118	23	49	57	21	22	290	132	61	42	185	420
Deletions[a]	2	2	2	3	1	0	10	50	27	32	28	137
Recall	0.95	0.95	0.95	0.93	0.98	1.0	0.96	0.74	0.84	0.83	0.82	0.80
Precision	0.26	0.64	0.46	0.41	0.67	0.66	0.47	0.53	0.69	0.79	0.40	0.57

[a]See Figure 5.10 for corresponding description of evaluation metrics.

Table 5.7. Spotting results after classification (2nd stage).

	Case Study 1							*Case Study 2*				
	HS	**CN**	**DR**	**LB**	**PU**	**PD**	**Total**	**CL**	**DK**	**SP**	**HD**	**Total**
Relevant[a]	43	43	43	43	43	43	258	196	165	186	153	700
Retrieved[a]	57	61	58	41	47	65	329	225	155	163	209	752
Recognised[a]	41	41	41	31	42	43	239	146	137	145	124	552
Insertions[a]	16	20	17	10	5	22	90	79	18	18	85	200
Deletions[a]	2	2	2	12	1	0	19	50	28	41	29	148
Recall	0.95	0.95	0.95	0.72	0.98	1.0	0.93	0.74	0.83	0.78	0.81	0.79
Precision	0.72	0.67	0.71	0.76	0.89	0.66	0.74	0.65	0.88	0.89	0.59	0.73

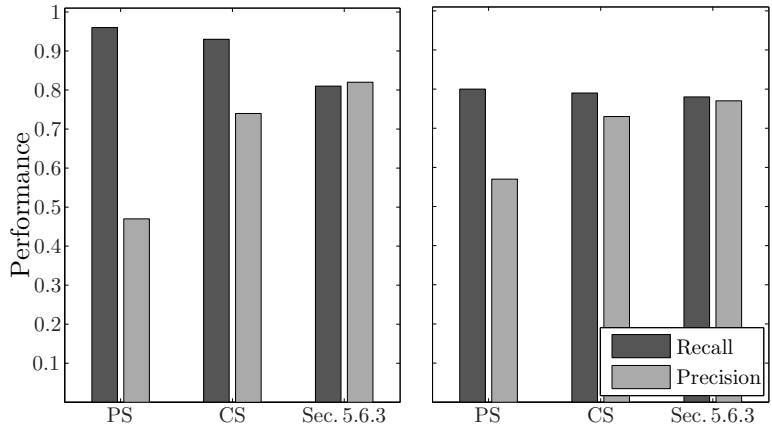
[a]See Figure 5.10 for corresponding description of evaluation metrics.

of 0.78 (without garbage model: 0.79) and a total precision of 0.77 (without garbage model: 0.73). Compared to the results of the classification without the garbage model this indicates an improvement of precision at almost constant recall.

Both concepts indicate that classification improvements with zero-class models can be achieved, however further work in this area is needed.

5.7. Discussion

Hidden Markov Models (HMMs) have proven to be applicable for recognition tasks in a variety of application domains, including gesture classification from inertial body-worn sensors. However, the spotting of gestures in a continuous data stream with HMMs is problematic due to their complexity and requirement for a zero-class. The similarity-based search in the preselection stage of our approach presents an elegant way to avoid the explicit modelling of a zero-class. In the HMM-based classification we exploit the competition of all trained models to select the most probable one. This requires that more than one gesture needs to be included in the classification stage, which can be seen



(a) Case Study 1 (b) Case Study 2

Figure 5.13. Summary of the total spotting results for the preselection (PS) and classification (CS) stages in case study 1 and 2. Additionally, the results of two extensions, discussed in Section 5.6.3, are shown.

as a limitation of our approach. However, for most applications, the spotting of several different motion events is aimed. Moreover, an explicit zero-class model can be added, when available, to improve the recognition. Initial results for two different zero-class extensions have been presented in this work.

The similarity-based search procedure used in the preselection stage permits different feature sets for individual gestures. Thus, the search can be tailored to the individual characteristic of a gesture. For example, consider game control gestures, as in [3], that are conducted in the horizontal or vertical plane only. Such gestures could be described more precisely by specific feature sets. This is an advantage over many established classification procedures, such as k-nearest neighbour classifiers or HMMs, which use the same features for all gestures to be recognised.

The section similarity search can be regarded as a natural extension of the frequently used sliding window approach for motion and activity detection, as e.g. in [34]. We introduced a size-variable search window to accommodate for the variability in the length of gestures and used a dynamic step size given by the segmentation points. While the trivial sliding window was mainly used for the detection of repetitive motions, such as hammering, the approach presented in this work was successfully evaluated for non-cyclic motion events in the two case studies.

The problem of human gesture recognition depends largely on the application domain: In contrast to artificial gestures used for human-computer control or repetitive motions in very specific activities, natural motions in activities of

daily living are more challenging to spot. This is due to the fact that control gestures can be constructed to provide strong discrimination, that is typically not the case for gestures being part of activities of daily living. Hence, such gestures contain more intra- and inter-person variability, making the spotting task more challenging. However, the presented results indicate, that our spotting procedure performs well for these types of gestures.

In a related work of the authors, one-hand gestures, specifically constructed for game control, were investigated [3]. The approach in that work differed from the current work in firstly, raw inertial sensor signals were used from a sensor attached to a glove at the hand and secondly, the gestures were designed to aim at discrimination and detection in a gaming scenario. In contrast, the current work aimed at recognising natural everyday life gestures with large fluctuations in length and execution from using sensor data from the lower and upper arm. Consequently, with the use of HMMs, a more complex approach was deployed in the current work to achieve the recognition.

The focus of the current work was to analyse the recognition performance using person-specific training. The case studies were designed to incorporate additional motions and gestures and maintain a low share of relevant gestures: 25.4% in case study 1 and 34.7% in case study 2. Both case studies evaluated four subjects each. An initial insight into the subject-specific variability was obtained from reviewing the precision-recall curves. However, a larger number of users should be evaluated in future works, to study the fluctuation in recognition performance and investigate non-personalised detection models in more depth.

The temporal phases of a gestures are onset, core and conclusion. Typically, onset and conclusion are variable transfer states between consecutive gestures. However, the core part is specific for a gesture. In the evaluated case studies most of the gestures were acquired with a defined start and ending position, but all contained a core part. For example, in the phone pickup gesture, the user's hand moved towards the receiver, picked it up and moved the receiver to the ear. While the movement may commence with the hand at an arbitrary position, the core is preserved in order to successfully complete the activity. The motion segments in the core phase and during the transitions involve direction changes in the segmentation signal. In our approach, segmentation points were created at these positions. Based on the preselection feature set, the section similarity search was used to test for gestures cores at every segmentation point. Hence, we expect that by using the segmentation and search procedure, gestures embedded in arbitrary transitions can be detected.

Looking at the individual results of case study 1 and case study 2, we observe lower spotting performance for those gestures included in case study 2. We assume that this is due to higher intra-person variability of those gestures. More specifically, we observed the following additional challenges for the spotting of gestures: 1) differences in the size and consistency of food pieces, 2) additional degrees of freedom produced by the tools used for the food intake

and 3) temporal aspects, such as the temperature change of the food and the natural satiety of the subject developed during the intake session. To overcome potential weaknesses in the spotting of gestures related to food-intake, we argue that the recognition of such gestures can be enhanced by combining different sensing modalities to develop a dietary monitoring system [2].

We expect that the presented spotting approach can be applied to other types of motion events. At the implementation level an appropriate motion parameter must be selected. This motion parameter shall describe the major properties of the motion event and lead to a reproducible and distinctive motion segmentation. We believe that this can be achieved for many applications.

5.8. Conclusion and Outlook

We conclude that our spotting scheme based on the concept of motion segments is a feasible strategy for the identification of motion events in a continuous signal stream. We demonstrated that our approach works well for arm-based motions, that are particularly difficult to recognise due to the inherent complexity of arm motions. Moreover, we have shown that our approach simplifies the rejection of non-relevant gestures. We argue that our method is likely to facilitate a wide range of real-life applications of context and activity recognition.

Bibliography

[1] O. Amft, H. Junker, and G. Tröster. Detection of eating and drinking arm gestures using inertial body-worn sensors. In B. Rhodes and K. Mase, ed., *ISWC 2005: IEEE Proceedings of the Ninth International Symposium on Wearable Computers.*, pp. 160–163. IEEE Press, October 2005. doi:10.1109/iswc.2005.17.

[2] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of chewing sounds for dietary monitoring. In M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, ed., *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing*, vol. 3660 of *Lecture Notes in Computer Science*, pp. 56–72. Springer Berlin, Heidelberg, September 2005. doi:10.1007/11551201_4.

[3] D. Bannach, O. Amft, K. S. Kunze, E. A. Heinz, G. Tröster, and P. Lukowicz. Waving real hand gestures recorded by wearable motion sensors to a virtual car and driver in a mixed-reality parking game. In A. Blair, S.-B. Cho, and S. M. Lucas, ed., *CIG 2007: Proceedings of the 2nd IEEE Symposium on Computational Intelligence and Games*, pp. 32–39. IEEE Press, April 2007. doi:10.1109/cig.2007.368076.

[4] L. Bao. Physical activity recognition from acceleration data under semi-naturalistic conditions. Master's thesis, Massachusetts Institute of Technology, Boston, USA, 2003.

[5] A. Y. Benbasat. An inertial measurement unit for user interfaces. Master's thesis, Massachusetts Institute of Technology, Boston, USA, September 2000.

[6] A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Philos T Roy Soc B*, 352(1358):1257–1265, August 1997.

[7] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR 1997: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 994–999, June 1997.

[8] H. Brashear, T. Starner, P. Lukowicz, and H. Junker. Using multiple sensors for mobile sign language recognition. In *ISWC2003: Proceedings of the Seventh IEEE International Symposium on Wearable Computers*, pp. 45–52, Oct. 2003. doi:10.1109/iswc.2003.1241392.

[9] O. Cakmakci, J. Coutaz, K. Van Laerhoven, and H. Gellersen. Context awareness in systems with limited resources. In *ECAI 2002: Proceedings of the third workshop on Artificial Intelligence in Mobile Systems (AIMS)*, 2002.

[10] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. In *Proceedings of the Fifth International Conference on Computer Vision*, pp. 624–630, June 1995. doi:10.1109/iccv.1995.466880.

[11] S. Chambers, S. Venkatesh, G. West, and H. Bui. Hierarchical recognition of intentional human gestures for sports video annotation. In R. Kasturi, D. Laurendeau, and C. Suen, ed., *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 2, pp. 1082–1085. IEEE Press, August 2002. doi:10.1109/icpr.2002.1048493.

[12] J. Deng and H. Tsui. An hmm-based approach for gesture segmentation and recognition. In *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 2, pp. 679 – 682, September 2000.

[13] W. Gao, J. Ma, J. Wu, and C. Wang. Sign language recognition based on HMM/ANN/DP. *Int J Pat Rec Artif Intelli*, 14(5):587–602, 2000.

[14] H. Junker, P. Lukowicz, and G. Tröster. Locomotion analysis using a simple feature derived from force sensitive resistors. In *Proceedings of the 2nd International Conference on Biomedical Engineering*, 2004.

[15] K. Kahol, K. Tripathi, and S. Panchanathan. Documenting motion sequences with a personalized annotation system. *IEEE Multimedia*, 13(1):37–45, Jan-March 2006. doi:10.1109/mmul.2006.5.

[16] K. Kahol, P. Tripathi, S. Panchanathan, and T. Rikakis. Gesture segmentation in complex motion sequences. In *ICIP2003: Proceedings of the International Conference on Image Processing*, vol. 2, pp. 105–108, September 2003.

[17] E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time series. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 289–296, 2001.

[18] N. Kern, B. Schiele, and A. Schmidt. Multi-sensor activity context detection for wearable computing. In *Proceedings of the European Symposium on Ambient Intelligence*, pp. 220–232, Eindhoven, The Netherlands, November 2003.

[19] C. Lee and X. Yangsheng. Online, interactive learning of gestures for human/robot interfaces. In N. Caplan and C. G. Lee, ed., *ICRA 1996: Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 4 of *IEEE Robotics and Automation Society*, pp. 2982–2987. IEEE Press, April 1996.

[20] H.-K. Lee and J. H. Kim. An HMM-based threshold model approach for gesture recognition. *IEEE T Pattern Anal*, 21(10):961–973, October 1999.

[21] J. C. Lementec and P. Bajcsy. Recognition of arm gestures using multiple orientation sensors: Gesture classsification. In *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems*, pp. 965–970, October 2004.

[22] R.-H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 558 – 567. IEEE Computer Society, April 1998.

[23] P. Lukowicz, J. A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner. Recognizing workshop activity using body worn microphones and accelerometers. In *Pervasive 2004: Proceedings of the International Conference on Pervasive Computing*, vol. 3001 of *Lecture Notes in Computer Science*, pp. 18–32. Springer, April 2004. ISBN 3-540-21835-1. doi:10.1007/b96922.

[24] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Comput Vis Image Und*, 81(3):231–268, 2001.

[25] P. Morguet. *Stochastic Modeling of Image Sequences for the Segmentation and Recognition of Dynamic Gestures*. PhD thesis, Technische Universität München, December 2000.

[26] G. Ogris, T. Stiefmeier, H. Junker, P. Lukowicz, and G. Troster. Using ultrasonic hand tracking to augment motion analysis based recognition of manipulative gestures. In B. Rhodes and K. Mase, ed., *ISWC 2005: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*, pp. 152–159. IEEE Press, October 2005. doi: 10.1109/iswc.2005.54.

[27] D. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. In B. Rhodes and K. Mase, ed., *ISWC 2005: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*, pp. 44–51. IEEE Press, October 2005. doi:10.1109/iswc.2005.22.

[28] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *P IEEE*, 77(2):257 – 286, Feb. 1989.

[29] C. Rao and M. Shah. View-invariance in action recognition. In *CVPR 2001: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 316–322, 2001.

[30] M. Shah and R. Jain. *Motion-Based Recognition*. Kluwer Academic Publishers, 1997.

[31] M. Stäger, P. Lukowicz, N. Perera, T. von Büren, G. Tröster, and T. Starner. Soundbutton: Design of a low power wearable audio classification system. In *ISWC 2003: Proceedings of the 7th IEEE International Symposium on Wearable Computers*, pp. 12–17, Oct 2003. doi:10.1109/iswc.2003.1241387.

[32] T. Starner. Visual recognition of american sign language using hidden markov models. Master's thesis, Massachusetts Institute of Technology, Boston, USA, February 1995.

[33] T. S. Wang, Y. Shum, Y. Xu, and N. Zheng. Unsupervised analysis of human gestures. In *IEEE Pacific Rim Conference on Multimedia*, pp. 174–181, 2001.

[34] J. Ward, P. Lukowicz, G. Tröster, and T. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE T Pattern Anal*, 28(10): 1553–1567, Oct. 2006. doi:10.1109/tpami.2006.197.

[35] J. A. Ward, P. Lukowicz, and G. Tröster. Gesture spotting using wrist worn microphone and 3-axis accelerometer. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pp. 99–104, Oct 2005.

[36] Y. Wu and T. Huang. Vision-based gesture recognition: A review. In *Proceedings of the International Gesture Workshop, France*, 1999.

[37] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *CVPR 1992: Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 379–385, 1992.

6

Analysis of chewing sounds for dietary monitoring

Oliver Amft, Mathias Stäger, Paul Lukowicz and Gerhard Tröster

UBICOMP 2005: Proceedings of the 7th International Conference on Ubiquitous Computing, LNCS Vol. 3660, 56–72, Springer Berlin, Heidelberg, 2005.
DOI: 10.1007/11551201_4

Abstract

The paper reports the results of the first stage of our work on an automatic dietary monitoring system. The work is part of a large European project on using ubiquitous systems to support healthy lifestyle and cardiovascular disease prevention. We demonstrate that sound from the user's mouth can be used to detect that he/she is eating. The paper also shows how different kinds of food can be recognised by analysing chewing sounds. The sounds are acquired with a microphone located inside the ear canal. This is an unobtrusive location widely accepted in other applications (hearing aids,

headsets). To validate our method we present experimental results containing 3500 seconds of chewing data from four subjects on four different food types typically found in a meal. Up to 99% accuracy is achieved on eating recognition and between 80% to 100% on food type classification.

6.1. Introduction

Healthy lifestyle and disease prevention are a major concern for large portions of the population. Considering the worrying trend of sky-rocketing health care costs and the ageing population, these are not just personal but also important socio-economic issues. As a consequence all concerned parties: individuals, health insurance and governments are willing to spend considerable resources on tools that help people develop and maintain healthy habits. In Europe a considerable portion of research funding in this area is directed at mobile and ubiquitous computing technology. Within this program our group is involved in the 34 Million Euro MyHeart project that includes 35 medical, design, textile and electronics related research institutions and companies.

The aim of the consortium is to develop schemes that combine long term physiological monitoring and behavioural analysis with a personalised direct or professional-observed feedback to help users reduce their risk of cardiovascular disease. As is well known, the three main aspects that need to be addressed are stress, exercise and diet. In the project our group focuses on the later. Our aim is to develop wearable sensing technology to aid the user in monitoring his eating habits. In this paper we report on results of the first stage of this work: using wearable microphones to detect and classify chewing sounds (called mastication sounds) from the user's mouth.

6.1.1. Dietary monitoring

Dietary monitoring includes a variety of factors starting from the diet composition to frequency, duration and speed of eating, all of which can be relevant health issues. Today such monitoring is almost entirely done 'manually' by user questionnaires. Electronic devices are at best used as intelligent log books that can derive long term trends, calculate calories from entered data and give simple user recommendations. The collection and entry of the data has to be done by the user which involves considerable effort. As a consequence, as anyone who has ever attempted a diet knows, compliance tends to be very poor.

Since prevention involves the adaptation of a healthier lifestyle, long term, quasi permanent monitoring (months or years) is needed to really make an impact on the risk of cardiovascular diseases. Thus any, even very rudimentary, tool that reduce the effort and interaction involved in data collection and entry could make a big difference.

6.1.2. Automating dietary monitoring

The ultimate goal of a system that precisely and 100% reliably determines the type and amount of all and any food that the user has consumed is certainly more of a dream then a realistic concept. However, we believe that with a combination of wearable sensors and a degree of environmental augmentation useful assistive systems are conceivable. On one hand, such systems could provide a rough estimate on the food consumption much like many today's physical activity monitoring devices provide only a rough guess of the caloric expenditure. On the other hand, it could be used as an entry assistant that, at the end of the day, would present the user with its best guess of when, how much, and what he has eaten and ask him to correct the errors and fill the gaps.

Overall we imagine such a non-invasive dietary monitoring support system to rely on the following three components:

1. Monitoring of food intake through appropriate wearable sensors. The main possibilities are
   (a) detecting and analysing chewing sounds,
   (b) using electrodes mounted on the base of the neck (e.g in a collar) to detect and analyse bolus swallowing,
   (c) using motion sensors on hands to detect food intake related motions.
2. Monitoring food preparation/purchase through appropriate environmental augmentation. Here, approaches such as using RFID-tags to recognise food components or communicating with the restaurant computer to get a description and nutrition facts of the order are conceivable.
3. Including user habits and high level context detection as additional information sources. Here, one could accentuate the fact that eating habits tend to be associated with locations, times and other activities. Thus information on location (e.g in the dining room sitting at the table), time of day, other activity (unlikely to eat while jogging) etc. provide useful hints.

6.1.3. Paper contributions

In the paper we concentrate on the first component of the envisioned system: food intake detection. Specifically, we consider the detection and classification of chewing sounds. To this end the paper presents the following results:

1. We show that good quality chewing sound signal can be obtained from a microphone placed in the ear canal. Since much of the acoustic signal generated by mechanical interaction of teeth and food during occlusion is transmitted by bone conduction, these sounds are actually much stronger

than the speech signal. At the same time the location is unobtrusive and proved acceptable in applications such as hearing aids or recent high end mobile phone headsets.

2. We show that chewing sequences can be discriminated from a signal containing a mixture of speech, silence and chewing.

3. We present a method that detects the beginning of single chews in a chewing sequence.

4. We show that chewing sound based discrimination between different kinds of food is possible with a high accuracy.

For the above methods we present an experimental evaluation with a set of four different food products selected to represent different categories of food that might be present in a meal. The experiments consists of a total of 650 chewing sequences, from 4 subjects that amount to a total of 3500 seconds of labelled data. We show that recognition rates of up to 99% can be achieved for the chewing segment identification and of between 80 and 100% for the food recognition.

Overall, while much still remains to be done, our work proves the feasibility of using chewing sound analysis as an important component in a diet monitoring system. An important aspect of our contribution is the fact that the type of information derived by our system (what has actually been eaten) is very difficult to derive using other means.

6.1.4. Related work

Activities of daily living are of central interest for high-level context-aware computing. Information acquisition can be realised by distributing sensors in the environment and on the human body. Realisation of intelligent environments have been studied, e.g. in the context of smart homes [16] and mobile devices [9]. These works are generally focused on enhancing the quality of life, e.g. for independent living [15, 18]. Smart identification systems have also been developed [21] which may provide information associated to nutrition phases, e.g. smart cups [2].

The interaction of chewing, acoustic sensation and perception of textures in food has been studied intensively in food science. Work in this area has been dedicated mainly to the relation of chewing sounds on the sensation of crispness and crunchiness. This was done by investigating air-conducted noises produced during chewing [25, 27] or by instrumental monitoring of the deformation under force [5, 7, 11, 12] and studying correlation with sensory perception [23, 26]. The loudness of a foodstuff during deformation depends mainly on the inner structure, i.e. cell arrangement, impurities and existing cracks [1]. Wet cellular materials, e.g. apples and lettuce, are termed wet crisp since the cell structures

contain fluids whereas dry crisp products, e.g. potato chips have air inclusions [8]. A general force deflection model has been proposed [28] interpreting the acoustic emissions as micro-events of fracture in brittle materials under compression.

Initially Drake [7] studied the chewing sound signal in humans when chewing crisp and hard food products. It was found that a normal chewing cycle after bringing the food piece to the mouth cavity can be partitioned into two adjacent phases: Gross cutting the ingested material and conversion in fine grained particles. This process is understood as a gradually decomposition of the material structure during chewing and is audible as a decline of the sound level [7]. A swallowable bolus is formed after a certain level of lubrication and particle size has been reached. A first attempt was made by DeBelie [6] to discriminate two classes of crispness in apples by analysing principal components in the sound spectrum of the initial bite.

Originating from the pioneering work on the auscultation of the masticatory system (system related to chewing) done by Brenman [3] and Watt [29] the stability of occlusion and has been assessed in the field of oral rehabilitation by analysing teeth contact sounds (gnathosonic analysis) [19]. Similarly the sounds produced by the temporomandibular joints during jaw opening and closing movements have been studied regarding joint dysfunction [31]. It is not expected that these sound sources provide a audible contribution to chewing of food materials in healthy subjects. However, these studies provide information regarding sound transducer types and mounting position that may be usable also for the analysis of chewing sounds. Recent investigations [19, 24] evaluated measurement methodology, applicable transducers types and positions.

6.2. Methodology

This section will give an overview of our approach. It is important to note that, as described in the introduction, we consider the sound analysis to be just one part of a larger dietary monitoring system. This means that sound analysis is not meant to solve the entire dietary monitoring problem by itself. Instead the goal of our work is to demonstrate that a significant amount of useful information *that is difficult to obtain through other means* can be extracted from chewing sound analysis. Furthermore, the question how it can be expected to interact with other context information is an important research question pursued by our group (although it is not the focus of this paper).

6.2.1. Approach

Nutrition intake can be coarsely divided into three phases: fracturing (tearing) the food mainly with the incisors, chewing of the pieces and swallowing of the bolus. Ultimately, all three phases should be analysed since the bolus formation process differs for characteristic food materials [10], e.g. a dry potato chip

differs in structure, fluid compartments and chewing from cooked pasta. Initial bites may have more distinctive properties [6], but occur less often and are not available for all food types. A combination of fracture sound and bolus production process features may permit the acoustic detection of food products.

In this paper, we concentrate on the longest phase. Therefore we have chosen to analyse the sound of normal chewing cycles, i.e. beginning after intake of the food piece up to and excluding swallowing of the bolus. We stopped with analysing the sound when the amplitude level decayed to approximately 5dB above the noise level.

Fig. 6.1 illustrates the overall structure of our approach. It consists of three main steps: signal acquisition, chewing segment identification and food type classification.

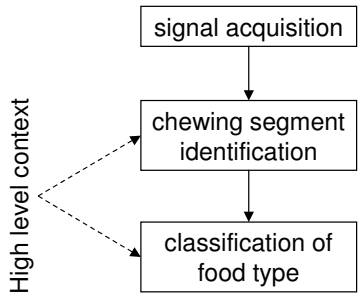


Figure 6.1. Approach to the analysis of chewing sounds

The challenge of signal acquisition is to identify a microphone position that combines good amplitude levels for the chewing sounds, with good suppression of other sounds at a location that is comfortable and socially acceptable to the user.

For chewing segment identification this paper considers only sound-related means. In particular, we investigate a classifier that can distinguish between a broad range of chewing sound and various speech/conversation sounds. In a wearable computing environment, other means are possible. E.g., food intake is usually accompanied by moving the arm up and bringing the hand close to the users mouth. The lower arm is then pointing away from the earths centre of gravity; something which can easily detected by an accelerometer mounted on the users wrist. However, the user can perform similar movements for other activities (e.g. scratching his chin) so other information from sensors in the environment might be needed (e.g. location information that the user is in the kitchen or the dining room).

Once a segment is classified as being a chewing sound, the type of food needs to be identified. Again, we focus on the audio analysis of the chewing sound. In doing so, we do not aim to be able to pick any of the thousands of possible food types. This would clearly be unrealistic. Instead we assume

(1) that we have a certain prior knowledge about the type of foods that are relevant to the particular situation and (2) that often it is sufficient to just be able to identify a general type of food or be able to say "could have been XY". The first assumption is not as far fetched as it might sound. The intelligent refrigerator/cardboard that knows what food is inside and what has been taken out (e.g. through RFID) is the prototypical ubiquitous application. In a restaurant credit card information or an electronic menu could be used to constrain the number of possibilities. Additionally, people have certain fairly predictable eating habits. The second point relates to the type of application that is required. As stated in the introduction, the system does not need be fully automated to be useful and to be an improvement over current 'manual' monitoring. Thus it is perfectly sufficient if at the end of the day the system can remind the user that for example "at lunch you had something wet and crisp (could have been salad) and some soft texture stuff (spaghetti or potatoes)" and asks him to fill in the details. From the above considerations we concentrate our initial work on being able to distinguish between a small set of predefined foods and on the distinction between certain food classes.

6.2.2. Experiments

The evaluation of all methods described in the remainder of the paper has been performed using the following experimental setup.

Test subjects: Four subjects (2 female, 2 male, mean age 29 years) were instructed to eat different food products normally, with the mouth closed during chewing. In this way the chewing phase of the nutrition cycle is covered: Beginning after intake of the food piece up to swallowing of the bolus (see Sec. 6.2.1).

By restricting our experiments to the chewing phase, we ensure that the recognition works solely on chewing. Specifically, we exclude swallowing and tearing sounds since these phases have different acoustic characteristics. Fracturing (tearing) and swallowing sounds are regarded as additional source of information and may be analysed independently. Since these events are not occurring at the same high frequency than chewing, they are considered less relevant.

The subjects had no denture, no acute teeth or facial pain and no known history of occlusion or temporomandibular joint dysfunction. Furthermore none of the subjects expressed a strong dislike of any food product in this study.

Test objects: The food products shown in Table 6.1 have been selected since they imitate typical components in a meal or daily nutrition. The food groups reflect the acoustic behaviour during chewing and not their nutrition value. They can be simply reproduced with a high fidelity. Furthermore some of the crisp-classified products have been referenced in texture studies before: Potato

chips [28] and apples [6]. Beside the dry-crisp and wet-crisp categories, a third acoustic group of "soft texture" foods have been included: Cooked pasta and cooked rice.

Table 6.1. Details for the food products and categorisation

Food product	Food group	Product/Ingredients/Preparation
Potato chips	dry-crisp	Zweifel, potato chips (approx. 3cm in diameter)
Apple	wet-crisp	type "Jonagold" and "Gala" washed, cut in pieces, with skin
Mixed lettuce	wet-crisp	endive, sugar loaf, frisée, raddichio, chicory, arugula
Pasta	"soft texture"	spaghetti (al dente)
Rice	"soft texture"	rice without skin

Initial evaluation of the sound data showed that the rice recordings were smallest in amplitude of all recorded foods. The potato chips produced the highest amplitude for all subjects. Fig. 6.5 illustrates a typical waveforms recorded for apples.

Table 6.2 depicts the inspected sound durations for the food products from all subjects. The number of single chews is the number given by the single chew detection algorithm explained in Sec. 6.5.1. The single chews per chewing sequence reflects the authors' experience that usually potato chips are destruct with only a few chews, whereas pasta or lettuce require several chews to masticate properly.

Table 6.2. Statistics of the acquired and inspected sound database for all food products

Food product	Time recorded and inspected	No. of chewing sequences	Detected chews	Chews per sequence
Potato chips	677 sec	179	979	5.5
Apple	1226 sec	245	1538	6.3
Mixed lettuce	1054 sec	152	1691	11.1
Pasta	630 sec	74	1290	17.4
Rice[a]	240 sec	-	-	-
Total	3827 sec	650	5498	

[a]Omitted because of small amplitude, see Sec. 6.4

Test procedure: A electret condenser microphone (Type Sony ECM-C115) was placed in the ear canal as described in Sec. 6.3. After positioning, the microphone fixation was checked to avoid interference between movements of the jaw and the microphone in the ear canal. A second microphone of the same type was used at collar level, at the side of the instrumented ear, as reference to detect possible environmental sounds during inspection. The waveforms were recorded at a sampling frequency of 44.1 kHz, 16 bit resolution.

All products were served on a plate. Cutlery was used for the mixed lettuce, pasta and rice. Subjects were instructed to take pieces, small enough to be ingested and chewed at once, as described above. The temperature of pasta and rice was cold enough to allow normal chewing.

6.3. Positioning of the microphone

Sound produced during the masticatory process can be detected by air- and bone-conduction. Frequency analysis of air-conduced sounds from chewed potato chips showed spectral energy between zero and 10 kHz [12] although the frequency range with highest amplitude for various crisp products are in the range of 1 kHz−2 kHz [4]. Bone-conducted sounds are transmitted through the mandibular bones to the inner ear. The soft tissue of mouth and jaw damp high frequencies and amplify at the resonance frequency of the mandible (160 Hz) when chewed with closed mouth [11].

Condenser or dynamic microphone transducers have been used in texture studies literature at various places with the goal to detect and reproduce human perception. Mainly the following positions were evaluated: In front of the mouth [7, 12], at the outer ear above the ear canal [26], a few centimetres in front of the ear canal opening [5], pressed against the cheek [5, 7] or placed over the ear canal opening [6, 7]. Gnathosonic studies used a stereo-stethoscope technique [29] and microphones [11] at the forehead or over the zygoma [30]. More recently a method using head-phones with the microphones positioned over the ear canal opening has been proposed [19].

Several positions for the microphone have been evaluated for this study as indicated in Table 6.3. This list includes some of the positions used in previous work. The evaluation of ubiquitous positions, not hindering the user's perception was emphasised. To this end, positions 1, 5 and 6 are favourable because their implementation can be hidden in human anatomy or in cloths.

Potential artifacts introduced by daily use could interfere significantly with the microphone function. This may affect position 5 since it has the disadvantage of being hidden under cloths or disturbed by cloth sounds. Position 1 has the advantage of being less affected by loud environmental noises since it is embedded directly into the ear canal: With a directional microphone oriented towards the eardrum, the intensity of any noise from the environment is reduced.

Table 6.3. Evaluated microphone positions

Microphone	Position
1	Inner ear, directed towards eardrum (Hearing aid position)
2	2cm in front of mouth (Headset microphone position)
3	At cheek (Headset position)
4	5cm in front of ear canal opening (Reference position for audible chewing sounds)
5	Collar (Collar microphone position)
6	Behind outer ear (Hidden by the outer ear, used by older hearing aid models)

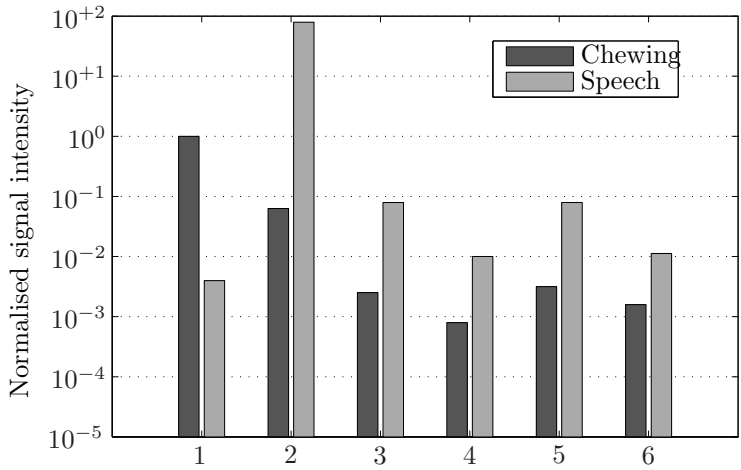


Figure 6.2. Signal intensity of different microphone positions (see Tab. 6.3)

The position of the microphone was evaluated while a subject was chewing potato chips and while the subject was speaking. The mean amplitude perceived at position 1 was used as reference for normalisation. Fig. 6.2 depicts the relation of the signal amplitude intensity shown on a logarithmic scale. It can be seen clearly that position 1 not only has the highest intensity for chewing sounds but it is also the only position with chewing sound intensity higher than speech intensity. Therefore for all further measurements position 1 was used.

A microphone at position 1 does not need to hinder the person, as modern hearing aids prove. Applicable microphones could be very small and combined with an earphone be used for other applications, e.g. mobile phones. For example, modern hearing aids already operate with a combined microphone/earphone.

6.4. Chewing segment identification

The identification of chewing segments in a continuous sound signal can be regarded as a base functionality and hence is of high importance for the detailed analysis of the masticated food type. We see mainly two different methods based on audio signal processing.

A: Intensity of audio signal:

In an environment, like a living room, with background music playing or in a quiet restaurant, the chewing sound picked up in the inner ear is much louder than a normal conversation or background music. This is indicated in the sample recording shown in Fig. 6.3.

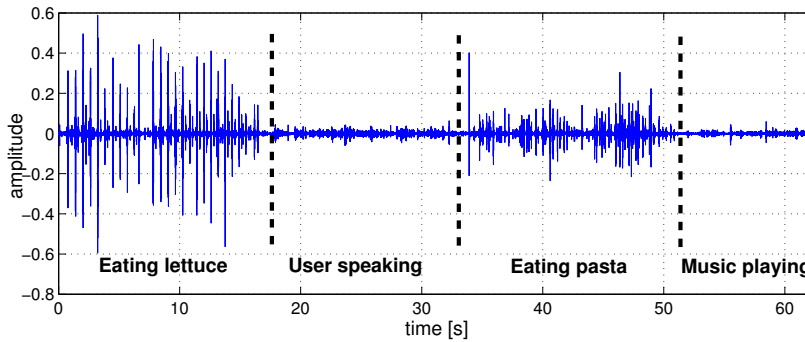


Figure 6.3. Chewing sound and speech recording in a room with background music

B: Chewing sound - speech classifier:

Despite the general suppression of the speech signal, loud speech can at times develop amplitude peaks similar to chewing signals. Therefore it is necessary to be able to separate these two classes. This is achieved by calculating audio features from a short signal segment of length t_w, averaging the features over N_{avg} segments and then finally classifying them with a previously trained classifier [22].

Features: We used features that are popular in the area of speech, audio and auditory scene recognition [13, 14, 17]. In the temporal domain, those were

zero-crossing rate and fluctuation of amplitude. Frequency domain features were evaluated based on a 512-point Fast Fourier Transformation (FFT) using a Hanning window. Here, the features included: frequency centroid, spectral roll-off point with the threshold of 0.93, fluctuation of spectrum and band energy ratio in 4 logarithmically divided sub-bands. Additionally 6 cepstral coefficients (CEP) were evaluated. Both time and frequency domain features were evaluated on a window of $t_w = 11.6$ ms. No overlap between the windows was used.

The features were averaged over N_{avg} windows to improve the recognition results. This method helped to bridge pause gaps between the chewing sounds. These gaps vary between 100 ms and 600 ms depending on the chewed material and the progression of decomposition (see Fig. 6.5). Longer pauses may be observed at the beginning of a chewing sequence for larger food pieces as well as before and after partial bolus swallowing.

Classifiers: A C4.5 decision tree classifier from the Weka Toolkit [32] was trained with the aforementioned features. The classifier was 10-fold cross-validated on a two class data set. The first class contained all food products as specified in Table 6.2 except cooked rice. Rice was excluded since individual classification of food products against speech signals showed weak results for rice. This was expected from the low signal-noise ratio of the rice sounds. The second class included various speech signal segments from several speakers as well as conversation of test subjects and the authors.

Since the accuracy of a classifier depends on the class distribution, the ROC curve (Receiver Operating Characteristic) is presented instead (see Fig. 6.4). ROC curves help to visualise classifier performance over the whole range of frequency of occurrence [20]; the best classifier is the one to the top-left corner. This is useful in our case since the number of occurrences of speech and chewing sounds may vary and may not be known beforehand. Clearly, the classifier that uses the CEP features dominates. This was expected since the CEP features help to pick out speech sequences. Furthermore, the number N_{avg} of averaging frames was varied. We found that the highest recognition rates can be achieved if N_{avg} is chosen so that the features are at least averaged over one single chew which takes about one second. In our case this occurs if $N_{avg} > 1 \text{ sec}/t_w = 86.2$.

6.5. Discrimination of foods products

6.5.1. Isolation of single chews

First trials in separating different food products with the same methods as in the previous section (i.e. calculating features over a large window) produced recognition rates around 60%. The reason for this is mainly due to the rather long pause between single chews, which produces the same audio signature for all food items.

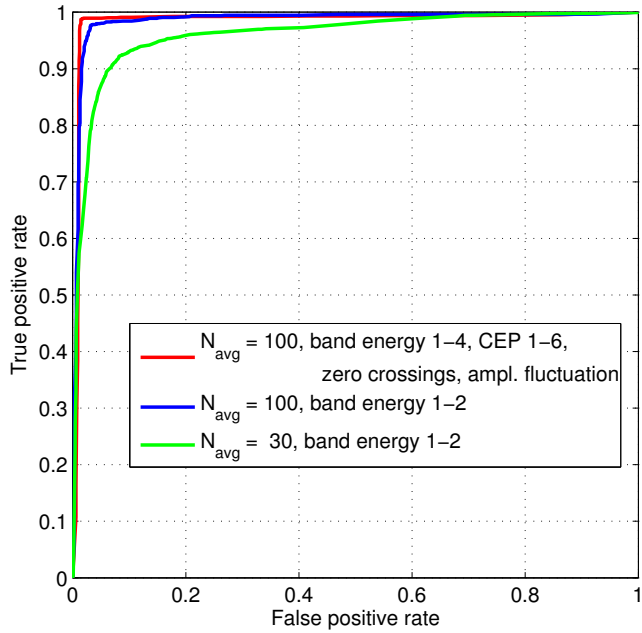


Figure 6.4. ROC curve for chewing sounds (positives) and speech sounds (negatives)

To overcome this problem we have looked in more detail at the temporal structure of a typical chewing sequence (see Fig. 6.5). It can be seen that the audio signal of one chew is mainly composed of four phases: The closing of the mandible to crush the material, a small pause, the opening of the mandible in which material that stick to the upper and lower teeth is uncompressed, and again a pause. The timing between those phases is given mainly by the mechanical properties of the food and the physical limitations of the mandible. All test subject showed almost the same timing for the same food, with the exception of a longer or shorter pause in phase 4 (fast/slow eater). The four phases are very well distinguishable in crispy food, in softer food like pasta the phases tend to merge. Still, the pause in phase 4 and the increase in amplitude at the beginning of phase 1 remain.

A relatively simple algorithms helps us the detect the beginning of each chew. The short-time signal energy in a 20 ms window is compared to a energy threshold and the resulting signal is set to 1 if the short-time signal energy is larger than the threshold and to 0 otherwise. The resulting signal is low-pass filtered with a 4th order butterworth filter. We found that a filter with a 3dB cut-off frequency of 4 to 5 Hz reliably responds to the pause in phase 4

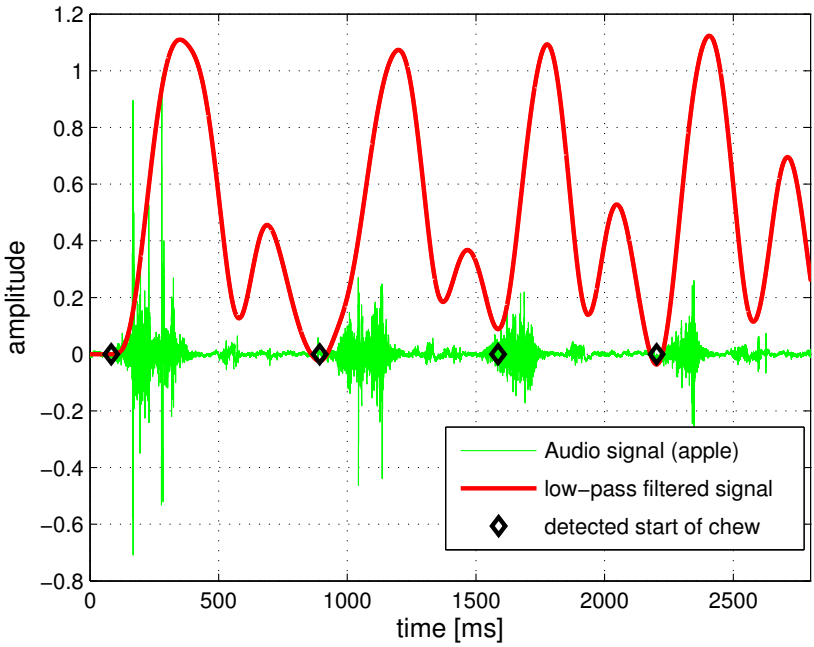


Figure 6.5. Sample sound signal observed for chewing an apple

while filtering out the shorter pause in phase 2. With help of the hill climbing algorithm the beginning of each chew is detected as shown in Fig. 6.5. We found that this algorithm can detect the start point of about 90% of all chews while producing only very little insertions.

6.5.2. Classification

Once the audio signal is segmented into single chews, the segments are classified using the same procedure as in Sec. 6.4. Several features were applied to a short window that was consecutively shifted. We found that a 11.6 ms window with a shift of 8.7 ms works best for our sound classes. The most promising features were: zero crossing rate, band energy ratios, fluctuation of amplitude, fluctuation of spectrum and bandwidth. The features were further averaged over the length of a single chew. The length of a single chew was used as an additional feature and helped to improve the recognition rate of especially the pasta, since soft-texture foods have shorter durations of chews. The features were then 10-fold cross-validated with a C4.5 decision tree classifier. Recognition rates range around 66% to 86% and the corresponding confusion matrix is listed in Table 6.4.

Since the material inside the mouth can not change between single chews,

Table 6.4. Confusion matrix for single chews

a	b	c	d	← classified as	Accuracy
669	170	25	115	a = Chips	68.34%
183	1024	41	290	b = Apple	66.58%
25	39	1112	114	c = Pasta	86.20%
125	293	95	1178	d = Lettuce	69.66%

a majority decision over a whole chewing cycle was performed. This measure resulted in an increase of recognition rate of 15 to 20% as shown in Table 6.5. It can be seen that there is some confusion between apple and lettuce which can be explained by them belonging into the same food category (see Table 6.1) and therefore having similar mechanical properties.

Table 6.5. Confusion matrix for chewing cycles

a	b	c	d	← classified as	Accuracy
156	12	1	10	a = Chips	87.15%
24	198	1	22	b = Apple	80.82%
0	0	74	0	c = Pasta	100.00%
4	21	0	127	d = Lettuce	83.55%

6.6. Conclusion and future work

6.6.1. Conclusion

The work presented in this paper has proven that chewing sound analysis is a valuable component for automated dietary monitoring systems. Specifically we have shown that:

1. A microphone location inside the ear can acquire good quality chewing sounds while suppressing many other sounds originating inside the oral cavity such as speech. At the same time it is a location that has been proven to be acceptable to users in other applications (e.g. hearing aids, headsets). Applicable microphones could be very small, not hindering the normal perception. Moreover, a combination of microphone and earphone for shared use with other applications, e.g. a mobile phone, could be employed.

2. Chewing sounds can be reliably separated from the main sound source inside the mouth cavity: speech.

3. Individual chews can be isolated and partitioned into phases with a simple low pass filter based algorithm

4. Audio analysis can be used to distinguish between a small predefined set of different food types as for example found in a single meal.

The food groups introduced in the experiments reflect the acoustic behaviour during chewing and not their nutrition value. The results show, that our approach is not limited to a specific group of foods. Moreover, it is possible to discriminate foods from the same group. The actual nutrition value can be derived either precisely from other monitoring components, e.g. RFID tags of packages, or as an estimate from a generic food database.

An important aspect of our work is the fact that information about the specific type of food which is being chewed is very difficult to derive using other sensor modalities. The only alternative we could think of is video analysis of the items inserted into the mouth. While theoretically feasible it has many problems of its own, in particular sensitivity to light conditions and background clutter as well as large computational complexity.

Overall the results presented in this paper provide crucial groundwork for further development that, we believe, will lead to complete automated dietary monitoring systems. Within the scope of the EU-funded MyHeart project we aim to have first versions of such a system within the next two to three years. Additionally, points 1 and 2 have implications beyond dietary monitoring as they allow a fairly accurate recognition of the fact that the user is eating. This in itself is an important context information.

6.6.2. Future work

On the sound analysis the next steps that we will undertake are:

1. Modelling temporal evolution of the signal from individual chews with hidden Markov models to further increase the recognition rates and allow similar food types to be distinguished.

2. Modelling the temporal evolution of the individual chewing signals over an entire chewing cycle to extract food type specific parameters. This shall include the number of individual chews needed, their length and the evolution of the sound intensity.

3. Performing studies about the robustness of the system by adding controlled levels of noise.

4. Performing more studies with more, different food types.

5. Performing studies to determine how the recognition performance degrades with increasing number of food types that need to be differentiated.

6. Using a hierarchical approach with an initial classification of the category (dry crisp, wet crisp etc.) and then a category specific algorithm for further recognition, to overcome the above limitation.

Furthermore, other components of a dietary monitoring system will also be investigated. In particular, we will look at the detection of swallowing motion with collar electrodes, analyse the hand motions related to food intake and integrate high level context information relevant to eating habits into the system.

Bibliography

[1] W. AlChakra, K. Allaf, and A. Jemai. Characterization of brittle food products: Application of the acoustical emission method. *J Tex Stud*, 27(3):327–348, July 1996. doi:10.1111/j.1745-4603.1996.tb00078.x.

[2] M. Beigl, H.-W. Gellersen, and A. Schmidt. MediaCups: Experience with design and use of computer-augmented everyday artefacts. *Computer Networks*, 35(4):401–409, March 2001. ISSN 1389-1286. doi:10.1016/s1389-1286(00)00180-8. Special Issue on Pervasive Computing.

[3] H. S. Brenman, R. C. Weiss, and M. Black. Sound as a diagnostic aid in the detection of occlusal discrepancies. *Penn Dent J*, 69(2):33–39, Feb 1966.

[4] D. Brochetti, M. Penfield, and S. Burchfield. Speech analysis techniques: A potential model for the study of mastication sounds. *J Tex Stud*, 23(2):111–138, 1992. doi:10.1111/j.1745-4603.1992.tb00515.x.

[5] C. Dacremont, B. Colas, and F. Sauvageot. Contribution of air- and bone-conduction to the creation of sounds perceived during sensory evaluation of foods. *J Tex Stud*, 22 (4):443–456, January 1991. doi:10.1111/j.1745-4603.1991.tb00503.x.

[6] N. DeBelie, V. De Smedt, and D. B. J. Principal component analysis of chewing sounds to detect differences in apple crispness. *Postharvest Biol Technol*, 18:109–119, 2000.

[7] B. Drake. Food crushing sounds. an introductory study. *J Food Sci*, 28(2):233–241, March 1963. doi:10.1111/j.1365-2621.1963.tb00190.x.

[8] J. Edmister and Z. Vickers. Instrumental acoustical measures of crispness in foods. *J Tex Stud*, 16(2):153–167, 1985.

[9] H. W. Gellersen, A. Schmidt, and M. Beigl. Multi-sensor context-awareness in mobile devices and smart artifacts. *Mob Netw Appl*, 7(5):341–351, October 2002. ISSN 1383-469X. doi:10.1023/a:1016587515822.

[10] J. Hutchings and D. Lillford. The perception of food texture - the philosophy of the breakdown path. *J Tex Stud*, 19(2):103–115, 1988. doi:10.1111/j.1745-4603.1988.tb00928.x.

[11] K. Kapur. Frequency spectrographic analysis of bone conducted chewing sounds in persons with natural and artifcial dentitions. *J Tex Stud*, 2(1):50–61, 1971. doi:10.1111/j.1745-4603.1971.tb00272.x.

[12] W. Lee, G. Schweitzer, G. Morgan, and D. Shepherd. Analysis of food crushing sounds during mastication: total sound level studies. *J Tex Stud*, 21:156–178, 1990. doi:10.1111/j.1745-4603.1990.tb00473.x.

[13] D. Li, I. K. Sethi, N. Dimitrovac, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recogn Lett*, 22(5):533–544, April 2001. doi:10.1016/s0167-8655(00)00119-7.

[14] S. Z. Li. Content-based audio classification and retrieval using the nearest feature line method. *IEEE T Speech Audi P*, 8(5):619–625, September 2000.

[15] A. Mihailidis, B. Carmichael, and J. Boger. The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. *IEEE Trans Inf Technol Biomed*, 8(3):238–247, Sept. 2004. doi:10.1109/titb.2004.834386.

[16] E. Mynatt, A.-S. Melenhorst, A.-D. Fisk, and W. Rogers. Aware technologies for aging in place: understanding user needs and attitudes. *IEEE Perv Comput*, 3(2):36–41, April-June 2004. doi:10.1109/mprv.2004.1316816.

[17] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa. Computational auditory scene recognition. In *ICASSP 2002: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1941–1944. IEEE Press, May 2002.

[18] M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Perv Comput*, 3(4):50–57, Oct-Dec 2004. doi:10.1109/mprv.2004.7.

[19] J. F. Prinz. Computer aided gnathosonic analysis: distinguishing between single and multiple tooth impact sounds. *J Oral Rehabil*, 27(8):682–689, Aug 2000.

[20] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD 1997: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 43–48, August 1997.

[21] K. Römer, T. Schoch, F. Mattern, and T. Dübendorfer. Smart identification frameworks for ubiquitous computing applications. In *PerCom 2003: Proceedings of the First IEEE International Conference Pervasive Computing and Communications*, 2003.

[22] M. Stäger, P. Lukowicz, N. Perera, T. von Büren, G. Tröster, and T. Starner. Soundbutton: Design of a low power wearable audio classification system. In *ISWC 2003: Proceedings of the 7th IEEE International Symposium on Wearable Computers*, pp. 12–17, Oct 2003. doi:10.1109/iswc.2003.1241387.

[23] A. Szczesniak. Texture: Is it still an overlooked food attribute? *Food Technology*, 44 (9):86–95, 1990.

[24] K. Tyson. Monitoring the state of occlusion - gnathosonics can be reliable. *J Oral Rehabil*, 25:395–402, 1998.

[25] Z. Vickers. Relationships of chewing sounds to judgements of crispness crunchiness and hardness. *J Food Sci*, 47(1):121–124, 1981. doi:10.1111/j.1365-2621.1982.tb11041.x.

[26] Z. Vickers. Sensory acoustical and force-deformation measurements of potato chip crispness. *J Food Sci*, 52:138–140, 1987.

[27] Z. Vickers and C. Christensen. Relationships between sensory crispness and other sensory and instrumental parameters. *J Tex Stud*, 11:291–307, 1980. doi:10.1111/j.1745-4603.1980.tb00327.x.

[28] J. F. V. Vincent. The quantification of crispness. *J Sci Food Agric*, 78:162–168, 1998.

[29] D. Watt. Gnathosonics - a study of sound produced by the masticatory mechanism. *Journal of Prosthetic Dentistry*, 16(14):73, 1966.

[30] D. Watt. *Gnathosonics and Occlusal Dynamics*. Praeger New York, 1981.

[31] S. E. Widmalm, W. J. Williams, and C. Zheng. Time frequency distributions of tmj sounds. *J Oral Rehabil*, 18(5):403–412, Sep 1991.

[32] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1 edition, 1999.

7

Temporal sequences in chewing sounds

Oliver Amft, Martin Kusserow and Gerhard Tröster

Full publication title: Automatic identification of temporal sequences in chewing sounds
BIBM 2007: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 194–201, IEEE Press, 2007.
DOI: 10.1109/BIBM.2007.25

Abstract

Chewing is an essential part of food intake. The analysis and detection of food patterns is an important component of an automatic dietary monitoring system. However chewing is a time-variable process depending on food properties. We present an automated methodology to extract sub-sequences of similar chews from chewing sound recordings. The approach is based on a chew-accurate segmentation of the sound signal, a multi-objective evolutionary search for temporal partitions in the sequence using NSGA-II and a validation of the best solution by classification.

We evaluate the method on chewing sound recordings from a four participant study, eating foods with different rheological properties. The proposed methodology allows to determine the most appropriate partitioning of the sequences and extract relevant sound features at the same time. Potato chips and chocolate showed a two-phase structure, for lasagna and apples a single-phase structure was derived. The results led to the hypothesis that a sequential structure can be found in chewing sounds from brittle or rigid foods.

7.1. Introduction

Food intake is a vital aspect of human health. The prevalence of over- and under-consumption as well as unbalanced meal composition surges the risk of chronic diseases such as obesity. Consequently, assistive systems that monitor food intake from non-invasive sensors could provide a valuable tool for dietary monitoring in risk groups. Manual dietary monitoring systems that are currently used require frequent user interaction, e.g. to log food type and time in consumption questionnaires, scan or take pictures of the foods consumed. Besides the high post-processing efforts of professional services to analyse and verify the data, all of these methods require intensive collaboration by the user.

We believe that the automation of dietary monitoring could alleviate this problem by reporting daily schedule and tentative food consumption using a wearable system. Towards automatic dietary monitoring, a system was proposed [3] that consists of three sensing domains: 1) identification of arm and trunk movements that characterise the food intake from inertial sensors [2], e.g. using fork and knife, a spoon or hand-only movements, 2) detection of swallowing from a sensing collar [4] and 3) characterisation of foods from chewing sounds. The latter sensing domain is the focus of this work.

The breakdown of foods during chewing generates sound emissions that are conducted by bones, skull and body tissue. For a wearable dietary monitor, sound from chewing can be recorded by a microphone or similar acoustic transducer in the ear canal or close to it. In this way, different foods have been classified from their acoustic profile during chewing [3, 9]. However chewing is a time-dependent process, modified by the changing properties of the food material in the mouth when fluids are pressed out, the food is mixed with saliva and a bolus is formed [17]. The intrinsic chewing movement pattern is generated by a brain stem pattern generator. The movement pattern is continuously modified by oral sensory feedback [18].

Chewing sequences start from an *initial bite* (shearing of a sample from a food piece with the incisors), followed by a variable number of rhythmic *chewing cycles* (compressing the sample using molars) until swallowing occurs. Recent studies confirmed that changes occur in movement as well as muscle activity of the masticatory apparatus within these sequences (see Section 7.1.1 for a discussion of related work). Changes in movement of the mandible were attributed to the modification of rheology parameters (hardness, fracturability, adhesiveness) of food [13]. Since both, movement and rheology changes during a chewing sequence, it can be hypothesised that different acoustic stages exist in the sequence as well. Such stages could relate to sensory changes occurring during the sequence. Moreover, if the existence of acoustic stages could be confirmed, sequential food classification models become feasible. Such models can achieve a better fit to the chewing sound pattern and potentially improve the scalability of food sound models.

In the present work we apply an automated methodology to extract sub-

sequences of chewing sequences from ear-canal chewing sound recordings. Sub-sequences are defined as a series of cycles in the chewing sequence with similar acoustic properties. Since the temporal evolution of the acoustic pattern and hence chewing sound stages are largely unknown, an unsupervised search and optimisation strategy was deployed to analyse the sub-sequences in a chewing sequence and select appropriate features for discriminating these structures at the same time. The solutions are qualified for compliance to a chewing sequence model and validated by classification on test data.

7.1.1. Related work in chewing monitoring

First insight into the sequence of chewing was obtained by kinematic and electromyographic (EMG) studies. At the beginning of the sequence hardness was found to control the frequency of chewing and activity of M. masseter and M. temporalis while in middle of the sequence product rheology described mandibular movement such as vertical amplitude [13]. Similar results were found in earlier studies, where burst duration and mean voltage of the EMG as well as vertical mandibular movement decreased during the sequence [15]. Chewing of materials with different rheology indicated, that changes in the activity pattern of M. masseter occur on the chewing cycle level during a sequence, however sequence parameters did not differ for materials of varying rheology [5].

Based on the initial investigation by Drake [11] chewing sound has been assessed predominately to study auditory and sensory perception of material texture in food science [1, 19]. From observing audio waveforms it was assumed that a normal chewing sequence could be partitioned into two phases: initial gross cutting of the ingested material and subsequent conversion in fine grained particles [11]. This process is understood as a gradual decomposition of the material structure during chewing and is audible as a decline of the sound amplitude in brittle foods [3].

The loudness of a foodstuff during chewing depends mainly on the cell arrangement, impurities and existing cracks [1]. Naturally grown foods, e.g. apples or lettuce, contain more liquids compared to dry products, e.g. potato chips, that have air inclusions [12].

Most previous works that targeted classification of chewing sounds used a small share of chewing cycles from each sequence only or analysed a sequence as one entity. None of these works used the complete chewing sequence, consequently phasing of the sequences was not investigated. De Belie et al. [9] used the sound spectrum of the initial bite to compute principal components and discriminate two classes of crispness in apples. More recently a classification of five snack food products was presented, based on the auditory signal of the initial bite and the first chew [10].

Analysing the initial bite or the chewing cycle refers to two separate investigation techniques. It was found that bite and chewing cycles differ regarding

movement and emitted sound [6, 16]. For our purpose of finding stages in the chewing sequence, all chewing cycles are included in this analysis, while the initial bite is manually excluded.

7.2. Methods

This section summarises the study procedure to acquire chewing data and presents the data analysis steps for extracting a partitioning from the chewing sequences. After summarising the feature processing from chewing cycles, the multi-objective search strategy is presented. Finally the procedure to identify and test the best partitioning solution is detailed.

7.2.1. Chewing study

Four participants (male students, natural dentition, aged 20 to 30 years) without known chewing or swallowing abnormality were recruited for the study. During a pre-recording interview the lab and measurement environment was shown to each participant for familiarisation. Two measurement sessions were carried out on separate days, with at least one day break in between. Each session was recorded around mid-day. Participants were asked to eat the following foods: Potato chips ("Chio chips Ready salted", ∼25 pieces), meat lasagna (∼250 g), one apple ("Jangold", ∼100 g), 12 pieces of chocolate ("Coop lait", total: 40 g). The meat lasagna was a commercial deep-frozen version, heated in an oven for 40 minutes.

All participants were familiar with the food types based on cultural origin. None of the participants expressed a dislike for any of the foods nor problems to chew or swallow the selected foods. Participants were sitting comfortably on a chair close to a table carrying the food items and a glass of water. They were asked to chew and swallow normally and allowed to move, drink and speak during the recording sessions between the chewing sequences. The recording duration was not constraint since the participants were eating/drinking at their individual pace.

Chewing sound was recorded using a miniature microphone (Knowles, TM-24546) embedded into an ear-pad. The sound signal was amplified and sampled at 44 kHz, 16 bit. Surface EMG from left and right M. masseter was recorded at 2 kHz, 24 bit.

An observer controlled the recording procedure during each session and annotated the chewing sequences. In a post-processing step all annotated sequences were reviewed, the start/end times adapted and swallowing events marked for exclusion by inspecting the signals. Synchronisation marks in the data (set during the recordings) were used to align audio and EMG data streams.

7.2.2. Chewing segmentation and feature processing

From the continuous chewing sequence recordings individual chewing cycles were extracted using the muscle activity derived from EMG. Signals from left and right M. masseter were bandpass filtered and combined for independence from the chewing side. Peak muscle activity was used to estimate teeth clench (complete occlusion, teeth in full contact) and hence the previous onset of the combined EMG signal determined the beginning of the current and end of the previous chewing cycle. The first and last cycles were determined using the sequence annotation. Only chews segmented in the bounds of the annotation were used for further analysis.

Three feature search spaces were computed from the segmented chewing cycle. For search space one, the complete segmented chewing cycle was used, for search space two, only an initial part of the chewing cycle was used (EMG onset to peak) and for search space three both previous spaces were combined. Spaces one and two contained 65 audio features (130 for feature space three), computed from time and frequency domain. The time domain features included: length of the segment, extrema, fluctuation, zero crossings and the integrated signal. The spectral features included: total and band energy, fluctuation, centroid, bandwidth, rolloff as well as auto-correlation and cepstral coefficients.

For the subsequent investigation observations were derived by computing the features from every chewing cycle. The total set of observations was split into a search and testing set. The testing set (10% of all observations) was used for result validation in the final classification.

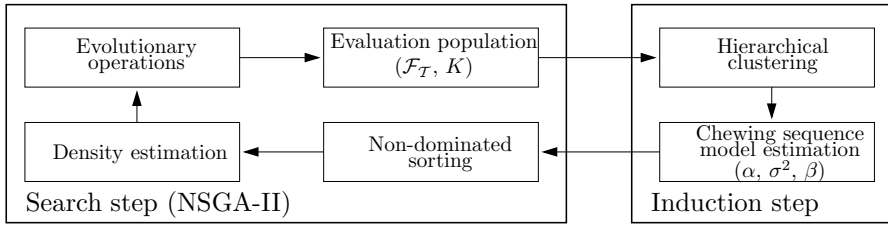


Figure 7.1. Outline of the search procedure for feature selection and chewing sequence partitioning.

7.2.3. Sequence partitioning using multi-objective search

Finding partitions in the temporal evolution of chewing sequences was regarded as a search problem that required the identification of related observation subsequences and the selection of discriminative features to model the partitioning. Fig. 7.1 provides an overview on the search framework, composed of a search and an induction step. In the search step potential features were selected. In the

induction step, these features were used to determine clusters in the observation data. The clustering result was validated against a chewing sequence model.

Search step

We selected the Non-dominated Sorting Genetic Algorithm, version 2 (NSGA-II) [8] as search algorithm and feature selection wrapper. NSGA-II belongs to the family of search heuristics based on evolutionary algorithms that can accommodate multiple search goals. A detailed introduction to evolutionary algorithms can be found in [7].

The algorithm keeps a diverse population of individual solutions and aims at finding non-dominated (Pareto-optimal) solutions by applying evolutionary operations (selection, mutation and crossover). With the diversity of individuals in the population a high robustness is achieved: locked oscillations and single solutions at local optima are avoided. Moreover, the algorithm promotes elitism by maintaining Pareto-optimal solutions, once found, in the following generations. The algorithm achieved a good performance when compared to similar methods [20].

For each individual solution, a binary bit vector encodes the feature set and expected sub-sequences (clusters) in the chewing sequences. At initialisation a uniform distribution of the bits was used to set the vector.

The genetic search was performed with a population size of 100 individuals, a mutation rate of 0.05 (uniform) and a crossover rate of 0.8 (single point). An independent search was performed for each of the three feature spaces. After 250 generations the feature selection was stopped and the Pareto-optimal solutions were evaluated and tested as described in 7.2.4.

Induction step

As induction step for the feature selection we used a hierarchical clustering of the chewing observations in each sequence and assessed the partitioning result using validity parameters of a chewing sequence model. The validity parameters were used in the search step (as search goals) and during the result selection (as quality measure).

The expected number of sub-sequences (specified in the search step) was used as clustering target. We analysed the range of 2 to 5 clusters.

The following chewing sequence model was defined to describe the search problem and to develop the model validity parameters. The model was applied for each food type individually.

A complete sequence of chewing cycles $\mathcal{S}_i$, ($\mathcal{S}_i \in \mathcal{S}$) of food type $\mathcal{T}$ consists of K unique sub-sequences, called phases $\mathcal{P}$. $\mathcal{S}$ is the set of all N sequences from $\mathcal{T}$. Each phase $\mathcal{P}_{i,j}$ consists of $M_{i,j}$ chewing observations described by feature vector $f_{i,j,n}$ from feature space $\mathcal{F}_\mathcal{T}$. The relation of the sets is described

by Eq. 7.1.

$$\begin{aligned} \mathcal{S} &= \{\mathcal{S}_1 \dots \mathcal{S}_N\} \\ \mathcal{S}_i &= \{\mathcal{P}_{i,1} \dots \mathcal{P}_{i,K}\}, \qquad i = 1 \dots N \\ \mathcal{P}_{i,j} &= \{f_{i,j,1} \dots f_{i,j,M_{i,j}}\}, \quad \forall i : j = 1 \dots K \end{aligned} \tag{7.1}$$

Phases can neither overlap nor repeat in a single chewing sequence $\mathcal{S}_i$. An observation $f_{i,j,n}$ belongs to exactly one phase $\mathcal{P}_{i,j}$.

Based on this model formulation, the following model validity parameters were defined to describe a chewing sequence: *phase count* α, *phase size variance* σ^2 and *phase transitions* β. The parameters were defined in order to obtain a minimisation problem.

1. The parameter *phase count* α relates the number of retrieved phases $\hat{K}$ and the number of expected phases K, normalised over all sequences of $\mathcal{S}$ (Eq. 7.2).

$$\alpha = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{K}}{K} \tag{7.2}$$

   The value of α is minimal (zero) if $\hat{K} = K$, i.e. if the number of phases found is equal to the number of expected phases in all sequences of $\mathcal{S}$.

2. The parameter *phase size variance* σ^2 depicts the variation in the number of chewing observations within each phase, normalised over all expected phases K and chewing sequences N (Eq. 7.3). M_j is the mean share of observations in all sequences assigned to phase P_j.

$$\sigma^2 = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=1}^{N} \frac{(\frac{M_{i,j}}{M_i} - M_j)^2}{N-1} \tag{7.3}$$

$$M_i = \sum_{j=1}^{K} M_{i,j} \; ; \qquad M_j = \frac{1}{N} \sum_{i=1}^{N} \frac{M_{i,j}}{M_i} \tag{7.4}$$

   The parameter σ^2 reflects the stability of the obtained partitioning: if $\sigma^2 = 0$, the relative number of chewing observations in each phase is constant for all phases of all chewing sequences.

3. The parameter *phase transitions* β is defined as ratio of existing phase changes in each sequence and the expected number of phase changes ($K-1$), normalised over all sequences in $\mathcal{S}$ (Eq. 7.5 and 7.6).

$$\beta = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\sum_{n=1}^{M_i - 1} d_{i,n}}{K-1} - 1 \right| \tag{7.5}$$

$$d_{i,n} = \begin{cases} 1 : cluster(f_{i,n}) \neq cluster(f_{i,n+1}) \\ 0 : otherwise \end{cases} \tag{7.6}$$

The parameter β measures the consistency of the phases in the sequences: the higher β, the less correspondence exist between the obtained grouping and the expected sequential partitioning (the more alternations exist). If $\beta = 0$, the number of transitions matches the expectation. If $\beta > 0$ the number of transitions does not match, because more or less than the expected transitions were obtained.

7.2.4. Result selection and testing

A selection strategy was developed to extract the best result from the solution space. Fig. 7.2 provides an overview on the applied selection steps.

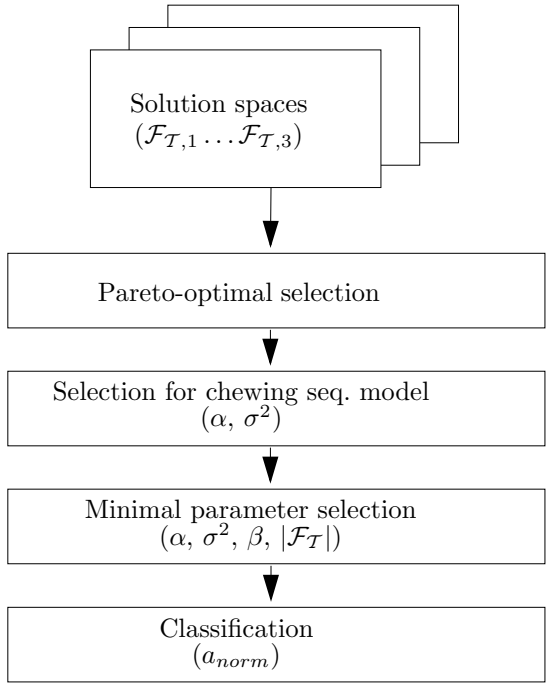


Figure 7.2. Outline of the selection and testing procedure.

All results from the independent search runs for each feature space were merged into a single solution space and their Pareto-optimal individuals were selected. Fig. 7.3 exemplarily visualises the obtained solution space for the three model validity parameters.

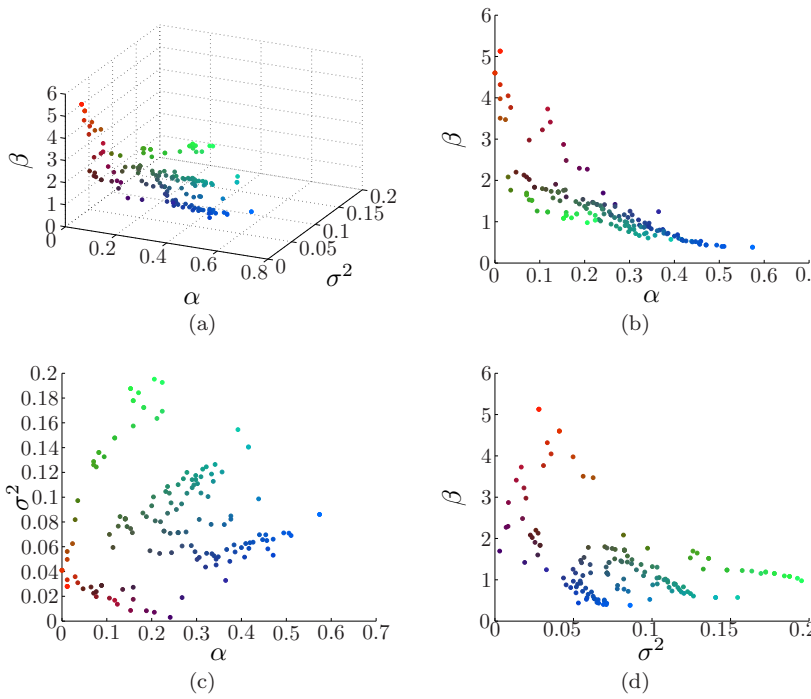


Figure 7.3. Visualisation of the solution space derived by the search procedure for food type *chocolate* after merging of the feature spaces. Each point corresponds to an individual solution. The optimal compliance with the chewing sequence model is at $\alpha = 0$, $\sigma^2 = 0$ and $\beta = 0$. The plots show: (a) All axes α, σ^2 and β; (b) Projection on α-β; (c) Projection on α-σ^2; (d) Projection on σ^2-β.

In the following step all solutions that weakly matched the chewing sequence model were removed. Eq. 7.7 shows the constraints applied for α and σ^2. Individual solutions that did not comply with these bounds were removed from the solution space. Parameter p_{err} limits the share of sequences that do not conform to the chewing sequence model and was set to $p_{err} = 0.2$.

$$\alpha \leq \frac{p_{err}}{K} \; ; \qquad \sigma^2 \leq \frac{p_{err}}{2} \tag{7.7}$$

The best candidate was chosen from the remaining solutions in a final selection step. The space was reduced by sequentially selecting the minimal value of each parameter α, σ^2, β and $|\mathcal{F}_\mathcal{T}|$. Individuals with a minimal α were chosen first, since the correct phase count is a prerequisite to achieve the partitioning target.

In order to verify the result, the identified solution was applied in a classification of phases in the testing observation set. Using the selected features for the best candidate, a nearest centroid classifier was trained on the training set using the candidate's clustering result as class association rule. For evaluating the testing result, the class associations were obtained by assuming the phase distribution of the best candidate. In this way, the solution was tested independently of the search procedure.

The partitioning into phases resulted in class skews (one phase contained more observations than another phase). Training a skewed classifier was avoided by selecting an equal number of training observations from all phases.

To compare classification results with an unequal number of test observations in each class, the normalised accuracy measure as used. For a multi-class classification the normalised accuracy was derived as mean of the class-relative accuracies:

$$a_{norm} = \frac{1}{K} \sum_{i=1}^{K} \frac{Recognised_i}{Relevant_i} \tag{7.8}$$

where K is number of expected phases, $Recognised_i$ and $Relevant_i$ are the number of correctly identified observations and the total number of observations for each phase category.

7.3. Results and discussion

7.3.1. Chewing segmentation

In total 11480 chewing cycles were extracted for the analysis from 602 annotated chewing sequences. Tab. 7.1 summarises the number of chewing sequences and cycles for each food product.

Table 7.1. Summary of the segmentation results.

Food type	Chewing sequences	Chewing cycles
Apple	102	2447
Potato chips	227	3015
Chocolate	90	1913
Lasagna	183	4105
Total	**602**	**11480**

7.3.2. Food-specific analysis

The partitioning was analysed for every food type, including the observations from all four study participants. The solution space spanned by the model

validity parameters is visualised in Fig. 7.3 for the food type *chocolate*.

The solution space diagram shows the distribution of the individuals around the optimal solution of the minimisation problem ($\alpha = 0$, $\sigma^2 = 0$ and $\beta = 0$). While the optimal solution was not achieved in the search, good solutions were found, using the result selection procedure.

The result selection from a solution space is considered to be a critical step for multi-objective approaches [14]. The procedure applied in this work aimed at evaluating the search goals individually, instead of combining the goals in a single weighting function. This strategy was reasonable due to the exploratory analysis approach.

In the food-specific evaluation a two-phase structure was obtained for all foods. Fig. 7.4 shows the average distribution of the phases (occurrence ratios) in the chewing sequences for the individual foods. The occurrence ratios indicate the relative share and position of each phase in the chewing sequences. For all food types a short first phase (30-40% of the sequence length) was found, followed by a second longer phase. For *potato chips* the longest initial phase was obtained, when compared to the other foods.

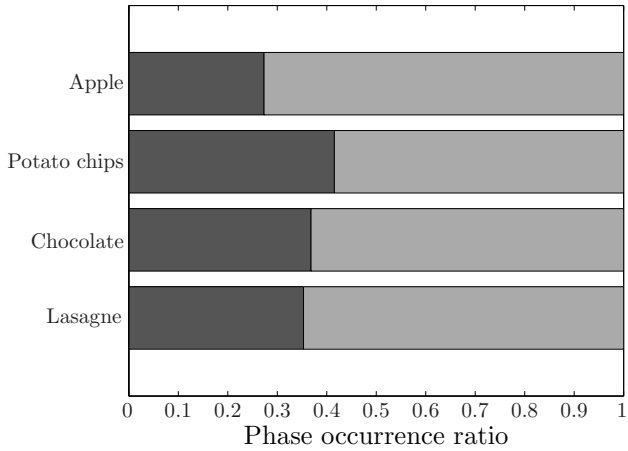


Figure 7.4. Average distribution of phases in the chewing sequences for all four foods obtained in the food-specific analysis. The plots indicate the location of the phases, but do not show insertions inside individual phases. For food types *potato chips* and *chocolate* the partitioning was confirmed by the testing procedure.

This finding confirms the assumption of an initial chewing phase that differs from the remaining sequence made by observing sound amplitude vs. time plots [11]. Tab. 7.2 summarises the phase ratios and quality of the solutions obtained. Both, search and testing performances of the selected candidates are indicated by the model validity parameters of the chewing sequence model.

A good quality, according to the parameters α and σ^2 was found for all four

Table 7.2. Results of the food-specific sequence partitioning analysis.

Search parameter	**Apple**	**Potato chips**	**Chocolate**	**Lasagna**
Expected phases (K)	2	2	2	2
Nr of features ($\|\mathcal{F}_\mathcal{T}\|$)	16	24	22	20
Feature space	2	2	2	2
Sequence model parameter	Candidates after search			
Phase count (α)	0.04	0.00	0.00	0.07
P. size variance (σ^2)	0.07	0.02	0.04	0.09
Phase transitions (β)	3.45	1.35	4.60	5.55
	Testing data			
Phase count (α)	0.07	0.02	0.00	0.08
P. size variance (σ^2)	0.09	0.05	0.03	0.09
Phase transitions (β)	2.86	1.07	4.80	5.63

food types. However, the high number of phase transitions (indicated by β) suggest that the phases had several insertions. This effect was very strong for the food type *lasagna*. The overall best quality for all parameters was achieved for *potato chips*.

The effect of applying the candidate solutions on testing data was very low on the model parameters: the values for all three model parameters remained in the same range. For all foods 16 to 24 features were selected from feature space two.

Whether the sequence structure of the candidates can be verified, was assessed by the classification performance on test data. Fig. 7.5 shows the classification result for all four foods, when applying the properties of the selected solution on the testing set. A normalised accuracy of $a_{norm} = 0.5$ would indicate a random classification. Therefore only the range $a_{norm} = 0.5 \ldots 1.0$ is shown.

The classification confirms the partitioning for *potato chips* and *chocolate*: a normalised accuracy of approx. 80% was achieved, indicating that the two-phase assumption holds, even on the testing data.

For *apple* and *lasagna* a low accuracy was obtained, indicating that the foods cannot be classified with the two-phase search solution. Since phase counts larger than two were rejected during the search stage already, we concluded, that the foods do not exhibit a sub-sequence structure at all.

Table 7.3. Results of the participant-specific sequence partitioning analysis.

Food type	Participant	Expected phases K	Nr of features $\lvert\mathcal{F}_\mathcal{T}\rvert$	Feature space	Phase ratio 1	Phase ratio 2	Sequence model parameters: Candidates α	Candidates σ^2	Candidates β	Testing α	Testing σ^2	Testing β	Norm. accuracy a_{norm}
Apple	A	—	—	—	—	—	—	—	—	—	—	—	—
	B	2	34	1	0.79	0.21	0.00	0.01	2.96	0.00	0.00	0.00	0.67
	C	2	29	2	0.44	0.56	0.00	0.04	10.80	0.00	0.16	9.00	0.61
	D	2	20	1	0.88	0.12	0.00	0.02	3.93	0.50	0.00	1.00	0.50
Potato chips	A	2	26	1	0.52	0.48	0.00	0.02	1.27	0.04	0.05	1.23	0.79
	B	2	61	3	0.72	0.28	0.00	0.01	2.09	0.00	0.00	3.20	0.79
	C	2	28	1	0.34	0.66	0.00	0.00	0.42	0.00	0.00	0.40	0.92
	D	2	65	3	0.84	0.16	0.00	0.01	1.72	0.00	0.01	3.60	0.82
Chocolate	A	2	17	2	0.46	0.54	0.00	0.01	2.09	0.00	0.06	1.50	0.81
	B	2	25	1	0.88	0.12	0.00	0.00	2.95	0.00	0.02	6.00	0.46
	C	2	13	1	0.16	0.84	0.00	0.00	1.95	0.00	0.00	2.33	0.75
	D	2	14	1	0.84	0.16	0.00	0.01	3.71	0.00	0.01	5.00	0.53
Lasagna	A	2	28	2	0.67	0.33	0.10	0.07	1.84	0.33	0.09	1.50	0.51
	B	2	26	2	0.55	0.45	0.05	0.08	3.70	0.11	0.14	2.56	0.56
	C	2	21	2	0.64	0.36	0.01	0.08	9.96	0.00	0.11	6.50	0.56
	D	2	27	1	0.91	0.09	0.00	0.00	2.60	0.25	0.03	4.50	0.57

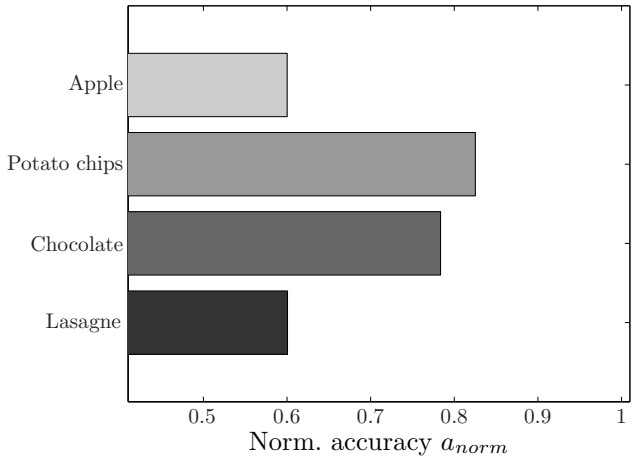


Figure 7.5. Classification result for all four foods when applying the search solution on the testing set. (A normalised accuracy of $a_{norm} = 0.5$ would indicate a random classification. Therefore the plot shows the range $a_{norm} = 0.5 \ldots 1.0$ only.)

7.3.3. Participant-specific analysis

The search and solution selection was performed for each participant individually to analyse the person-related structure of the chewing sequences. Tab. 7.3 summarises the results for all four foods and participants A–D. Except for food type *apple* from participant A, two-phase solutions were found for all combinations of food and participant using the selection procedure. For *apple* from participant A none of the solutions matched the required criteria.

The classification accuracy a_{norm} achieved during testing gives an indication whether the two-phase result can be generalised on the testing data. Except for *chocolate*, the result of the food-specific analysis were confirmed: For *potato chips* a two-phased partitioning was obtained with good classification rates ($0.79 \ldots 0.92$) for all participants. For *apple* and *lasagna* the rates are generally lower. While the performance for *apple* was above random (~ 0.6), the two-phase partitioning for *lasagna* often performed less well.

For *chocolate* participants A and C achieved a classification performance of 0.81 and 0.75 respectively. However, no phase structure was found for the remaining two participants and *chocolate*. In order to analyse this effect in more detail, the study should be extended by additional participants.

7.3.4. Result discussion

The phase distribution and selected features of the participant-specific analysis showed a higher variability when compared to the food-specific analysis. It was

assumed that this is a result of the lower number of test observations and hence less averaging effect in the participant-specific analysis.

In contrast to the food-specific analysis, applying the search solution on the testing data in the participant-specific analysis had a strong effect on the model validity parameters. For *apple* and *lasagna* this indicated the absence of sub-sequences, for the other food types this result was not expected. Although an overall large set of observations was available, the splitting into search and test set (test set size was 10% of all observations) may have led to a reduction of the result stability. Even to confirm the results of the food-specific analysis, a test with different observations should be performed. However, these result confirms the benefit of the applied testing procedure.

The modelling of phases in chewing sequences for an exploratory search is a challenging task. While the developed optimisation goals (phase count, phase size variance and phase transitions) proved to be vital parameters for a valid partition, some limitations of the automatic phase extraction approach remain to be solved.

Regarding the search procedure, the indirect control by rating the solutions after the clustering step, produced invalid solutions, that could have been avoided. To this end clustering should consider the sequence of the observations. However, to achieve this, a highly domain specific grouping algorithm would be needed.

From the phase transitions result (parameter β), large values were found for both food- and participant-specific investigations. While the parameter correctly indicates a partitioning that does not correspond to the analysed number on phases, it is sensitive to insertions. Single assignments of observations to a different phase, which is non-critical for the overall integrity of the partitioning and could be ignored, was not distinguishable from an completely invalid partitioning. This aspect influenced the selection of optimal solutions and could mislead conclusions. Again with the help of our selection and testing procedure this issue was minimised, since β was used as the last of all three model parameters and all results were verified by the classification test.

7.4. Conclusion

We presented an automatic method to extract partitions from chewing sequences, that follow a sequential order and can be identified using sound features. Our approach relied on a search and selection procedure, followed by a verification on test data. The search was performed using a NSGA-II wrapper for selecting appropriate features and expected sub-sequences in combination with an induction step. The induction was composed of hierarchical clustering of the chewing observations and analysing the search quality with respect to a chewing sequence model.

Three parameters were derived that describe a chewing sequence model. The parameters were used for the qualification of the search results and the final selection of the valid solutions. In the present work the model parameters *phase count*, *phase size variance* and *phase transitions* were used.

The sequence structure of four food types was analysed in recordings from four participants, regarding food-specific and participant-specific behaviour. A two-phase structure was found from the food- and participant-specific analysis of the food types *potato chips* and *chocolate*. The search results were verified in a classification test, by assuming the retrieved features and the partition structure in a food sequence model. A classification accuracy of approx. 80% was achieved for both foods. A person-dependency was found for *chocolate*, where no valid phasing was obtained for two of the participants, while for the remaining tow a good classification was possible. We assumed that this variability was caused by the small and fixed test observation set, since the food-specific analysis of all participants returned a classification performance of 78%.

Overall, the food- and participant-specific evaluations returned at maximum two phases for all foods. A common distribution of the phases was found in the food-specific analysis. For all food types, a short first phase (30-40% of the sequence length) was derived, followed by a second longer phase. For the participant-specific analysis different distributions were obtained depending on both, food and person.

Out of the three model validity parameters described above, the most important goal was *phase count*, since the number of actually retrieved phases was vital for the application of the sequence model. Solutions that had large values for this goal ($\alpha \geq 0.2$), typically did not perform well. A selection procedure was designed that reflected this observation, by sequentially limiting the share of sequences that did not conform to the chewing sequence model.

The parameter *phase transitions* was found to be most difficult to minimise, since the phases contained insertions in many sequences of the foods. The insertions were attributed to the natural variability in the data that could not be captured by the clustering algorithm. In the current work, the impact of this issue was minimised by the selection and testing procedure. For further investigations alternate solutions to obtain a requested number of transitions should be evaluated.

The frequent selection of feature space two in the food-specific analysis indicated that the chewing cycles were not stationary and hence supportive information was extracted from the intra-chew signal variation (as captured by the feature space) when compared to the entire chew (feature space one).

For the food types *apple* and *lasagna* no stable partitioning was found, indicating that both foods have a non-sequential sound pattern that alternatively could be described by a single phase for each sequence. Following these findings, it can be hypothesised that out of the different foods analysed, only dry and rigid foods have a clear sequential structure. Consequently, foods that

are soft, e.g. *lasagna* or based on a fibre-structure as *apple* do not change their sound pattern in an ordered sequence.

We consider the classification rates for *potato chips* and *chocolate* as a comparably good result in relation previous investigations of food sound classification. For bite and first chew classifications from five snack foods up to 18% classification error were reported [10]. In our previous work on classifying foods from chewing recordings, rates between 66% and 86% were achieved, depending on the food type [3]. For an application in automatic dietary monitoring the result of the current work will support the development of food-adapted classifiers. For the above food types an independent model of the two phases could be helpful to boost the system performance, while for other (single-phase) foods one model is sufficient.

Bibliography

[1] W. AlChakra, K. Allaf, and A. Jemai. Characterization of brittle food products: Application of the acoustical emission method. *J Tex Stud*, 27(3):327–348, July 1996. doi:10.1111/j.1745-4603.1996.tb00078.x.

[2] O. Amft, H. Junker, and G. Tröster. Detection of eating and drinking arm gestures using inertial body-worn sensors. In B. Rhodes and K. Mase, ed., *ISWC 2005: IEEE Proceedings of the Ninth International Symposium on Wearable Computers.*, pp. 160–163. IEEE Press, October 2005. doi:10.1109/iswc.2005.17.

[3] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of chewing sounds for dietary monitoring. In M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, ed., *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing*, vol. 3660 of *Lecture Notes in Computer Science*, pp. 56–72. Springer Berlin, Heidelberg, September 2005. doi:10.1007/11551201_4.

[4] O. Amft and G. Tröster. Methods for detection and classification of normal swallowing from muscle activation and sound. In E. Aarts, R. Kohno, P. Lukowicz, and J. C. Trainini, ed., *PHC 2006: Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–10. ICST, IEEE digital library, November 2006. doi:10.1109/pcthealth.2006.361624.

[5] I. Ashida, H. Iwamori, S.-Y. Kawakami, Y. Miyaoka, and A. Murayama. Analysis of physiological parameters of masseter muscle activity during chewing of agars in healthy young males. *J Tex Stud*, 38(1):87–99, February 2007. doi:10.1111/j.1745-4603.2007.00087.x.

[6] C. Dacremont, B. Colas, and F. Sauvageot. Contribution of air- and bone-conduction to the creation of sounds perceived during sensory evaluation of foods. *J Tex Stud*, 22 (4):443–456, January 1991. doi:10.1111/j.1745-4603.1991.tb00503.x.

[7] S. De, A. Ghosh, and S. K. Pal. *Genetic Algorithms for Pattern Recognition*, chapter 1, pp. 1–23. CRC Press, Boca Raton, FL, USA, 1996. ISBN 0849394678.

[8] K. Deb, S. Agrawal, A. Pratab, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. J. Merelo, and H.-P. Schwefel, ed., *Proceedings of the Conference on Parallel Problem Solving from Nature VI*, number 1917 in Lecture Notes in Computer Science, pp. 849–858, Paris, France, 2000.

[9] N. DeBelie, V. De Smedt, and D. B. J. Principal component analysis of chewing sounds to detect differences in apple crispness. *Postharvest Biol Technol*, 18:109–119, 2000.

[10] N. DeBelie, M. Sivertsvik, and J. DeBaerdemaeker. Differences in chewing sounds of dry-crisp snacks by multivariate data analysis. *J Sound Vib*, 266(3):625–643, September 2003.

[11] B. Drake. Food crushing sounds. an introductory study. *J Food Sci*, 28(2):233–241, March 1963. doi:10.1111/j.1365-2621.1963.tb00190.x.

[12] J. Edmister and Z. Vickers. Instrumental acoustical measures of crispness in foods. *J Tex Stud*, 16(2):153–167, 1985.

[13] K. D. Foster, A. Woda, and M. A. Peyron. Effect of texture of plastic and elastic model foods on the parameters of mastication. *J Neurophysiol*, 95(6):3469–3479, Jun 2006. doi:10.1152/jn.01003.2005.

[14] A. A. Freitas. A critical review of multi-objective optimization in data mining: a position paper. *SIGKDD Explor Newsl*, 6(2):77–86, 2004. ISSN 1931-0145. doi:10.1145/1046456.1046467.

[15] K. Kohyama and L. Mioche. Chewing behavior observed at different stages of mastication for six foods studied by electromyography and jaw kinematics in young and elderly subjects. *J Tex Stud*, 35(4):395–416, October 2004. doi:10.1111/j.1745-4603.2004.tb00603.x.

[16] C. Lassauzay, M. A. Peyron, E. Albuisson, E. Dransfield, and A. Woda. Variability of the masticatory process during chewing of elastic model foods. *Eur J Oral Sci*, 108(6):484–492, Dec 2000. doi:10.1034/j.1600-0722.2000.00866.x.

[17] P. J. Lillford. The materials science of eating and food breakdown. *MRS Bulletin*, 25 (12):38–43, December 2000.

[18] J. P. Lund and A. Kolta. Generation of the central masticatory pattern and its modification by sensory feedback. *Dysphagia*, 21(3):167–174, Jul 2006. doi:10.1007/s00455-006-9027-6.

[19] Z. Vickers. Relationships of chewing sounds to judgements of crispness crunchiness and hardness. *J Food Sci*, 47(1):121–124, 1981. doi:10.1111/j.1365-2621.1982.tb11041.x.

[20] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the strength pareto evolutionary algorithm. Technical Report 103, ETH Zürich, Gloriastrasse 35, CH-8092 Zurich, Switzerland, 2001.

8

Bite weight estimation

Oliver Amft, Martin Kusserow and Gerhard Tröster

Full publication title: Bite weight estimation using acoustic recognition of chewing
submitted to IEEE Transactions on Biomedical Engineering, June 2008.

Abstract

Manual recording of eating behaviour, food amount in particular, is a cumbersome burden for individuals following a diet management program. This work investigates the estimation of bite weight from chewing microstructure based on techniques developed for Automatic Dietary Monitoring.

Chewing activity was acoustically recorded at the ear. A recognition procedure was developed for detecting chewing cycles in continuous sound data and identifying food type. Based on the recognition result, timing and structure variables of the chewing process were extracted to predict bite weight.

Bite weight estimation was investigated in 50 variables from habitual food intake of eight healthy individuals. The recognised food type was used to select the bite weight prediction model. Three foods of similar acoustic properties but different material structures were investigated: potato chips, lettuce, and apples.

Food-specific multivariate linear models were deployed to predict bite weight. Mean prediction error was lowest for apples (19.4%) and largest for lettuce (31%). The bite weights estimated from chewing sound were compared to an estimation based on Electromyography. We conclude that bite weight estimation using acoustic chewing recordings is a feasible approach for solid foods.

8.1. Introduction

The goal of Automatic Dietary Monitoring (ADM) is to simplify reporting of eating behaviour for personalised weight and diet coaching programs, and many further applications that require dietary supervision, such as obese patients under clinical observation. Such eating behaviour information includes meal schedule and consumed food type of each meal for monitoring durations of several days to years, as for long-term diet coaching [32].

Currently, information on eating behaviour is acquired through self-reports, that are not feasible to maintain for longer time periods than one week [34]. Besides meal schedule and food type the respondent is typically asked to provide the amount of food consumed. While food weight provides important data on the balance of nutrient composition and portion size, weighting every food item adds a substantial burden for the individual to follow a normal lifestyle. This continuous manual effort is detrimental for reporting compliance [24]. Misreporting of intake amount results in an estimation bias that depends on various social and personal aspects [16, 31]. Typically, the reporting becomes inaccurate after a short time and weighting may be omitted completely, even because of an adapted perception of desirable intake patterns [16, 29].

The ADM approach assists an individual in dietary monitoring by extracting and fusing information from on-body sensors that monitor different diet-related activities. These activities include upper body and arm movement during intake [17], chewing [1] and swallowing [2]. In chewing, a particularly vital source of information is the chewing microstructure for each bite taken. That is the sequence of chewing cycles (closing and re-opening movement of the jaw) used to decompose a food piece from ingestion into the mouth until swallowing [35]. Our previous investigations have shown that bone-conducted food breakdown sounds can be recorded by a miniature microphone at the ear canal [1]. Moreover, food category and individual chewing cycles could be recognised from acoustic pattern models. These acoustic patterns correspond to food texture properties [3].

The goal of this work was to analyse the potential for estimating food weight in individual bites. In particular, our focus was to predict bite weight under unconstrained food size selection and freestyle chewing. For this purpose an acoustic recognition of chewing cycles was used to derive structural and timing variables of the chewing microstructure. These variables included the number of chews to consume a food piece, duration of chews, chewing speed, and total chewing time. Bite weight estimation from the sound-based recognition was compared to a second chewing cycle recognition approach using muscle activity recorded from surface Electromyography (EMG). In total 50 microstructure variables were analysed from three foods and eight healthy individuals.

Moreover, to demonstrate the applicability of the sound-based bite weight estimation for ADM, we present the recognition approach to identify and categorise chewing cycles and food type in continuous sound data. The recognised

food type was subsequently used to select a bite weight prediction model. Performance of this recognition architecture was quantitatively analysed using cross-validation and compared to sound-energy and EMG detections.

8.1.1. Chewing microstructure

Human chewing adapts to food type according to the individual's physiology and capabilities as well as taste [25] Food ingested to the mouth excites different oral receptors that convey sensory information on the material properties to the brain stem. Most important stimuli are food texture related, such as crispness and hardness, size and shape related as well as flavour related [22, 35].

The continuous adaptation to these stimuli target an efficient food breakdown and creating a food bolus, feasible for swallowing [37]. However, intra-individually, this process is fairly constant. Using controlled settings and food stimuli, no significant differences were found in several repetitions, analysing mandibular movement parameters, muscle activity and chewing microstructure [7, 20]. This stability is the key aspect to derive personalised bite weight estimation models in this work.

Some variables of the chewing microstructure were reported to alter in response to variations of the bite size for constant food [35]. The most consistent reports exist for variables measuring the number of chewing cycles and chewing sequence duration from ingestion to swallowing. Both variables increased with bite size for artificial food [8] and three natural foods using constrained sizes [13]. Moreover the movement trajectory of single chewing cycles change with bite size [27, 30]. A prominent effect is an increase in vertical jaw amplitude and inter-arch distance with increasing bite size during the first chewing cycles [9, 27]. The reported relations were mostly tied to correlation and linear regression analyses with bolus dimensions.

The *estimation* of bite weight from the chewing microstructure remains broadly unexplored. Neither are reports available that analyse the relevance of larger microstructure variable sets. To this end it is unclear whether changes within chewing sequences occur that relate to bite weight. Most investigations assessed the chewing microstructure to analyse masticatory performance using standardised food items [15, 27]. Weight estimation was neither investigated for fixed induced bite sizes nor in the habitual freestyle settings. Consequently, it is not clear how consistently habitual bite weight of different natural foods is reflected in the chewing behaviour.

8.1.2. Food selection

Food texture provides vital features for an intra-individual food discrimination using chewing sounds [3, 10]. This relation to texture and material structure was investigated in a detection of chewing cycles based on texture groups [3].

Texture groups can be defined in various forms, since they relate to the food perception and jargon of human panelists [14]. For the purpose of this paper, we refer to two texture groups and utilise three different foods: (1) *wet* structures from naturally grown foods, such as *apple* and *lettuce*, and (2) *dry* foods, such as *potato chips*.

Although these foods are different in composition, our previous investigations have shown that bone-conducted breakdown sounds of crisp textures with wet and dry structures are often confused during detection. Nevertheless, a robust categorisation of the food type is needed in our approach to select the bite weight estimation model. Consequently, the recognition procedure was optimised to discriminate these foods from their acoustical pattern.

Hence, the foods selected in this work serve two purposes. Firstly it allowed to evaluate the discrimination of two similar sound emission groups (dry and wet structures). Secondly, we studied the habitual bite size selection and the bite weight prediction in different weight ranges. Regarding the second goal, the foods were chosen to cue different weight selections: from very low weight potato chips to apples that are typically consumed in mouthful bites.

8.2. Experimental procedure

8.2.1. Study protocol

Eight volunteer students (two female, six male) aged between 20 to 35 years were recruited from different ETH departments through advertisements. All participants had natural dentition and no known history of chewing or swallowing abnormalities. Further exclusion criteria were known disorders or audible sounds of the temporomandibular joints as well as known food allergies. A pre-recording interview was conducted with each participant in the measurement room for familiarisation. The recording procedure was explained, however the specific goal of this investigation was not mentioned. Participants were invited for a individual recording session around mid-day.

Participants were asked to eat the following foods: potato chips (Chio chips "Ready salted", ∼25 pieces, 20 g), mixed lettuce (containing endive, sugar loaf, frisée, raddichio, chicory, arugula, ∼55 grams) and one apple ("Jangold", ∼110 g). The food amounts indicate approximate values, participants were allowed to eat the foods, chew and swallow in their habitual style. From the potato chips a few chips were taken for each bite with the hand, lettuce was consumed using fork and knife and apples were eaten by taking bites from the entire skinned fruit. The fruit core was not consumed.

All participants were familiar with the food types. None of the participants expressed a dislike or problems to chew or swallow the selected foods. Participants were allowed to move, drink water and speak during the recording sessions. The recording duration was not constraint since the participants were eating/drinking at their individual pace. Informed consent was obtained from

each participant. The study protocol was reviewed and approved by the ETH ethics committee.

8.2.2. Data recording

Chewing sound was recorded using a miniature microphone (Knowles, TM-24546) embedded in an ear-pad device. The occlusion of the pad was kept low, so that participants could still hear room-level conversation at the applied side. The room was controlled for a constant noise level of an office environment. The sound signal was amplified and sampled at 44 kHz, 16 bit. Surface EMG was recorded bilaterally from Ms. masseter at 2 kHz, 24 bit and bandpass filtered. The plate weight was recorded at ∼1 Hz with a resolution of 0.1 grams using a weight scale build into the table. The scale performed an automatic measurement stabilisation.

An observer controlled the recording procedure during each session and annotated the chewing sequences and swallowing events. In a post-recording step all annotated sequences were reviewed, the start/end times adapted and swallowing events marked for exclusion by inspecting the signal waveforms.

Figure 8.1 shows a sample plot of sound and EMG data for one chewing sequence. Individual chewing cycles are visible as change in signal amplitude of audio and EMG. The mandible closing phase of each cycle is marked as shaded area. The entire chewing sequence corresponds to the intake of one piece of apple from ingestion into the mouth until final swallow. While intermediate swallowing may occur none was observed in the depicted example.

8.2.3. Chewing annotation

To derive food pattern models, the annotation (location in the recorded data and food type) of chewing cycles was needed. A manual annotation of chewing cycles (mandible closing phase) was performed in a post-recording step by reviewing the sound and EMG waveforms and listening to the chewing sounds. To differentiate the phase annotation from the complete chewing cycle, we denote the annotated phase as chewing event where necessary.

The chewing event annotation was performed by one observer, in order to keep the annotation as consistent as possible. The method is accurate in identifying every chewing cycle until the food bolus is swallowed, however it is expensive and time-consuming for large chewing data sets. To alleviate this effort, a share of all chewing sequences were annotated, in total approx. 300–400 chewing events for each subject. The remaining chewing sequences were marked as continuous sections where chews may potentially exist. Intermediate swallows were excluded in both cases. The recognition procedure was adapted to this annotation method, as detailed in Section 8.5.1. In total 7301 chewing cycles were identified and annotated in 473 chewing sequences. Including the continuously marked sections, a total of 504 chewing sequences were recorded.

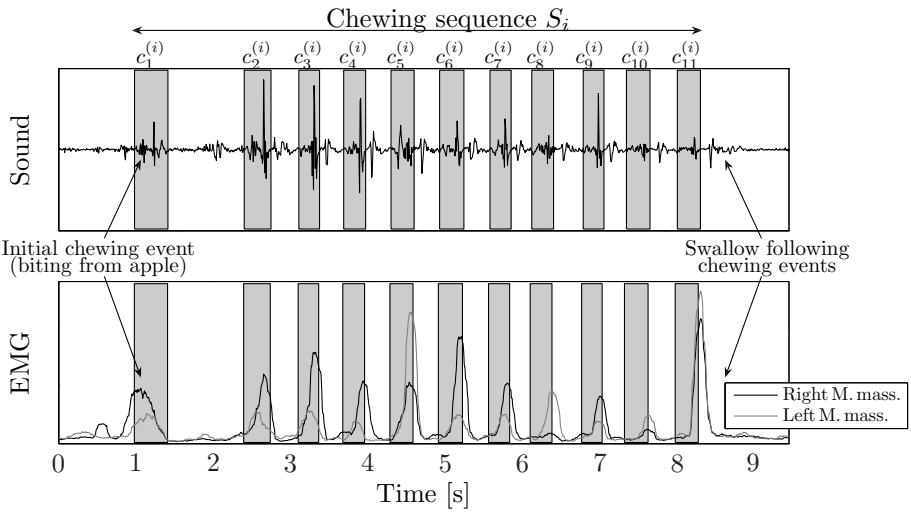


Figure 8.1. Illustration of sound (upper plot) and averaged rectified EMG (lower plot) waveforms for a chewing sequence S_i from apple. Individual chewing cycles $c_j^{(i)}$ are visible as change in signal amplitude of audio and EMG. The mandible closing phases are marked as shaded areas.

The total length of the data set was 8.64 h, the average length per participant was 64.83 min (SD 14.6 min).

8.3. Recognition of chewing

Figure 8.2 illustrates the complete evaluation procedure to detect chewing events, identify food type and predict bite weight. In order to recognise chewing events a feature similarity search (FSS) was applied for each food and subject. The food type was subsequently classified for each detected chewing event and a chewing sequence voting was applied to determine the food type of each sequence. Finally, by comparison of all concurrently detected events, those with the highest model confidence were retained in a final fusion step (COMP). Below, the recognition procedure is presented in detail. The remaining steps to estimate bite weight are discussed in Section 8.4.

8.3.1. Feature similarity search

FSS is an online event-oriented pattern detection algorithm, based on a variable-length feature pattern search. The method is a generalisation of the classic sliding-window algorithm and has been discussed in previous works, e.g. [3]. It was adapted here for detection chewing events of a specific food in

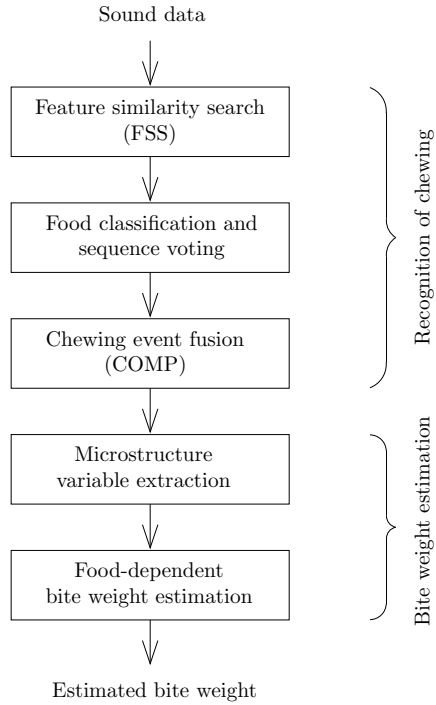


Figure 8.2. Analysis procedure to recognise chewing events and estimate bite weight. The steps related to *Recognition of chewing* were used to detect chewing events and identify the food type (see Section 8.3 for details). The steps related to *Bite weight estimation* were used to extract the microstructure variables and estimate bite weight (see Section 8.4 for details).

sound data. This sound data may as well contain other arbitrary noises that embed the chewing events (NULL class).

A chewing structure model was used to capture the relation of chewing sequences $\mathcal{S}$ of one food type and chewing events (Figure 8.1 illustrates the concept). The chewing sequence of one food $\mathcal{S}_i$, ($\mathcal{S}_i \in \mathcal{S}$) consists of M_i chewing events $c_j^{(i)}$:

$$\mathcal{S}_i = \{c_1^{(i)} \dots c_{M_i}^{(i)}\} \tag{8.1}$$

Viewed as event, $c_j^{(i)}$ has unique temporal parameters: time of occurrence $t_{i,j}$ and duration $l_{i,j}$. The set $\mathcal{F}^+$ of feature vectors $\mathbf{f}$ was used to describe the chewing event, with: $c_j^{(i)} \rightarrow \mathbf{f}(t_{i,j}, l_{i,j})$. The sound data that embeds the event is referenced as feature set $\mathcal{F}^-$ (NULL class).

A normalised Euclidean distance function d was used to evaluate feature vector $\mathbf{f}(t,l)$ at every position t and potential duration l. The result is a distance $D(t)$, indicating the similarity of $\mathbf{f}(t,l)$ to training features derived from $\mathcal{F}^+$. A training set $\mathcal{X}_{Train}$ from $\mathcal{F}^+$ and $\mathcal{F}^-$ was used for every food to obtain training features and search bounds for l:

$$\begin{aligned} \mathcal{X}_{Train} = \quad & \{\mathcal{F}^+_{Train}, \mathcal{F}^-_{Train}\}, \\ with \quad & \mathcal{F}^+_{Train} \subset \mathcal{F}^+ \, , \mathcal{F}^-_{Train} \subset \mathcal{F}^- \end{aligned} \tag{8.2}$$

A distance threshold D_{Thres} was derived during training by evaluating the detection sensitivity on $\mathcal{X}_{Train}$. The pattern models used in this work were optimised for a high sensitivity, to retrieve at least 90% of the training set events. By applying function d and threshold D_{Thres} on test data, an estimation of chewing events, $\hat{c}_{FSS}$ was obtained. The threshold was used further to normalise distances and derive a confidence for each event $C(\hat{c}_{FSS})$ [3].

Using a fixed-time segmentation, the time between each evaluation of d was set to $\Delta t = 1/8\,\text{sec}$. While this choice limits the temporal resolution of retrieved chewing events it reduces processing requirements, compared to an evaluation for every sampled data point. The resolution was acceptable for the bite weight estimation, since temporal information below Δt was not expected. Similarly, parameter l was limited to multiples of Δt.

Cross-validation in continuous data

To account for a natural dataset variability a ten-fold cross-validation was performed to select training and validation set for FSS and all subsequent recognition steps. The dataset was partitioned into ten sections, while avoiding an intersection with a chewing sequence. For each iteration of the cross-validation, nine data sections were used for training and one for validation. Hence, each section was used once for validation.

Feature selection

A set of 264 features was selected based on chewing sound data of an earlier study [3]. The set consisted of (1) log-band spectral energy, cepstral coefficients, linear predictive coefficients (10 features each), and (2) skewness, kurtosis, and tristimulus, in total: 33 features. Laws to compute these audio features are detailed in [26]. All features mean and variance were computed for the annotated chewing event and three evenly-divided slices of each event using a sliding window of 512 samples without overlap.

To select an adapted feature subset we deployed a feature relevance and independence filtering. While this procedure was not confirmed to be particulary optimal among selection methds, it could be adapted to the FSS data description problem. For FSS, two classes (correct events and embedding data) have a large skew, consequently the selection should not consider a class prior.

In a first step, relevance of a feature $f_n \in \mathbf{f}$ was determined. A weight w was computed from the feature distribution in correct events $\mathcal{F}^+$ and embedding data $\mathcal{F}^-$, $(1 \leq n \leq |\mathbf{f}|)$:

$$w(f_n) = \left| P(f_n^+ \in \mathcal{F}_{Train}^+) - P(f_n^- \in \mathcal{F}_{Train}^-) \right| \tag{8.3}$$

The feature distribution was computed using a histogram with a bin size $N^{(1/3)}$, where N is the number of training feature vectors in $\mathcal{X}_{Train}$.

The second step refines the feature relevance ranking by using an indication for the independence between features [36]. This step aims to select a relevant feature subset while minimising redundancy. Independence I was determined from correct events ($\mathcal{F}_{Train}^+$) using Spearman's correlation coefficient ρ [21] $(1 \leq m \leq |\mathbf{f}|,\, m \neq n)$:

$$\rho(f_n, f_m) = 1 - 6 \sum_N \frac{(f_n, f_m)^2}{N(N^2 - 1)} \,, \tag{8.4}$$

$$I(f_n, f_m) = \sqrt{1 - \rho(f_n, f_m)^2} \,, \quad f_n, f_m \in \mathcal{F}_{Train}^+ \,. \tag{8.5}$$

We used an iterative scheme to select features based on [36]. Starting with the highest relevance-weighted feature, in each iteration i one feature was selected that obtains the highest combined weight w_C, when evaluated against previously selected features $f_{m1}, \ldots f_{mi}$:

$$w_C(f_n) = w(f_n) \times I\big(f_n, \{f_{m1}, \ldots f_{mi}\}\big), \; 1 < i \leq i_{Max}. \tag{8.6}$$

The procedure was terminated when a maximum number of features, specified by i_{Max}, had been selected. For this work 20 features were used.

8.3.2. Food classification and sequence voting

The FSS detection of individual foods of similar texture leads to confusions between the foods. Typically, lettuce chewing events are confused with potato chips and vice versa. This prohibits the direct inference of food type from the detection result. Consequently, we applied an additional food classification step using the detected chewing events. The food type was determined from a majority vote of all classified events in a chewing sequence. This approach is reasonable, since the food type does not change within one chewing sequence. Moreover, the weight estimation approach required the food type information for each chewing sequence only.

A nearest centroid classifier was trained based on a Fisher's linear discriminant feature transformation [11]. The dataset cross-validation and the features from the detection were reused for this classification.

8.3.3. Chewing event fusion

Temporal event overlaps, as a result of independent FSS instances for each food, were merged using a event comparison fusion (COMP). This method filters the combined event detection results and retains those events from all temporal overlaps, that have the highest confidence $C(\hat{c}_{FSS})$. COMP fusion was introduced for continuous recognition in [3].

8.4. Bite weight estimation

8.4.1. Microstructure variable extraction and relevance analysis

A set of 50 variables was extracted from the recognised microstructure of each chewing sequence. Eight basic variables were defined, as summarised in Table 8.1. This basic set was applied to each entire chewing sequence, three evenly-partitioned sections as well as the first five and the first three chewing events only. Additionally, from the first chewing event two variables were obtained: event duration ($l(i,1)$) and mean signal energy. We computed variables for sections within chewing sequences to investigate the microstructure precisely and evaluate whether recognition or estimation could be simplified by using a few chewing events only.

The correlation of each variable v_n with bite weight W was analysed using Spearman's correlation coefficient ρ, corresponding to Eq. 8.4. The correlation result of the variables was summarised in a measure of variable relevance $w_V(v_n)$ for all subjects:

$$w_V(v_n) = \sum_{Subjects} \mid \rho(v_n, W) \mid . \tag{8.7}$$

8.4.2. Food-dependent bite weight estimation

The food classification result was used in this step to select a bite weight estimation model. We deployed a multiple linear regression approach of the form:

$$\hat{W}_i = a_0 + \sum_{j=1}^{n} a_j x_{ij} \ , \tag{8.8}$$

for the bite weight prediction. The microstructure variables are represented by $x_1 \dots x_n$, n is the total number of variables included in the estimation model. The result $\hat{W}_i$ denotes estimated bite weight for chewing sequence $\mathcal{S}_i$. The coefficients $a_0, a_1, \dots a_n$ were found by a least-squares fit.

Two evaluation methods were investigated, (1) a stepwise regression fit to select informative variables, and (2) a leave-one-out weight prediction to

Table 8.1. Basic microstructure variable set. This set was computed for each chewing sequence $\mathcal{S}_i$ (see Eq. 8.1) and sections within $\mathcal{S}_i$ (see Sec. 8.4). Iterator j is used to denote all chewing events within $\mathcal{S}_i$.

Pos	Description	Law for sequence $\mathcal{S}_i$ $(\forall i : 1 \leq j \leq \hat{M}_i)$
1	Nr of chewing events	$\hat{M}_i$
2	Total chewing duration	$\sum_j l(i,j)$
3	Mean event duration	$\sum_j l(i,j)/\hat{M}_i$
4	Variance of event duration	$\sum_j (l(i,j) - \bar{l}_i)^2/\hat{M}_i$
5	Slope of event duration	$\underset{x}{\text{argmin}} \sum_j (l(i,j) - jx)$
6	Chewing speed	$\hat{M}_i / \sum_j l(i,j)$
7	Trend in chewing speed	$\underset{x}{\text{argmin}} \sum_j (1/l(i,j) - jx)$
8	Mean signal energy $\bar{e}_i$	$\sum_j e(c(i,j))$

analyse the model prediction error. For both evaluations variable subsets were preselected based on the variable correlation analysis (Section 8.4.1 above).

8.5. Results

8.5.1. Recognition of chewing

The chewing event detection performance was analysed using the metrics Precision and Recall, see [17] for a detailed description. To account a chewing event as recognised a soft-alignment procedure was applied, that allows a jitter between the detected section bounds and annotation [3]. In this work, detections with a jitter of $< 50\%$ with respect to annotation were accepted as recognised.

All individually annotated chewing events were considered as ground truth. In the detection performance analysis, events were not considered when they were reported within continuously marked chewing sequences (according to Section 8.2.3).

The FSS detection, food classification and COMP fusion steps were evaluated on testing data using the cross-validation procedure detailed in Section 8.3. The detection performance was compared to a trivial sound energy detection. For this purpose, the FSS algorithm was fed with a energy feature only. Figure 8.3 shows the food-specific precision-recall performance for all participants, at different steps of the recognition procedure as well as the sound energy detection. The performance was evaluated by applying confidence thresholds on the set of retrieved chewing events. Best performance is found towards high precision and recall. The EMG detection result in Figure 8.3 is further discussed in Section 8.5.2.

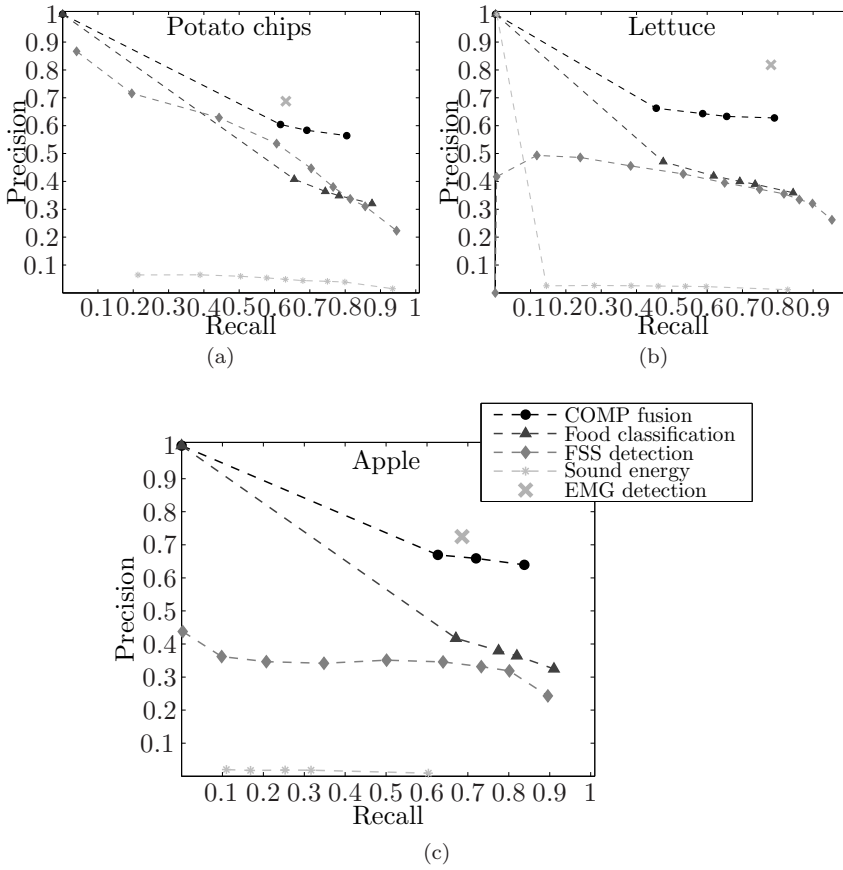


Figure 8.3. Food-specific validation performance of major recognition steps (confidence threshold sweeps): *FSS detection*, *food classification*, and *COMP fusion* in comparison to a sound energy threshold applied on the sound data. *EMG detection* shows the performance of an EMG threshold detection w.r.t. to annotation, as described in Sec. 8.5.2. Best performance is found towards the top-right corner (high precision, high recall). (a): Potato chips, (b): Lettuce, (c): Apple.

The results show good recognition performance was achieved, despite the very similar sound patterns of the selected foods. In particular the additional classification helped to refine the detection result. For all foods, a recall larger than 80% was achieved for the COMP fusion, demonstrating that chewing events can be spotted and classified from sound data. The precision was larger than 60% for all foods.

Starting from the FSS detection an increase in precision is obtained in

particular for the COMP fusion step. The performance result obtained with this approach is noticeable, since recall was only minimally depressed by the additional steps (<10% for apple and <20% for lettuce and potato chips).

As expected, the energy threshold cannot achieve a high recall and incurs many insertion errors (precision <10% in this evaluation). This weak generalisation performance was attributed to arbitrary noise in the dataset and the natural variability in chewing sound energy.

Figure 8.4 visualises the food classification and majority voting accuracy for chewing sequences. The normalised accuracy was used and any class skew was removed from the training set. The results show a very good discrimination of the foods in each sequence with an average accuracy above 90%.

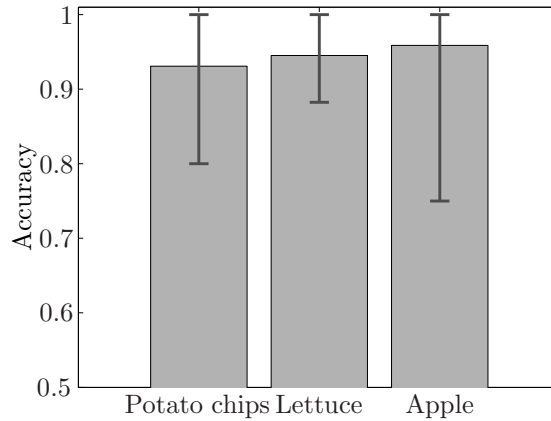


Figure 8.4. Normalised accuracy of chewing event classification and chewing sequence majority voting. Min-Max values shows participant-specific result variation.

8.5.2. Bite weight estimation

Figure 8.5 visualises the cumulative intake curves for all foods and participants. The smallest bite weights were recorded for potato chips, the largest for apples. The largest participant-specific weight variances were observed for apples.

Figure 8.3 shows the performance of sound-based recognition besides an EMG-based chewing detection. The EMG detection was obtained from the M. masseter activity, according to methods described in chewing movement investigations [5, 15]. This detection is performed by applying a threshold on the rectified and averaged EMG signal in regions annotated as chewing sequence. The threshold was set to the signal level before chewing onset plus 1 SD.

Precision of the EMG detection was less than 10% better than the final sound-based precognition performance, except for lettuce. Conversely, the

sound-based precognition method achieved at least 10% higher recall. These results indicate a similar performance of both. The weaker sound-based result for lettuce was attributed to persisting confusions of the sound-based precognition.

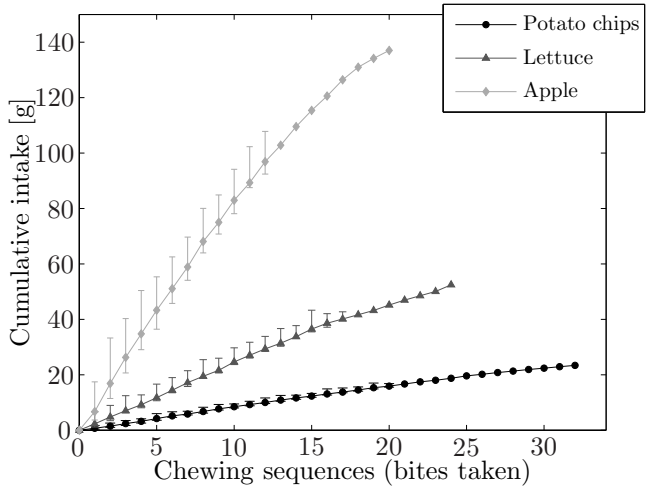


Figure 8.5. Cumulative intake curves for all foods in this study. Min-Max values show participant bite weight variation.

Variable relevance analysis

Correlations of the microstructure variables with the bite weight were analysed to determine variables that were commonly relevant in the chewing behaviour of all participants. Relevance was derived using Eq. 8.7.

Figure 8.6 shows the resulting relevance map. Highest variable relevance $w_V \geq 0.6$, hence good individual correlations, were observed for the number of chewing events and chewing duration except for potato chips. For apple the highest participant-specific correlations were observed (up to 0.96) and overall relevance of chewing event count and chewing duration were highest ($w_V \geq 0.7$). None of the further variables was consistently relevant for more than one food and participant.

A good agreement was observed between the correlations for sound-based recognition and EMG detection. EMG correlations were higher for apple in sections First 5 and First 3.

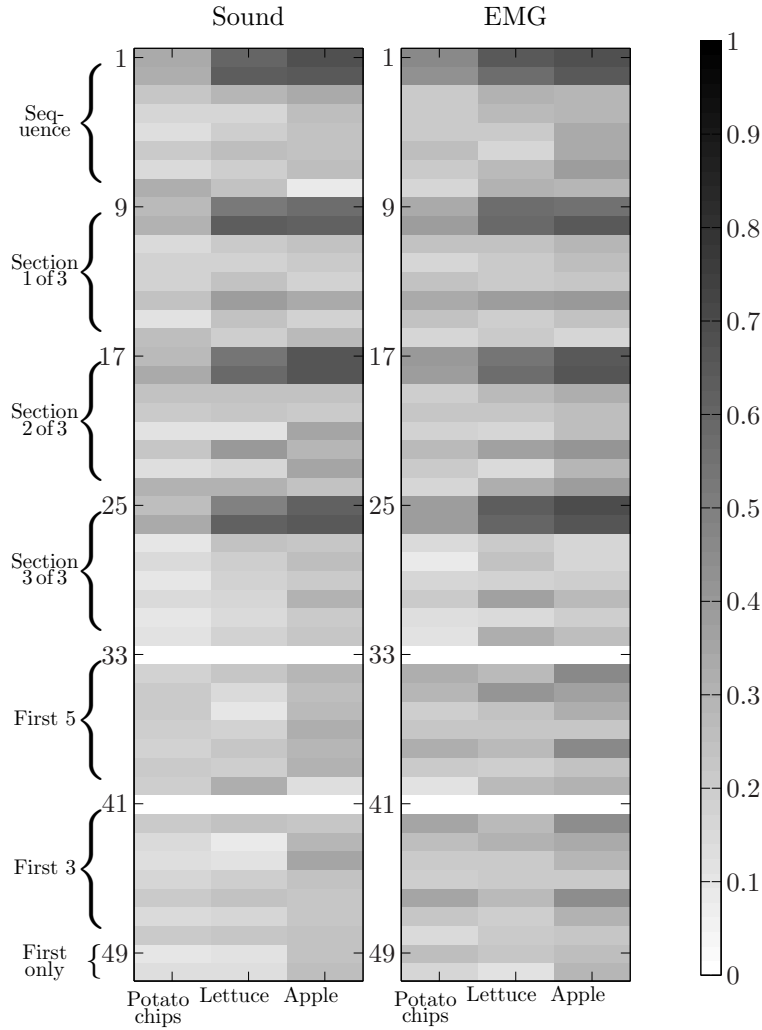


Figure 8.6. Microstructure variable relevance derived using Eq. 8.7 for sound-based recognition (left) and EMG detection (right) of all foods. According to Tab. 8.1 eight variables were derived for each of the following sections: the entire chewing sequence, three partitions (Sections 1 to 3), the first five and three chewing events. Two further variables were derived for the first chewing event only.

Stepwise variable selection

Stepwise variable exclusion was performed using linear regression fitting. The most important variables according to the relevance analysis discussed before were selected as starting model. In particular, the following variables were in-

cluded: number of chewing events and chewing duration for the entire sequence and all three sections of the sequence (in total eight variables).

Variables from the entire sequence and first section were most frequently left in the model for all foods and participants. The mid-section variables were least frequently retained. We concluded that alternatively to the entire sequence variables, the first section could be used for the weight prediction.

Leave-one-out prediction performance

The bite weight was predicted using a subset of four variables that were manually determined from related literature and confirmed by the relevance analysis. The subset contained the number of chewing events and chewing duration for the entire sequence and Section 1 of 3. This choice was made since the variables were often retained in the selection models and a small variable set was sought.

Table 8.2 shows the prediction error for both, sound-based recognition and EMG detection approaches. Moreover, the performance of an inter-individual model using sound-based and a prediction assuming a constant weight (average weight from 2nd and 3rd chewing sequence) are shown. The leave-one-out verification scheme was used for all results, except the constant weight prediction.

The constant weight prediction assumes that the bite weight does not change between the bites of a specific food and that the 2nd and 3rd sequences represent stable weight averages. Hence, the constant weight prediction marks a baseline for predictions based on chewing microstructure detections.

Overall, lowest prediction errors were achieved for the sound-based prediction for apple with an average error of $\leq$19%. This result demonstrates the effectiveness of the sound-based prediction compared to the constant weight prediction (error: 62%) and EMG (error: 28%).

For lettuce the sound-based prediction error was 31%, compared to 29% for EMG. For potato chips both sound- and EMG-based prediction incurred an error of 27-28%, confirming the low variable relevance result. As expected, the inter-individual models cannot capture the microstructure adequately. For all results, sound and EMG show a dependency on the detection performance: higher detection fidelity is reflected in lower prediction errors.

Figure 8.7 illustrates the sound-based prediction of bite weights as a cumulative weight prediction for one participant. For comparison the constant weight is shown. The sound-based prediction closely followed the actual weight, while the constant weight provides best estimates for low weight variations, such as in potato chips.

Table 8.2. Leave-one-out bite weight prediction using chewing microstructure information (participant-specific). Bite weight estimation errors are shown for sound-based recognition, EMG detection, inter-individual sound-based recognition, and for a constant weight[a].

Metric	**Foods**		
Mean (SD)	**Potato chips**	**Lettuce**	**Apple**
Sequences ($\hat{M}_i$)	26.9 (4.4)	20.0 (3.3)	14.9 (2.9)
Bite weight W [g]	0.8 (0.2)	2.3 (0.8)	7.8 (1.5)
	Sound recognition		
Absolute error [g]	0.2 (0.1)	0.6 (0.2)	1.4 (0.4)
Relative error [%]	**27.7 (9.5)**	**31.0 (5.5)**	**19.4 (4.3)**
	EMG detection		
Absolute error [g]	0.2 (0.1)	0.6 (0.2)	1.9 (1.1)
Relative error [%]	**26.5 (9.0)**	**28.9 (4.0)**	**27.8 (14.6)**
	Sound rec. (inter-individual)		
Absolute error [g]	0.2 (0.2)	0.8 (0.6)	2.3 (1.8)
Relative error [%]	**31.7 (30.6)**	**40.2 (38.2)**	**37.2 (37.1)**
	Constant weight[a]		
Absolute error [g]	0.3 (0.1)	0.9 (0.3)	3.3 (1.8)
Relative error [%]	**41.1 (25.8)**	**50.5 (29.8)**	**62.2 (33.8)**

[a]Average weight of 2nd and 3rd chewing sequence.

8.6. Discussion

8.6.1. Paper methodology

Both energy density and portion size influence energy intake independently [19]. The focus of this work was to evaluate the estimation of bite weight, being the smallest granularity of food intake. Our work builds on the relationship between chewing microstructure and bite weight. This relation is an area of ongoing research. To this end, the current work pioneers in considering natural foods, habitual consumption protocol and introducing a novel wearable measurement system.

Specific challenges faced in this investigation included the expensive data annotation, the continuous detection of chewing sound patterns and a tight acoustic relation of the foods. The latter raised the challenge to discriminate foods. As a consequence of these challenges, only a small number of foods was evaluated. The foods were chosen following our interest to monitor fruits and vegetables consumption, studying the acoustic discrimination of similar food-textures, and analysing the weight estimation from habitual bite choices of different material densities. However, we expect that the concept will be feasible for larger food groups or other specific foods.

Although expensive, the choice to manually annotate the chewing events

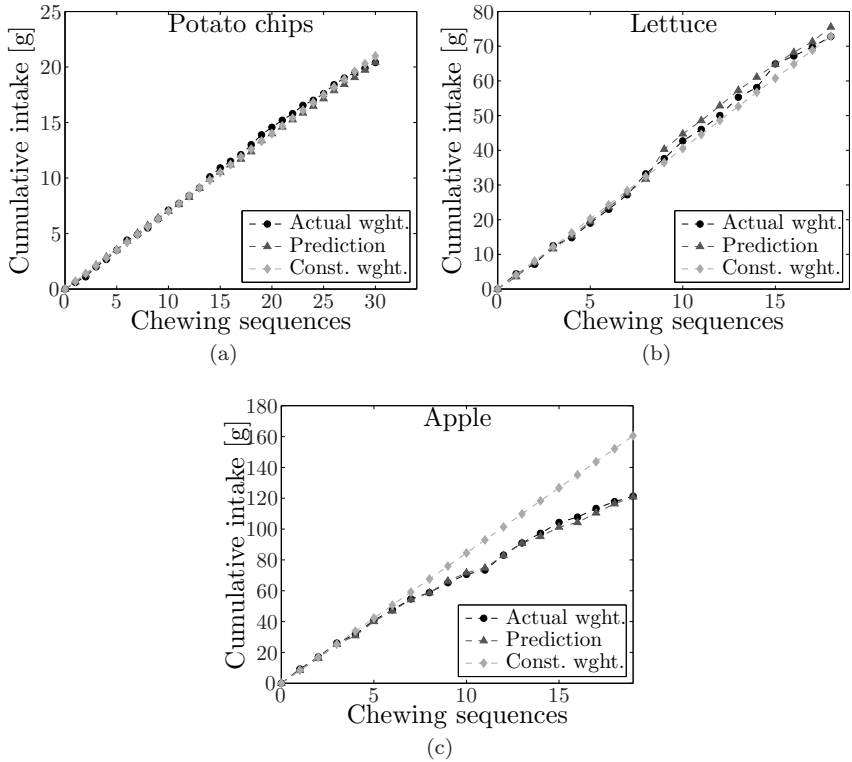


Figure 8.7. Prediction of cumulative weight intake: sound-based prediction and constant weight (average bite weight of 2nd and 3rd chewing sequence) for one subject and all foods. (a): Potato chips, (b): Lettuce, (c): Apple.

was made, to exclude potential errors from automatic chewing segmentation procedures. These were observed, e.g. for EMG-based chewing detection [18]. Nevertheless did the EMG detection perform well in our investigation.

8.6.2. Bite weight estimation

Both correlation and weight estimation results show that bite weight is not equally reflected in the chewing microstructure. Especially for low weight, such as lettuce (mean bite weight: 2.3 g) and potato chips (mean bite weight: 0.8 g), the sound-based prediction incurred larger errors (∼30%) compared to apple (error: 19.4%). We concluded that chewing behaviour does not adapt to these bite weights as it does for larger weights. This is partially confirmed by a recent study on gum chewing of different weights [33]. A 1 g gum bolus resulted in the largest within-subject variability, suggesting that the oral sensation is

less sensitive to these low weight stimuli. Consequently, the chewing pattern is less predictive for the bite weight of these foods.

In contrast, the habitual bites taken from apple were larger (mean bite weight: 8.3 g) and more varying in size. These bite weights fall into the range of 4 to 18 g. For this weight range similar boluses were found after gum chewing, indicating an adaptation of the oral management to these weights [28]. For apples, this prediction performance approaches the weight variation of the fruit itself. Hence it also reaches the reporting quality of food reports that use qualitative amount descriptions. However, further investigations are required, that focus on foods in this weight category.

The variable relevance analysis showed that individuals adapt their chewing behaviour to the food in a similar way. We observed consistent correlations of microstructure variables with bite weight, up to $0.7 - 0.96$ for apple. This is novel result for these natural foods and habitual bite selection. The result is supported by correlations found when comparing fixed volume food items and artificial foods [13].

We observed that the within-sequence chewing cycle distribution is not related to weight. This is confirmed by low correlations obtained for the microstructure variables chewing speed trend and the consistently positive correlations of all three sequence sections with weight. Moreover, the results indicates that the chewing speed variation does not dependent on bite weight.

An increase in vertical movement of the mandible was observed when bolus size increased [9, 27]. However, closing velocity and muscle activity increased as well [4, 6]. The current investigation showed that in naturally variable bite shapes, these changes did not alter microstructure variables such as chewing speed. In agreement with this result, overall chewing cycle duration was maintained with changing gum bolus sizes [6].

An restriction of the current work is that the chewing detection did not consider the mandible reopening phase. However, we do not expect to derive further supportive information from this phase.

While altering the lubrication of foods, such as buttering toast [12], lowers the total number of chews, the change in absolute numbers remained low. Deterioration, e.g. in apples adds an uncertainty in a mass density change of $\sim$10% [23].

However, if food preparation steps or deterioration modify the material structure, the acoustic food recognition would reject the food category. Practically deployed, the food recognition system may offer a selection of the most likely foods and toppings. Once a user selection was made, the corresponding food model can be used for the amount estimation. The impact of food modification and deterioration requires further investigations to derive model bounds and integrate those in recognition and weight estimation.

8.7. Conclusion

Our work is driven by the stringent need for novel automatic diet monitoring solutions that provide an estimate of consumed food amounts. We combined the acoustic recognition of chewing cycles and foods with the estimation of bite weights in this current work. Moreover, this work presented a refined recognition procedure to recognise specific foods robustly in continuous sound data. Bite weight was predicted, based on prediction models selected according to the food type recognition. The approach was evaluated quantitatively and compared to EMG-based detection and estimation results.

Food type was recognised in continuous data among three foods with very similar acoustic properties as well as various arbitrary sounds. The sound-based detection achieved a performance that closely matched the EMG-based detection (avg. precision 60%-70%). Sound-based food classification achieved a normalised accuracy of 94%.

The weight prediction results indicated that assessing bite sizes using the chewing microstructure is a feasible approach. We showed that a simple constant bite size assumption fails with prediction errors of 60%, while the sound-based recognition achieved 19.4% prediction error for apples. We observed that the detection performance is a predictor for the bite weight estimation: weaker recognition of lettuce and potato chips resulted in larger weight prediction errors.

We concluded that microstructure information is vital to estimate food amount with a weight above 4 g. The approach is applicable for solid foods, hence further work is needed to evaluate larger food sets. Our future work will also address the ambulatory evaluation of the bite weight estimation.

Bibliography

[1] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of chewing sounds for dietary monitoring. In M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, ed., *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing*, vol. 3660 of *Lecture Notes in Computer Science*, pp. 56–72. Springer Berlin, Heidelberg, September 2005. doi:10.1007/11551201_4.

[2] O. Amft and G. Tröster. Methods for detection and classification of normal swallowing from muscle activation and sound. In E. Aarts, R. Kohno, P. Lukowicz, and J. C. Trainini, ed., *PHC 2006: Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–10. ICST, IEEE digital library, November 2006. doi:10.1109/pcthealth.2006.361624.

[3] O. Amft and G. Tröster. Recognition of dietary activity events using on-body sensors. *Artif Intell Med*, 42(2):121–136, February 2008. doi:10.1016/j.artmed.2007.11.007.

[4] K. Anderson, G. S. Throckmorton, P. H. Buschang, and H. Hayasaki. The effects of bolus hardness on masticatory kinematics. *J Oral Rehabil*, 29(7):689–696, Jul 2002. doi:10.1046/j.1365-2842.2002.00862.x.

[5] I. Ashida, H. Iwamori, S.-Y. Kawakami, Y. Miyaoka, and A. Murayama. Analysis of physiological parameters of masseter muscle activity during chewing of agars in healthy young males. *J Tex Stud*, 38(1):87–99, February 2007. doi:10.1111/j.1745-4603.2007.00087.x.

[6] R. Bhatka, G. S. Throckmorton, A. M. Wintergerst, B. Hutchins, and P. H. Buschang. Bolus size and unilateral chewing cycle kinematics. *Arch Oral Biol*, 49(4):559–566, July 2004. doi:10.1016/j.archoralbio.2004.01.014.

[7] W. E. Brown, M. Shearn, and H. J. H. Macfie. Method to investigate differences in chewing behaviour in humans: Ii. use of electromyography during chewing to assess chewing behavior. *J Tex Stud*, 25(1):17–31, March 1994. doi:10.1111/j.1745-4603.1994.tb00752.x.

[8] P. H. Buschang, G. S. Throckmorton, K. H. Travers, and G. Johnson. The effects of bolus size and chewing rate on masticatory performance with artificial test foods. *J Oral Rehabil*, 24(7):522–526, Jul 1997.

[9] D. G. Daet, M. Watanabe, and K. Sasaki. Association between the interarch distance and food bolus size in the early phase of mastication. *J Prosthet Dent*, 74(4):367–372, Oct 1995. doi:10.1016/s0022-3913(05)80376-2.

[10] N. DeBelie, M. Sivertsvik, and J. DeBaerdemaeker. Differences in chewing sounds of dry-crisp snacks by multivariate data analysis. *J Sound Vib*, 266(3):625–643, September 2003.

[11] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2000.

[12] L. Engelen, A. Fontijn-Tekamp, and A. van der Bilt. The influence of product and oral characteristics on swallowing. *Arch Oral Biol*, 50(8):739–746, Aug 2005. doi:10.1016/j.archoralbio.2005.01.004.

[13] F. A. Fontijn-Tekamp, A. van der Bilt, J. H. Abbink, and F. Bosman. Swallowing threshold and masticatory performance in dentate adults. *Physiol Behav*, 83(3):431–436, Dec 2004. doi:10.1016/j.physbeh.2004.08.026.

[14] A. Giboreau, C. Dacremont, C. Egoroff, S. Guerrand, I. Urdapilleta, D. Candel, and D. Dubois. Defining sensory descriptors: Towards writing guidelines based on terminology. *Food Qual Prefer*, 18(2):265–274, March 2007. doi:10.1016/j.foodqual.2005.12.003.

[15] R. González, I. Montoya, J. Benedito, and A. Rey. Variables influencing chewing electromyography response in food texture evaluation. *Food Rev Int*, 20(1):17–32, March 2004. doi:10.1081/fri-120028828.

[16] R. J. Hill and P. S. Davies. The validity of self-reported energy intake as determined using the doubly labelled water technique. *Br J Nutr*, 85(4):415–430, Apr 2001. doi:10.1079/bjn2000281.

[17] H. Junker, O. Amft, P. Lukowicz, and G. Tröster. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recogn*, 41(6):2010–2024, June 2008. doi:10.1016/j.patcog.2007.11.016.

[18] E. K. Kemsley, M. Defernez, J. C. Sprunt, and A. C. Smith. Electromyographic responses to prescribed mastication. *J Electromyogr Kines*, 13(2):197–207, April 2003. doi:10.1016/s1050-6411(02)00065-2.

[19] T. V. E. Kral and B. J. Rolls. Energy density and portion size: their independent and combined effects on energy intake. *Physiol Behav*, 82(1):131–138, Aug 2004. doi:10.1016/j.physbeh.2004.04.063.

[20] C. Lassauzay, M. A. Peyron, E. Albuisson, E. Dransfield, and A. Woda. Variability of the masticatory process during chewing of elastic model foods. *Eur J Oral Sci*, 108(6):484–492, Dec 2000. doi:10.1034/j.1600-0722.2000.00866.x.

[21] E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks, Revised*. Prentice Hall, Englewood Cliffs, NJ, 1998.

[22] P. J. Lillford. The materials science of eating and food breakdown. *MRS Bulletin*, 25(12):38–43, December 2000.

[23] D. Mitropoulos and G. Lambrinos. "delicious pilafa" apple density changes as a quality index of mass loss degradation during storage. *J Food Quality*, 30(4):527–537, August 2007. doi:10.1111/j.1745-4557.2007.00140.x.

[24] P. M. O'Neil. Assessing dietary intake in the management of obesity. *Obes Res*, 9 Suppl 5:361S–366S; discussion 373S–374S, Dec 2001.

[25] R. Orchardson and S. W. Cadden. *The Scientific Basis of Eating*, vol. 9 of *Front Oral Biol.*, chapter Mastication, pp. 76–121. Basel, Karger, 1998. doi:10.1159/000061108.

[26] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, Ircam, France, April 2004.

[27] M. A. Peyron, K. Maskawi, A. Woda, R. Tanguay, and J. P. Lund. Effects of food texture and sample thickness on mandibular movement and hardness assessment during biting in man. *J Dent Res*, 76(3):789–795, Mar 1997.

[28] J. F. Prinz and M. R. Heath. Bolus dimensions in normal chewing. *J Oral Rehabil*, 27(9):765–768, Sep 2000.

[29] D. A. Schoeller. Limitations in the assessment of dietary energy intake by self-report. *Metabolism*, 44(2 Suppl 2):18–22, Feb 1995. doi:10.1016/0026-0495(95)90204-x.

[30] H. Shiga, C. S. Stohler, and Y. Kobayashi. The effect of bolus size on the chewing cycle in humans. *Odontology*, 89(1):49–53, Nov 2001. doi:10.1007/s10266-001-8185-0.

[31] K. R. Westerterp and A. H. C. Goris. Validity of the assessment of dietary intake: problems of misreporting. *Curr Opin Clin Nutr Metab Care*, 5(5):489–493, Sep 2002.

[32] R. R. Wing and S. Phelan. Long-term weight loss maintenance. *Am J Clin Nutr*, 82(1 Suppl):222S–225S, Jul 2005.

[33] A. M. Wintergerst, G. S. Throckmorton, and P. H. Buschang. Effects of bolus size and hardness on within-subject variability of chewing cycle kinematics. *Arch Oral Biol*, 53 (4):369–375, April 2008. doi:10.1016/j.archoralbio.2007.10.012.

[34] J. C. Witschi. Short-term dietary recall and recording methods. In W. Willett, ed., *Nutritional Epidemiology*, vol. 4, pp. 52–68. Oxford University Press, 1990.

[35] A. Woda, K. Foster, A. Mishellany, and M. A. Peyron. Adaptation of healthy mastication to factors pertaining to the individual or to the food. *Physiol Behav*, 89(1, Making Sense of Food):28–35, August 2006. doi:10.1016/j.physbeh.2006.02.013.

[36] Q. Xu, M. Kamel, and M. M. A. Salama. Significance test for feature subset selection on image recognition. In *International Conference on Image Analysis and Recognition*, vol. 3211 of *Lecture Notes in Computer Science*, pp. 244–252. Springer, 2004. doi: 10.1007/b100437.

[37] Y. Yamada, K. Yamamura, and M. Inoue. Coordination of cranial motoneurons during mastication. *Respiratory Physiology & Neurobiology*, 147(2-3):177–189, Jul 2005. doi: 10.1016/j.resp.2005.02.017.

9

Recognition of swallowing

Oliver Amft and Gerhard Tröster

Full publication title: Methods for detection and classification of normal swallowing from muscle activation and sound.
PHC 2006: Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare, ICST, 1–10, 2006.
DOI: 10.1109/PCTHEALTH.2006.361624

Abstract

Swallowing is an important part of the dietary process. This paper presents an investigation to detect and classify normal swallowing during eating and drinking from Electromyography and microphone sensors. The non-invasive sensors are selected in order to integrate them into a collar-like fabric for continuous monitoring of swallowing activity over a day. We compare methods for the detection of individual swallowing events from continuous sensor data. Furthermore we present a classifier comparison for the swallowing event properties volume and viscosity. The methods are evaluated on experimental data and a performance analysis is shown. Moreover we present a class skew analysis based on the metrics precision and recall.

9.1. Introduction

The prevalence of chronic diseases related to lifestyle and behaviour as well as the aging population leads to a surge of healthcare costs all over the world. Consequently new concepts and methods are needed to fight diseases such as obesity, hypertension and cardio-vascular diseases. It is envisioned that long-term behavioural monitoring and coaching can contribute vastly to the problem of maintaining or achieving a healthy lifestyle and therefore reducing the risks of these diseases.

Relevant lifestyle aspects related to the afore-mentioned diseases include exercise and dietary behaviour. Our work aims at developing methods to monitor dietary behaviour automatically. We believe that wearable systems can provide valuable insight into daily eating behaviour, that is difficult to achieve by other means. The work on swallowing detection presented in this paper is considered one part of a wearable dietary monitoring system, since swallowing is inherently linked to eating and drinking activities.

9.1.1. Automatic dietary monitoring

Dietary monitoring includes a variety of aspects such as timing and frequency of eating activities, rate of intake as well as type and amount of foodstuff. Information about these parameters on a daily basis provide insight into the dietary activities and can be integrated in lifestyle feedback and reminders that have a relevant impact, e.g. to maintain a lunch duration of at least 15 minutes. Currently dietary activities are studied exemplary by entering the information manually into questionnaires. This involves a considerable effort of study participants and managers.

We believe that the absolute error-free estimation of amount and calories of every possible nutrient is rather visionary, using non-invasive sensors. However, a rough estimation of food type, e.g. ratio of fluid and solid nutrient combined with the timing information, e.g. event schedule and meal durations over the day, already provides a solid basis for behavioural monitoring. Although focusing on wearable sensors we expect that additional information can be obtained in combination with a supportive environment, e.g. food products with RF-identification tags, intelligent shopping lists or dietary monitoring tables.

We target a non-invasive wearable system relying on information from the following three sensing domains: 1) the identification of characteristic arm and trunk movements associated with food intake using inertial sensors [1], 2) the analysis of food chewing sounds from an ear microphone [2] and 3) the detection of swallowing from body-worn sensors. The focus of this paper is on the latter.

9.1.2. Swallowing process

Swallowing is a frequent human activity. It is estimated that normal swallowing occurs approx. 600 to 2000 times per day in healthy persons [10].

The swallowing act is often partitioned into three distinctive phases [10]: 1) the oral preparation, 2) the pharyngeal, and finally 3) the oesophageal phase. During the oral phase a food piece is transformed to a swallowable bolus. This may involve chewing and forming a bolus by tongue movements (depending on the food texture) and initiating the swallowing reflex, which starts the pharyngeal phase. In the oral phase the bolus type is sensed with regard to volume and viscosity. Henceforth the swallowing apparatus may adapt to the bolus [4].

The pharyngeal phase is formed by the bolus travelling through the pharynx and passing the upper oesophageal sphincter. During this phase a sequence of muscle activations is used to propel the bolus and protect the trachea from contamination. The following oesophageal phase is composed of peristalsis contractions that move the bolus towards the stomach.

Since the oral phase is involved with the variable process of chewing, it is less informative for the detection of swallowing. The pharyngeal and oesophageal phase are expected to be more specific since these are not controlled voluntarily. However the latter cannot be accessed with non-invasive methods due to the spine and trachea covering the oesophagus. Hence the pharyngeal swallowing phase is addressed with non-invasive sensors.

9.1.3. Paper contributions

The work presented in this paper aims at utilising non-invasive sensing modalities to detect and identify swallowing at the pharyngeal phase. Specifically, the following contributions are made:

1. We propose sensor modalities and locations that support the identification of individual swallows and present a experimental methodology to evaluate the feasibility of these sensor types during daily activities. Moreover we address the restricted sensor positioning that stems from the goal to integrate the sensors into a collar-like fabric.

2. We compare the performance of two swallowing event detection approaches on continuous sensor data. Here, the goal is to separate the swallowing events from sensor noise incurred from everyday activities and the various other functions of the pharynx.

3. We evaluate classifiers for the discrimination of bolus volume and viscosity and present classification results that indicate the discriminative information extracted from the chosen sensor modalities.

The work presented here is a first attempt to detect and classify swallowing events automatically and evaluate different procedures. The envisioned detection system shall not hinder the user's perception and a deployment in non-clinical environments is aimed. Specifically we rely on surface electromyography (EMG) detection of muscle activation patterns and sounds associated with the swallowing event. We evaluate the different classification and event detection algorithms on recordings from 5 subjects and a total of 868 annotated swallows.

9.1.4. Related Work

A number of clinical assessment methods have been developed to analyse the complex interaction of swallowing with phonation and respiration at the throat level. The most important invasive methods include videofluoroscopy, e.g. [14], manometry, e.g. [25] and wire-electrode based EMG, e.g. [12].

A number of non-invasive assessment methods have been evaluated during pharyngeal swallowing, including sensing of muscle activations by EMG, e.g. [15, 16], listening to the throat sounds (cervical auscultation) by stethoscope [22] and stethoscope acoustic transducers or sealed microphones [6]. As alternative to the acoustic analysis, tissue vibrations have been analysed [31]. However no clear advantage of the vibration based analysis was reported, except that the vibration sensor is more robust against environmental noises at the expense of a much higher device cost.

Some works aimed at sensing the larynx movement by using movement sensors at the neck, e.g. [7]. However the detection performance is strongly depending on gender with weak results at the less prominent female larynx. Furthermore it was shown that the simple sensor incurs errors from neck and tongue movements as well as larynx movements during speaking or externally applied pressure when used during daily activities.

Several other approaches have been proposed for the analysis of swallowing, mostly in combination with previously mentioned invasive methods, including ultrasound [4, 29], pharyngeal impedance sensing methods [13, 17, 24, 32] and impedance plethsmography [23].

Different automatic feedback systems for the detection of swallowing abnormalities such as dysphagia have been proposed. Most of the abnormality detection approaches rely on EMG or vibration sensors, e.g. [20]. These systems classify the subjects based on isolated swallowing events that have been identified and marked manually by an expert.

A few attempts have been made to detect swallowing from continuous sensor readings. Pehlivan et al. [26] proposed a device for counting swallows and compared swallow counts in normal subjects and Parkinson's disease subjects during eating and drinking. The device is based on the mechanical sensing approach using a piezoelectric sensor attached to the larynx. A manual

pre-segmentation for nutrition phases was applied. Speech was specifically excluded.

Das et al. [9] deployed an ensemble of neural networks to discriminate normal and dysphagic swallows from vibration recordings. In their approach swallows were recorded in a controlled environment largely avoiding sensor artefacts. Persisting artefacts were segmented by modelling them specifically with the neural networks.

The approach of Limdi et al. [21] was based on EMG intensity detection and aimed at informing the user of elevated swallowing rates. Sukthankar, Reddy et al. [30] used EMG and vibration sensors and aimed at dysphagia rehabilitation. However both works did not present a performance evaluation of their approaches for the continuous detection problem.

The pharyngeal phase of swallowing is influenced by the type of swallowed foodstuff. During chewing and tongue movement the bolus is sensed by various receptors in the oral cavity. Specifically volume, mass and viscosity of the bolus modify the central neurological pattern generator [4, 12]. Dantas et al. [8] found that bolus transit time through the pharynx increases with viscosity. This effect was captured in duration and amplitude parameters of EMG recordings from the submental and infra-hyoid regions [8, 11, 27]. Moreover it was found in these studies that transit time and EMG features are largely unaffected by bolus volume.

Chichero et al. [5] and Boiron et al. [3] reported a dependency of swallow sound features on bolus volume. However the studies disagree on the type of interaction.

9.1.5. Paper structure

The remaining of the paper is structured as follows: Section 9.2 describes our event detection and classification approach as well as the conducted experiments. Section 9.3 summarises the results of the swallowing event detection from continuous data. Section 9.4 provides the results of the bolus type classification. Finally, Section 9.5 provides a conclusion followed by an outlook on further work.

9.2. Methodology

This section provides an overview on our approach to detect and classify swallowing. Furthermore the experimental protocol to acquire evaluation data is described.

9.2.1. Approach

As described in the introduction of this paper our detection and classification targets the analysis of pharyngeal swallowing using non-invasive sensors at-

tached to the user's neck. Fig. 9.1 illustrates the overall concept of our approach to the problems of sensor data acquisition, event detection and classification. The following sections of this paper will evaluate solutions for these problems.

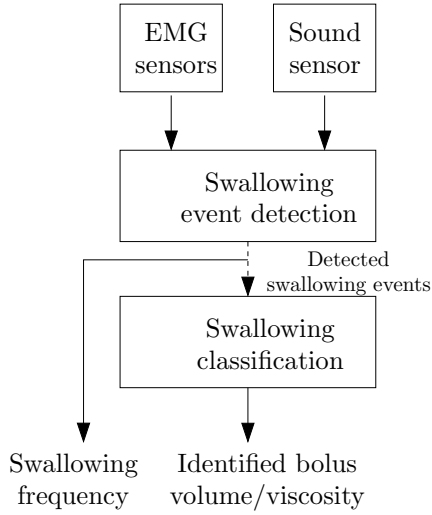


Figure 9.1. Approach to search and identify swallowing events. In this paper the swallowing classification is supported by manual annotation information.

Swallowing data acquisition is related to the problem of selecting appropriate non-invasive sensors that provide means for extracting information on swallowing events and the bolus characteristics viscosity and volume from the pharyngeal swallowing phase. Following the findings in [27] we recorded EMG to capture the viscosity variability and sound to analyse the volume/density variability of the pharynx [5]. Details of the experimental procedure and sensor placement are described in the following Section 9.2.2.

The swallowing event detection aims at extracting signal sections that contain individual swallows from a continuous stream of sensor data. Specifically the challenge is to distinguish swallowing events from sensor noise and artifacts, recorded when wearing the system during daily activities. By selecting a experimental procedure that includes non-swallowing activities, e.g. speaking, head turning, chewing, we aimed to cover these situations.

We evaluate two different methods for the swallowing event detection: 1) using a simple signal intensity measure applied to the rectified EMG amplitude and 2) using a pattern search based on a similarity measure of a data section. Here the pattern of a swallowing event is described using features derived form the sensor data. Section 9.3 presents the procedures and the evaluation results in detail.

We analyse the feasibility to discriminate bolus viscosity and volume using the sensor data in Section 9.4. Our approach is based on an isolated classification of individual voluntary swallows. For the investigation in this paper a manual annotation was applied to isolate the swallows. Fused EMG and sound feature sets from time-domain and combined frequency- and time-domain were evaluated by analysing the classification performances.

9.2.2. Experiments

Test subjects and materials

Five subjects (3 male, 2 female, aged 20 to 30 years) without known swallowing abnormalities were instructed to eat and drink different foodstuff items: 5 and 15 ml of water, a spoonful of yogurt and a piece of bread (approx. 2 cm^3). The items are summarised in Tab. 9.1. The material size was controlled by syringe for the water and visually for the spoonful of yogurt and the bread pieces. Additionally, reference samples from all foods were weighted.

The subjects were asked to aim at swallowing the nutrient items in one piece after chewing and manipulating the bolus as usual. None of the subjects expressed a dislike for any of the covered nutrients nor problems to swallow the selected bolus sizes. Subjects were sitting conveniently on a chair close to a table carrying the nutrients. They were allowed to move, chew and speak normally during the recording sessions. Naturally, during the short pharyngeal swallowing phases on speaking was audible. The environment was controlled for low and constant noise level during the swallowing events.

Sensor selection and location

Surface EMG from submental (SM-EMG) and infra-hyoid (IH-EMG) regions were recorded by gel electrodes at 24 bit, 2 kHz and bandpass filtered. Swallowing sound was recorded by an electret condenser microphone (type Sony ECM-C115), placed inferior midline from the cricoid cartilage. The microphone was secured and sealed with medical tape, following the protocol of previous investigations [5, 6, 31]. Sound data was recorded at 16 bit, 22 kHz. For the individual analysis steps the sample rate of EMG and sound was reduced.

Fig. 9.2 illustrates the positioning of the sensors. These positions have been used by previous investigations on EMG [27] and sound [5]. The SM-EMG electrode set was included in the recordings mainly for comparison and swallowing event inspection purposes.

Recording procedure

The nutrient properties are listed in Tab. 9.1. The subjects were instructed to eat/drink items from each of the nutrient categories to obtain at least 15 swallows per session. To account for physiologic variations two sessions

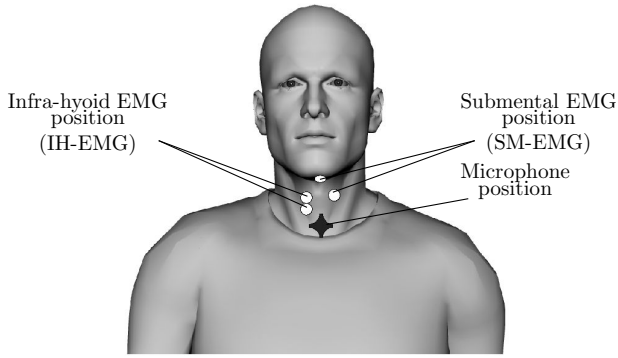


Figure 9.2. Schematic sensor positioning at the neck.

were recorded on different days. The recording duration was not constraint since the subject were eating/drinking at their individual speeds, selecting the food category for each individual swallow.

An observer was verifying the procedure during each session and annotating the food category as well as begin/end of each swallowing event. Additionally all recording sessions were videotaped for later verification of the annotated events. To simplify the online annotation, subjects were instructed to indicate swallowing to an observer by raising the hand and stop chewing shortly before swallowing. In a post-processing step all annotated events were reviewed and the begin/end times were adapted by the observer inspecting the signals. In situations were the swallowing event could not be clearly identified, the sound data was played back and/or the recorded video was analysed. In some situations spontaneous swallowing or multiple swallowing occurred before/after swallowing the food item. It was assumed that these swallows were used to clear the oral cavity and resulted in a tiny bolus or saliva swallow. These swallows were annotated as 5 ml water swallows.

To achieve a data level alignment artificial synchronisation events have been inserted on both data streams (EMG and sound) during the recordings. In the post-processing step the synchronisation events were used to adapt the alignment of the data streams.

Tab. 9.2 summarises the recorded and inspected swallowing events. To obtain a realistic data set additional data, resembling daily activities, were recorded with all subjects wearing the sensors. These activities included, speaking and conversation, background noise, head turning, tilting, nodding and chewing. In total 868 swallowing events were recorded and inspected from 4.85 hours of sensor data.

Table 9.1. Description of evaluation data: nutrients and properties.

Food	**Item volume**	**Viscosity category**	**Total swallowing events**	**Annotation duration** (mean ± SD)	**SM-EMG activation**[a] (mean ± SD)	**IH-EMG activation**[a] (mean ± SD)
Water	5 ml	fluid	353	1.8 s ± 0.5 s	0.49 ± 0.13	0.55 ± 0.17
Water	15 ml	fluid	205	2.0 s ± 0.5 s	0.55 ± 0.13	0.59 ± 0.16
Yogurt	∼7 ml (filled spoon)	semifluid	171	1.8 s ± 0.5 s	0.54 ± 0.14	0.56 ± 0.15
Bread	∼2 cm^3 (prepared pieces)	solid	139	2.0 s ± 0.6 s	0.54 ± 0.14	0.64 ± 0.17

[a]Rectified EMG signal duration above minimum+1 SD within the annotation section.

Table 9.2. Summary of evaluation data from 5 subjects.

Total swallowing events	868
Total swallowing duration	1632 s (27.2 min)
Total length of dataset	17465 s (4.85 hours)

9.3. Detection of swallowing events

In order to analyse and classify individual swallows, data sections containing the swallowing events need to be extracted from the continuous stream of sensor data.

The challenge to detect swallowing events can be formulated as follows: the envisioned system shall be continuously worn during daily activities, however swallowing events occur comparably rarely, embedded in non-swallowing phases (NULL class). A method aiming at detecting swallows shall be effective in retrieving correct events and omitting non-swallow phases while maintaining a low processing effort. The approach presented here attempts to isolate swallowing events for later analysis. Consequently the methods are optimised to reduce event misses at the expense of increased false positives. Moreover, the swallowing phases have a variable length as the event durations in Tab. 9.1 indicate.

9.3.1. Signal intensity detection

EMG signals are usually rectified and averaged for human inspection. In this way muscle contractions can be spotted visually as peaks in the waveform. We utilise a similar approach for detecting muscle contractions during swallowing events automatically: by sweeping a threshold on the rectified EMG amplitude possible events are obtained as signal sections, where the amplitude is above

the threshold. Selecting the threshold controls the system performance, e.g. the rate of false positives and false negatives.

We used the IH-EMG data at a resolution of 256 Hz. The rectified IH-EMG was obtained by averaging the absolute signal amplitude using a sliding window of 32 samples, one sample step size.

Since this method is not sensitive to the data pattern of a swallowing event, except for the signal intensity, it incurs errors and can be used to qualify the evaluation data. For the IH-EMG intensity, more detection errors correspond to more sensor artifacts from chewing and other pharyngeal activities. Furthermore the thyrohyoid muscle targeted with the infra-hyoid surface electrode position is covered by other muscle layers, disturbing the activation detection.

9.3.2. Feature similarity detection

The feature similarity approach relies on a two-step procedure of signal segmentation and similarity search. In the first step segmentation points are determined that reduce the subsequent search effort for the similarity analysis. We used the Sliding-Window And Bottom-up (SWAB) algorithm [19]. This algorithm partitions a continuous stream of sensor data very robustly by sequentially testing the approximation of the signal through linear regression lines and using the boundaries of these approximations as segments. The duration of the signal segments describe the signal variation over time, with shorter segments for highly fluctuating signals and longer segments for relatively monotone phases. We applied this algorithm to the rectified averaged IH-EMG signal (256 Hz signal resolution, mean window size of 32 samples, one sample step size). Fig. 9.3 illustrates the obtained segmentation boundaries at a sample signal.

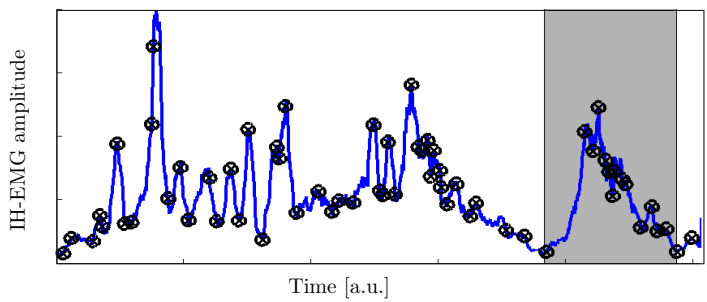


Figure 9.3. Rectified IH-EMG sample signal of yogurt eating (mean-filtered). The SWAB segmentation points are marked with ⊗. The shaded section indicates an annotated swallow.

The second step utilises the IH-EMG segmentation points to search for swallowing event sections using a feature similarity measure. The search is

performed by analysing the similarity of a data section under investigation compared to a trained pattern. For a given segmentation point, the history of sensor data is analysed from a lower up to an upper search bound. These bounds are determined in the training step from minimum/maximum overlaps between the annotated events and the segmentation points.

The similarity of the sensor data is determined from the Euclidean distance between the features of the data section under investigation during the search and the trained pattern. This approach has been applied previously to a classification problem of movement data from inertial sensors [1, 18].

The results of the feature similarity search (FSS) is a list of data sections with an associated Euclidean distance. From the training data an optimal distance threshold is determined that retains the best matching sections with the manual annotation.

The FSS procedure was applied to features from IH-EMG and sound data individually and combined using feature-level fusion. The features from feature set 1 (time domain, see Tab. 9.5) were used for the evaluation of the similarity searches. With regard to the potentially low mobile processing performance a low data resolution was used for both IH-EMG (128 Hz) and sound (4 kHz).

Furthermore two event fusion methods were tested: 1) a comparison of the individual IH-EMG and sound event detections and 2) a second-pass similarity search.

The comparison of sensor-specific event detections aims at selecting the front of best events from the individual similarity searches. For this procedure a detection confidence was determined by normalising the sensor-specific event distances with the corresponding similarity training threshold. In this way the event detection results of independent similarity searches can be compared. The best events were selected by a sliding window procedure.

For the fusion using a second-pass similarity search we applied an additional training step based on the event confidences of the individual similarity results. The training data from the first-pass similarity was reused for this training. The confidence was determined in the same way as for the comparison method.

9.3.3. Evaluation procedure

Training and testing was performed on the subject-specific data sets. To account for variations in the data set a 4-fold cross-validation procedure was used to determine training and testing data set for both detection procedures, IH-EMG intensity and FSS. For the training 3 of 4 data parts were used. Evaluation was performed on the left out data part. This procedure was repeated until all 4 parts were used for testing once. The partition boundaries were adapted to avoid intersecting swallowing data sections.

To analyse performance, we utilised the metrics *Precision* and *Recall* commonly used for evaluation in Information Retrieval. These metrics are derived as follows:

$$Recall = \frac{TP}{P} = \frac{\text{Recognised swallows}}{\text{Relevant swallows}} \tag{9.1}$$

$$Precision = \frac{TP}{TP + FP} = \frac{\text{Recognised swallows}}{\text{Retrieved swallows}} \tag{9.2}$$

Relevant swallows corresponds to the manually annotated number of swallowing events in a class (positives, P). Retrieved swallows represents the number of swallowing events that are returned by the algorithm. This includes both, true positives (TP) and false positives (FP). Finally, recognised swallows refers to the correctly returned number of swallowing events (true positives, TP).

9.3.4. Detection results

The results of all investigated detection methods are summarised in Tab. 9.3: IH-EMG intensity (Intensity), FSS for IH-EMG, sound and the feature level fusion, as well as the event fusion methods comparison (COMP) and second-pass. For all methods a threshold was chosen to achieve high recall. For the comparison both recall and precision must be considered.

Table 9.3. Summary for the subject-specific detection performance.

	Intensity	**FSS**	**FSS**	**FSS**	**COMP**	**2nd pass**
Sensors	IH-EMG	IH-EMG	SND	IH-EMG &SND	IH-EMG &SND	IH-EMG &SND
Relevant	868	868	868	868	868	868
Retrieved	8065	4128	4368	3961	1853	1660
Recognised	645	715	634	726	567	491
FN	223	153	234	142	301	377
FP	7420	3413	3734	3235	1286	1169
Recall	0.74	0.82	0.73	0.84	0.65	0.57
Precision	0.08	0.17	0.15	0.18	0.31	0.30

IH-EMG intensity

An overall recall of 0.74 was achieved on the evaluation data. However the method retrieves many false positives (low precision value). The weak detection result can be accounted to the fluctuating signal with high amplitude values for arbitrary muscle contractions and artifacts.

Feature similarity

The SWAB algorithm obtained 76803 segments for the 4.85 hours of evaluation data. The jitter between segmentation boundary and manual swallowing

event annotations were analysed for all food categories. The mean jitter was below 0.12 s (SD: 0.11 s) for all 868 events.

An overall recall of 0.82, 0.73 and 0.84 was achieved for IH-EMG, sound and the feature-level fusion respectively. However FSS retrieved far less false positive errors (higher precision value), compared to the IH-EMG intensity method. This is illustrated in the threshold sweep of Fig. 9.4. While the intensity method reaches an acceptable recall level, the precision does not increase above 0.1.

The precision-recall comparison in Fig. 9.4 furthermore presents the modelling performance of the two event fusion methods: similarity comparison (COMP) and second-pass similarity. While both methods retrieve far less false positives (increased precision), the number of recognised swallowing events decrease, when compared to the sensor-specific similarity searches. In the depicted example, the second-pass similarity provides (as intended) a far better result, resembling the performance of the best sensor-specific similarity result while the comparison method incurs more recognition errors.

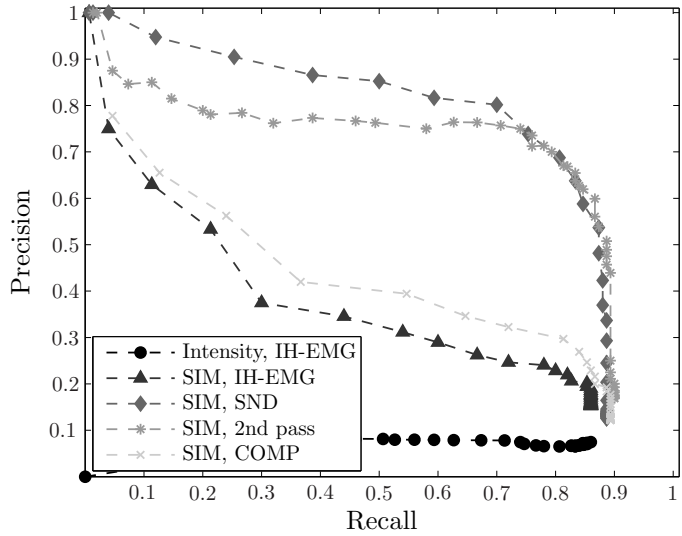


Figure 9.4. Precision-recall comparison (threshold sweep) of the IH-EMG intensity and FSS methods of a trained data section containing 150 relevant swallows from one subject. Best performance is found towards the top-right corner (high precision, high recall).

9.4. Classification of swallowing events

As reviewed in the introduction, different interactions of bolus volume and viscosity with EMG and sound features have been tested in the clinical settings of previous studies. In this section we evaluate the following hypotheses: 1) IH-EMG supports the discrimination of the bolus viscosity independent from volume and 2) sound supports the discrimination of the bolus volume. Our analysis is based on the isolated classification using manually derived swallowing annotation.

9.4.1. Evaluation procedure

In order to investigate the hypotheses described above the recorded food categories were grouped into classes according to Tab. 9.4. We assumed here that the chosen foodstuffs represent the typical variations in foods with regard to viscosity (fluid, semifluid and non-fluid) as well as volume (see Tab. 9.1).

Table 9.4. Class groupings of nutrient categories.

Hypothesis		Food group classes
Volume	1	Class 1: 5 ml water, 2 cm^3 bread pieces Class 2: spoonful yogurt Class 3: 15 ml water
	2	Class 1: 2 cm^3 bread pieces, 5 ml water, spoonful yogurt Class 2: 15 ml water
Viscosity	1	Class 1: 5 ml+15 ml water Class 2: spoonful yogurt Class 3: 2 cm^3 bread pieces
	2	Class 1: 5 ml+15 ml water Class 2: spoonful yogurt, 2 cm^3 bread pieces

Two feature sets (summarised in Tab. 9.5) were computed from the sensor data. Feature set 1 is based on time domain properties of the sensor streams, feature set 2 contains set 1 and additional frequency domain features. Feature set 1 was processed at relatively low sampling resolution of 128 Hz for EMG and 4 kHz for sound. For feature set 2, higher frequencies were used: 2 kHz for EMG and 16 kHz for the sound.

The feature sets were evaluated using three classifiers of different complexity: 1) Naive Bayes (NB), 2) k-Nearest Neighbour (KNN) and 3) Hidden Markov Models (HMMs). For NB a feature pre-processing using Linear Discriminant Analysis (LDA) was applied, LDA+NB. The LDA filter method permitted the integration of larger feature sets and improved the discrimination performance in some situations as described below.

Table 9.5. Feature sets for EMG and sound data.

Set	Feature description
Set 1	Absolute sum of 4 partitions, sum of pos/neg. deviation, Absolute sum of signal greater than 1SD, Length of absolute signal greater than 1SD, Peak count, peak distance, peak maximum
Set 2	All features from feature set 1, Spectral power, bandwidth, Centre of gravity, roll-off point, Spectral fluctuation, Sum of lin. band energy (4 bands), Log. band energy (4 bands),

For the KNN classifier $k = 10$ was chosen, however only a minimal performance degradation was observed for $k = 5$. For the HMMs, continuous left-right models with 5 states were used for each class with one Gaussian mixture per feature. Continuous features were derived according to feature set 2. To reduce the influence of the varying training performance, 10 instances of each HMM were trained and tested on the training data. The best performing set was used for the evaluation.

Training and testing was performed on the subject-specific data sets. To account for variations in the partitioning of classifier training and testing data set a 10-fold cross-validation procedure was used. For training 9 of 10 parts of all instances were used. Evaluation was performed on the left out data part in order to test every instance exactly once.

The chosen nutrient groups resulted in class skews (one class contained more instances than another class). To avoid training a skewed classifier an equal number of training instances was used for all classes and the test instances were adapted to satisfy the cross-validation procedure as described before.

To compare the classification results the normalised accuracy was used:

$$\textit{Normalised accuracy} = \frac{1}{2} * \left(\frac{TP}{P} + \frac{TN}{N} \right). \tag{9.3}$$

The normalised accuracy is robust against skew with a given (trained) classifier [28]. In our evaluation, the classifiers were trained with an equal class distribution. The measures are derived from the two-class confusion matrix as seen from one class: true positives (TP), all positives (P), true negatives (TN) and all negatives (N).

Class skew analysis

We present a performance analysis that incorporates the class skew based on class-wise precision and recall metrics. The class skew analysis simulates different class distributions using the classification result of the full evaluation dataset. The procedure starts with all relevant instances from class 1 and adds instances from the second class sequentially. This procedure is repeated for class 2 by stepwise removing instances from class 1. For each class distribution precision and recall was computed. The results are presented in the following class skew plots. Precision and recall were derived for each class in the same way as for the event detection described in Section 9.3.

9.4.2. Classification results

The class distribution for three volume and viscosity categories as presented in hypotheses 1 for volume and viscosity respectively (see Tab. 9.4), performed weak on all tested classifiers, sensor streams and feature sets. Therefore we concentrated on the evaluation of hypotheses 2 (classes for low and high volume/viscosity).

Volume classification

Fig. 9.5 illustrates the classification performance using the class skew precision-recall plot procedure as described before. The midpoint of each curve shows the performance for the class distribution in the evaluation data set. A natural distribution may be found to contain a large variation in swallowing volume, depending on nutrient, taste and physiology. According to the actual distribution the classifiers produces a result along the curves. Best performance is found towards the top-right corner (high precision, high recall).

The classification result of LDA+NB using EMG and sound (individually and by feature-level fusion) and from one KNN is shown in Fig. 9.5. For these results feature set 1 was utilised. Fig. 9.6 shows a comparison of the different classifiers using feature set 2. The best performing LDA+NB classifier from Fig. 9.5 is shown for reference. Overall the LDA+NB procedure with features from feature set 1 (time domain) performs marginally better than the KNN using feature from set 2. The HMMs did not improve the recognition rate compared to the best LDA+NB.

From the graphs it can be seen that the sound data contributes largely to the discrimination result while the individual IH-EMG or combined IH-&SM-EMG classification performs relatively less using LDA+NB. Best results are obtained from the feature-level fusion of IH-EMG and sound. Although more complex, feature set 2 did not improve the result. The classification performances for the bolus volume are summarised in Tab. 9.6 using the normalised accuracy metric.

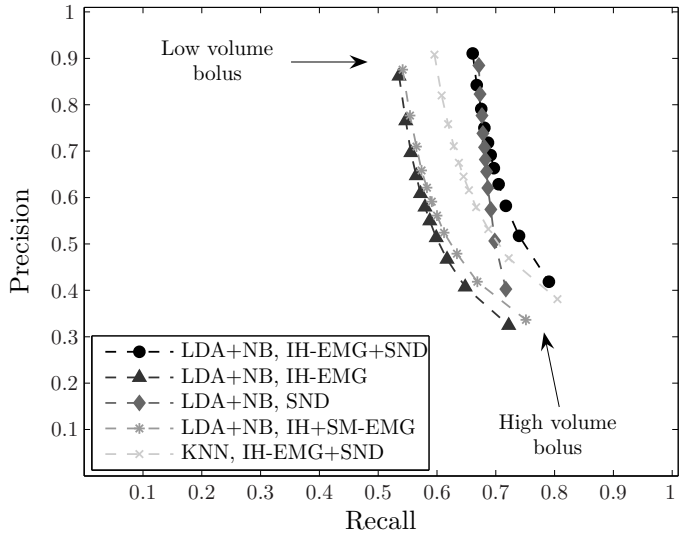


Figure 9.5. Class skew plot of the low vs. high volume bolus classification result using LDA+NB and KNN. Best performance is found towards the top-right corner (high precision, high recall).

Viscosity classification

Fig. 9.7 illustrates the classification performance using the class skew precision-recall plot procedure. Similar to the volume analysis, the midpoint of each curve shows the performance for the ratio between low and high viscosity in the evaluation data set. A natural distribution may be found to contain more low viscosity swallows than obtained in the experiments of this investigation (see Tab. 9.1). Using the non-skewed classifiers this would shift the result towards higher precision at a reduced recall.

The classification result of LDA+NB using EMG and sound (individually and by feature-level fusion) and from one KNN is shown in Fig. 9.7. For these classification results feature set 1 was utilised. The evaluation of the different classifiers using feature set 2 revealed a KNN using IH-EMG and sound features as best-performing classifier for low viscosity swallows. Similar to the volume analysis, the LDA+NB classifier performed well with feature set 1 using IH-EMG and sound. The 5-state HMMs did not improve the recognition rate compared to the best the KNN and LDA+NB classifiers.

The classification performances for the bolus viscosity classification are summarised in Tab. 9.7 using the normalised accuracy metric.

Table 9.6. Performance summary for low vs. high volume bolus classification.

	LDA+NB		**LDA+NB**		**LDA+NB**		**LDA+NB**		**KNN**		**HMM**		**HMM**		**KNN**		**KNN**	
Sensors	IH-EMG &SND		IH-EMG		SND		IH-&SM-EMG		IH-EMG &SND		IH-EMG &SND		SND		IH-EMG &SND		SND	
Feature set	1		1		1		1		1		2		2		2		2	
Class	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Relevant	663	205	663	205	663	205	663	205	663	205	663	205	663	205	663	205	663	205
Retrieved	481	387	412	456	503	365	410	458	435	433	399	469	432	436	416	452	453	415
Recognised	438	162	355	148	445	147	359	154	395	165	373	179	374	147	384	173	404	156
FN	225	43	308	57	218	58	304	51	268	40	290	26	289	58	279	32	259	49
FP	43	225	57	308	58	218	51	304	40	268	26	290	58	289	32	279	49	259
Norm. acc.	0.73		0.63		0.69		0.65		0.70		0.72		0.64		0.71		0.69	

Table 9.7. Performance summary for low vs. high viscosity bolus classification.

	LDA+NB		**LDA+NB**		**LDA+NB**		**LDA+NB**		**KNN**		**HMM**		**HMM**		**KNN**		**KNN**	
Sensors	IH-EMG &SND		IH-EMG		SND		IH-&SM-EMG		IH-EMG &SND		IH-EMG &SND		SND		IH-EMG &SND		SND	
Feature set	1		1		1		1		1		2		2		2		2	
Class	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Relevant	558	310	558	310	558	310	558	310	558	310	558	310	558	310	558	310	558	310
Retrieved	448	420	475	393	424	444	466	402	482	386	402	466	274	594	429	439	404	464
Recognised	377	239	353	188	341	227	365	209	408	236	344	252	215	251	376	257	333	239
FN	181	71	205	122	217	83	193	101	150	74	214	58	343	59	182	53	225	71
FP	71	181	122	205	83	217	101	193	74	150	58	214	59	343	53	182	71	225
Norm. acc.	0.72		0.62		0.67		0.66		0.75		0.72		0.60		0.75		0.68	

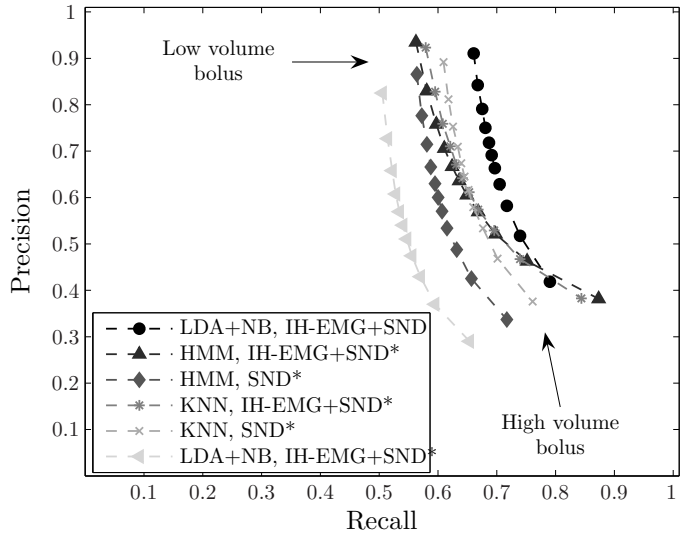


Figure 9.6. Class skew plot of the low vs. high volume bolus classification result using KNN and HMM. Classifiers using feature set 2 are marked (*). The best performing LDA+NB classifier from Fig. 9.5 is shown for reference. Best performance is found towards the top-right corner (high precision, high recall).

9.5. Discussion and conclusions

The work presented in this paper aimed at 1) detecting individual swallowing events in continuous data from EMG and sound and 2) classifying swallows regarding volume and viscosity properties.

9.5.1. Swallowing detection

For the detection of swallowing events from continuous data two approaches were presented: signal intensity thresholding and a feature similarity search. The method based on the signal intensity threshold recalled the swallowing events well (recall: 0.7), at the expense of high false positive errors (precision: 0.08). Comparably, the evaluated feature similarity methods retrieved almost half of the false positive errors, while achieving a similar recall.

The feature similarity search based on the IH-EMG signal performed better than using sound (recall and precision). However the overall result of sound is acceptable considering that the IH-EMG segmentation was used. The feature-level fusion of IH-EMG and sound similarity searches marginally improved the detection result, compared to the IH-EMG search alone.

To further improve the detection two event fusion methods were developed and tested. Both methods improved the precision clearly. However this was

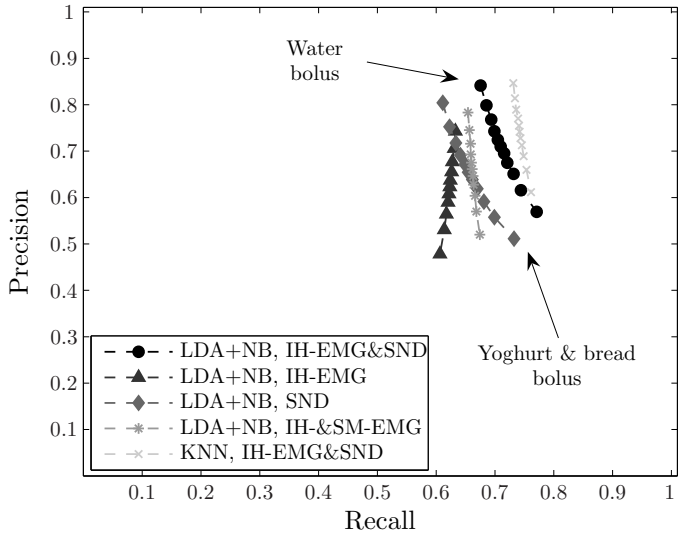


Figure 9.7. Class skew plot of the low vs. high viscosity bolus classification result using LDA+NB and KNN. Best performance is found towards the top-right corner (high precision, high recall).

achieved at the expense of a reduced recall for both methods. The second-pass similarity algorithm aimed at combining the best results from the sensor-specific searches. Although a good training performance was achieved the method failed to generalise on the test data.

In conclusion, both the IH-EMG and the sound provide important information for the swallowing event detection. The feature similarity based approach to detect swallows is clearly advantageous compared to the signal intensity method.

9.5.2. Swallowing classification

Two independent classification strategies for individual swallows were analysed: classification of bolus volume and classification of bolus viscosity. Initially the evaluated foodstuffs were grouped into three classes. However the classification result was very weak, indicating that no appropriate discriminative power was found in the sensor data and the chosen features. Therefore we concentrated on the discrimination among two classes of low and high volume as well as viscosity.

This classification revealed that the sound provides important information for volume as well as viscosity discrimination. This was expected from our initial hypothesis for the volume only. The classification result from EMG alone

was weak for both, volume and viscosity classification from the infra-hyoid and the submental positions. Best result were achieved from a feature-level fusion of IH-EMG and sound data.

We found that the combination of LDA+NB classifier performed well given the simpler time-domain feature set. This set included static features aimed at modelling the temporal pattern of the sensor data by partitioning the complete swallow into segments. These features improved the classification result. Although without LDA, the KNN classifier performed well in the evaluation. The HMMs reached the recognition rate of the best performing static classifiers.

In conclusion, a recognition rate of 0.73 to 0.75 was achieved for the volume and viscosity classifications. Although this is not an ideal performance we believe that it contributes to the envisioned dietary monitoring system. A tentative classification on individual swallowing events can be integrated since the system will be worn for entire meal consumption sessions.

9.6. Further work

From the results achieved in this work the following goals for future investigations can be derived:

1. Testing the methods on data from further subjects and additional nutrients to evaluate the robustness of the system and to verify the current findings regarding the classification of bolus volume and viscosity.

2. Evaluating the use of further sensors to improve the detection performance.

3. Studying the detection performance of double-swallowing and sequential swallowing specifically.

Bibliography

[1] O. Amft, H. Junker, and G. Tröster. Detection of eating and drinking arm gestures using inertial body-worn sensors. In B. Rhodes and K. Mase, ed., *ISWC 2005: IEEE Proceedings of the Ninth International Symposium on Wearable Computers.*, pp. 160–163. IEEE Press, October 2005. doi:10.1109/iswc.2005.17.

[2] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of chewing sounds for dietary monitoring. In M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, ed., *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing*, vol. 3660 of *Lecture Notes in Computer Science*, pp. 56–72. Springer Berlin, Heidelberg, September 2005. doi:10.1007/11551201_4.

[3] M. Boiron, P. Rouleau, and E. H. Metman. Exploration of pharyngeal swallowing by audiosignal recording. *Dysphagia*, 12(2):86–92, 1997.

[4] G. Chi-Fishman and B. C. Sonies. Effects of systematic bolus viscosity and volume changes on hyoid movement kinematics. *Dysphagia*, 17(4):278–287, 2002. doi:10.1007/s00455-002-0070-7.

[5] J. A. Y. Cichero and B. E. Murdoch. Acoustic signature of the normal swallow: characterization by age, gender, and bolus volume. *Ann Otol Rhinol Laryngol*, 111(7 Pt 1):623–632, Jul 2002.

[6] J. A. Y. Cichero and B. E. Murdoch. Detection of swallowing sounds: methodology revisited. *Dysphagia*, 17(1):40–49, 2002.

[7] C. Danbolt, P. Hult, L. T. Grahn, and P. Ask. Validation and characterization of the computerized laryngeal analyzer (CLA) technique. *Dysphagia*, 14(4):191–195, 1999.

[8] R. O. Dantas, M. K. Kern, B. T. Massey, W. J. Dodds, P. J. Kahrilas, J. G. Brasseur, I. J. Cook, and I. M. Lang. Effect of swallowed bolus variables on oral and pharyngeal phases of swallowing. *Am J Physiol*, 258(5 Pt 1):G675–G681, May 1990.

[9] A. Das, N. P. Reddy, and J. Narayanan. Hybrid fuzzy logic committee neural networks for recognition of swallow acceleration signals. *Comput Methods Programs Biomed*, 64(2):87–99, Feb 2001.

[10] D. M. Denk, H. Swoboda, and E. Steiner. Physiology of the larynx. *Radiologe*, 38(2):63–70, Feb 1998. In German.

[11] R. Ding, J. A. Logemann, C. R. Larson, and A. W. Rademaker. The effects of taste and consistency on swallow physiology in younger and older healthy individuals: a surface electromyographic study. *J Speech Lang Hear Res*, 46(4):977–989, Aug 2003.

[12] C. Ertekin and I. Aydogdu. Neurophysiology of swallowing. *Clin Neurophysiol*, 114(12):2226–2244, Dec 2003.

[13] H. Firmin, S. Reilly, and A. Fourcin. Non-invasive monitoring of reflexive swallowing. In *Speech, Hearing and Language: work in progress*, vol. 10. Department of Phonetics and Linguistics, University College London (UCL), 1997.

[14] J. Gates, G. G. Hartnell, and G. D. Gramigna. Videofluoroscopy and swallowing studies for neurologic disease: a primer. *Radiographics*, 26(1):e22, 2006. doi:10.1148/rg.e22.

[15] V. Gupta, N. P. Reddy, and E. P. Canilang. Surface EMG measurements at the throat during dry and wet swallowing. *Dysphagia*, 11(3):173–179, 1996.

[16] M.-L. Huckabee, S. G. Butler, M. Barclay, and S. Jit. Submental surface electromyographic measurement and pharyngeal pressures during normal and effortful swallowing. *Arch Phys Med Rehabil*, 86(11):2144–2149, Nov 2005. doi:10.1016/j.apmr.2005.05.005.

[17] H. Imam, S. Shay, A. Ali, and M. Baker. Bolus transit patterns in healthy subjects: a study using simultaneous impedance monitoring, videoesophagram, and esophageal manometry. *Am J Physiol Gastrointest Liver Physiol*, 288(5):G1000–G1006, May 2005. doi:04.

[18] H. Junker, P. Lukowitz, and G. Tröster. Continuous recognition of arm activities with body-worn inertial sensors. In *ISWC 2004: Proceedings of the Eighth International Symposium on Wearable Computers*, pp. 188–189, 2004.

[19] E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time series. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 289–296, 2001.

[20] L. Lazareck and Z. Moussavi. Classification of normal and dysphagic swallows by acoustical means. *IEEE Trans Biomed Eng*, 51(12):2103–2112, Dec. 2004. doi:10.1109/tbme.2004.836504.

[21] A. Limdi, M. McCutcheon, E. Taub, W. Whitehead, and I. Cook, E.W. Design of a microcontroller-based device for deglutition detection and biofeedback. In *EMBS 1989: Proceedings of the Annual International Conference of the IEEE Engineering in Engineering in Medicine and Biology Society.*, vol. 5, pp. 1393–1394. IEEE Press, November 1989. doi:10.1109/iembs.1989.96257.

[22] W. J. Logan, J. F. Kavanagh, and A. W. Wornall. Sonic correlates of human deglutition. *J Appl Physiol*, 23(2):279–284, Aug 1967.

[23] A. Moreau-Gaudry, A. Sabil, G. Benchetrit, and A. Franco. Use of respiratory inductance plethysmography for the detection of swallowing in the elderly. *Dysphagia*, 20(4):297–302, 2005. doi:10.1007/s00455-005-0031-z.

[24] T. I. Omari, N. Rommel, M. M. Szczesniak, S. Fuentealba, P. G. Dinning, G. P. Davidson, and I. J. Cook. Assessment of intraluminal impedance for the detection of pharyngeal bolus flow during swallowing in healthy adults. *Am J Physiol Gastrointest Liver Physiol*, 290(1):G183–G188, Jan 2006. doi:10.1152/ajpgi.00011.2005.

[25] J. B. Palmer, J. C. Drennan, and M. Baba. Evaluation and treatment of swallowing impairments. *Am Fam Physician*, 61(8):2453–2462, Apr 2000.

[26] M. Pehlivan, N. Yuceyar, C. Ertekin, G. Celebi, M. Ertas, T. Kalayci, and I. Aydogdu. An electronic device measuring the frequency of spontaneous swallowing: digital phagometer. *Dysphagia*, 11(4):259–264, 1996.

[27] J. L. Ruark, G. H. McCullough, R. L. Peters, and C. A. Moore. Bolus consistency and swallowing in children and adults. *Dysphagia*, 17(1):24–33, 2002.

[28] M. Stäger, P. Lukowicz, and G. Tröster. Dealing with class skew in context recognition. In *IWSAWC 2006: 6th International Workshop on Smart Appliances and Wearable Computing*, Jul 2006.

[29] D. Stevens. Ultrasound swallow. *Br Med J*, 2(6154):1789–1790, 1978.

[30] S. M. Sukthankar, N. P. Reddy, E. P. Canilang, L. Stephenson, and R. Thomas. Design and development of portable biofeedback systems for use in oral dysphagia rehabilitation. *Med Eng Phys*, 16(5):430–435, Sep 1994.

[31] K. Takahashi, M. E. Groher, and K. Michi. Methodology for detecting swallowing sounds. *Dysphagia*, 9(1):54–62, 1994.

[32] Y. Yamamoto, T. Nakamura, Y. Seki, K. Utsuyama, K. Akashi, and K. Jikuya. Neck electrical impedance for measurement of swallowing. *El Eng Jp*, 130(4):35–44, Jan 2000.

10

Probabilistic parsing of dietary activity events

Oliver Amft, Martin Kusserow and Gerhard Tröster

BSN 2007: Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks, IFMBE Proceedings Vol. 13, Springer, 242–247, 2007.
DOI: 10.1007/978-3-540-70994-7_41

Abstract

Dietary behaviour is an important lifestyle aspect and directly related to long-term health. We present an approach to detect eating and drinking intake cycles from body-worn sensors. Information derived from the sensors are considered as abstract activity events and a sequence modelling is applied utilising probabilistic context-free grammars. Different grammar models are discussed and applied to dietary intake evaluation data. The detection performance for different foods and food categories is reported. We show that the approach is a feasible strategy to segment dietary intake cycles and identify the food category.

10.1. Introduction

Nutrition is a key aspect of our everyday life and health. While pure over-consumption in time frames of months and years leads to the predominant overweight and obesity, many other forms of malnutrition exist. Often malnutrition is a confounding factor for developing chronic illnesses. Since nutrition is related to daily living behaviour, modifying eating behaviour requires changing lifestyle.

Besides caloric value, nutrition behaviour includes a variety of aspects such as duration and frequency of eating and drinking activities, rate of intake as well as the type of food itself. Information about these parameters on a daily basis provide insight into the dietary activities and can be integrated in lifestyle coaching, e.g. reminders to maintain a lunch duration of at least 15 minutes.

Our work aims at developing methods to monitor dietary behaviour automatically using wearable systems. In this paper we present an approach to infer eating and drinking activity as well as food categories from activity events derived in three on-body sensing domains.

10.1.1. Automatic dietary monitoring

We expect that by utilising wearable systems useful assistive systems for dietary monitoring are conceivable. Such systems could provide a rough estimate on the food consumption and could provide valuable insight into daily eating behaviour. This includes a rough estimation of food type, e.g. ratio of fluid and solid nutrient combined with the timing information, e.g. event schedule and meal durations over the day.

We target non-invasive wearable systems relying on information from the following three sensing domains: (1) the identification of characteristic arm and trunk movements associated with food intake using inertial sensors [1], (2) the analysis of food chewing sounds from an ear microphone [2] and (3) the detection of swallowing from collar-worn sensors [3]. These sensing domains are modelled as activity event sources by appropriate continuous pattern detectors. These events constitute the input for the event sequence detection presented in this work.

10.1.2. Decomposition of hierarchical activity

While many human actions may not be feasibly sensed and modelled as a whole, they can be described as a hierarchical activity process. Consequently, such an activity process is composed of separate sub-activities, often aligned in a sequential order. Given that a sufficient abstraction was chosen, patterns of identified sub-activities can be recognised from sensor data. An example for such an activity consisting of a sub-activity event sequence are dietary intake cycles. These cycles consist of movements to prepare a food piece and

manoeuvre it to the mouth, e.g. using fork and knife, chewing the food with multiple closing and opening cycles of the jaw and eventually swallowing the food bolus. Usually several intake cycles are used to consume a food product or meal. The combination of these sub-activities in their correct order forms the superior activity *eating*.

Sequences of sub-activities that are linked to form a meaningful action suggest the analogy to linguistic terms, e.g. words (=sub-activity events) and sentences (=action, consisting of sub-activity event sequences). Following the example of an intake cycle described above, a syntax is given by the fact that foods may be chewed and swallowed only after they have been prepared and moved to the mouth. We hypothesise that sub-activities follow a grammatical structure and henceforth can be interpreted as computable language. Given that this hypothesis holds, the high-level segmentation of intake cycles can be achieved and moreover, structure parameter such as number of chews and food category estimates per intake cycle become available.

The detection of the event sequences, in linguistic terms the parsing of symbols, has to deal with the following main problems: (1) the input sequence may not follow the assumed language syntax in all situations and (2) the input sequence may be partially incorrectly detected by the event pattern detectors. Both problems violate the applied grammar and a standard language parser would simply give up. Obviously the applied parsing method and grammar has to cope with such situations, however accounting for the violations. As a solution a probabilistic context-free grammar (PCFG) parser is used in this work.

10.1.3. Probabilistic parsing of activity events

A grammar G can generally be described by $G = (T, NT, P, S)$. Here T is a set of terminal symbols, NT is a set of non-terminal helper symbols, P is a set of production rules of the grammar and S is the start symbol.

The prototype production rule of a context-free grammar is described in Eq. 10.1. These production rules require that the left hand side corresponds to a non-terminal symbol X that is expanded by the set of terminal and non-terminal symbols $(NT \cup T)^*$ at the right hand side when required. This concept of production rules permits the modelling of embedded symbol sequences by parsing from outside to inside instead of left to right.

$$X \rightarrow \lambda,\ with\ X \in NT,\ \lambda \in (NT \cup T)^* \tag{10.1}$$

A context-free grammar is extended to a PCFG by assigning a probability P to each production rule. This principle is shown in Eq. 10.2.

$$X \rightarrow \lambda\ \ [P] \tag{10.2}$$

Conceptually, this probability is conditional on the selection of the non-terminal symbol X for derivation. The aspect of "contextual freeness" is reflected by the independence of the production rules X_i in a complete PCFG, Eq. 10.3.

$$\forall i : \quad \sum_j P\left(X_i \rightarrow \lambda_j\right) = 1 \tag{10.3}$$

While several problems can be tackled with this approach, we concentrate on the scoring task: we intend to estimate the probability that a symbol sequence was generated by a certain grammar. For this task J. Earley developed a parsing algorithm [5]. This algorithm was extended to probabilistic processing by Stolcke [10].

Further in this section, related works for activity sequence modelling and activity parsing are discussed. Section 10.2 describes our detection approach in the three on-body sensing domains and introduces the activity event parsing method. In Section 10.3 a experimental procedure is sketched to acquire and analyse evaluation data. Section 10.4 reports the achieved performance of the event parsing approach. Finally Section 10.5 summarises the work and Section 10.6 provides an outlook on future research.

10.1.4. Related works

Many attempts have been made to decompose activities into individual events of varying granularity and apply learning machines to identify the events individually. However the combined detection of the activity events sequences is favourable to reason about the superior activities. The methods applied at this level include Hidden Markov Models (HMMs), Bayesian networks, PCFGs and combinations thereof.

For HMMs different solutions have been proposed to model higher-level temporal structures including hierarchical HMMs and layered HMMs. Generally these HMM-based solutions require high training efforts, e.g. the availability of a large training corpus and extensive parameter search in order to tune the large amount of model parameters. Layered HMMs attempt to reduce this complexity by training layers independently [9]. Bayesian networks are by far the most flexible framework for reasoning and have been applied to the recognition of human activities, e.g. [4]. Moreover combinations with other reasoning approaches, such as PCFGs have been attempted, e.g. [7].

Research work relying on PCFGs for activity recognition have been presented in the domain of image recognition mainly, e.g. Moore and Essa [8], Ivanov and Bobick [6] and Yamamoto et al. [11]. Moore and Essa used the Earley-Stolcke parsing algorithm to detect activities in the card game Black Jack. The identification of player strategies were targeted. The authors proposed a complex error handling concept. Ivanov and Bobick demonstrated single recognition results for music conduction and activity surveillance at a

parking lot. A simpler error handling was used in this work. Yamamoto et al. applied PCFGs to the Japanese tea ceremony and tracked the correct activity execution.

10.2. Sensing and detection principle

The sensing domains used for our approach in dietary behaviour monitoring are further detailed in this section. Moreover the event detection method using PCFGs is introduced and basic dietary behaviour models are presented.

10.2.1. On-body sensing domains

To analyse dietary behaviour, we evaluated three sensing domains that are obviously related to dietary intake activities and provide insight into the eating micro-structure. For each sensing domain different sensing modalities are used to detect activity events in continuous data. The following event types are derived:

1. Movement events from inertial sensors, e.g. gestures of the arms during drinking or eating with specific tools.
2. Chew events from an ear-worn microphone sensor.
3. Swallow events from neck muscle contraction (Electromyography electrodes) and a stethoscope microphone integrated into a sensor-collar, worn at the neck.

Fig. 10.1 illustrates the applied sensing modalities and their respective positioning. Inertial sensors have been attached to the lower and upper arms and the back, a microphone is worn at the ear and the sensor-collar at the neck. Data acquisition using this setup is further described in Section 10.3.

Events for each sensing domain are discriminated into different categories: for motion events we distinguished gestures using fork and knife, using a spoon, drinking from a glass or bottle and simple hand-only gestures. A similar approach was taken for the chew and swallow events to discriminate the food texture and food bolus consistency respectively.

The search for individual event types is regarded as independent pattern detection problem. This detection was discussed in previous works [1–3]. For the remainder of this paper a correct detection of these events was assumed in order to analyse the grammar modelling feasibility.

We refer to a sequence of the events containing motion E_M, chew E_C and swallow events E_S as an intake cycle with the number of occurrences N_M, N_C, N_S, Eq. 10.4.

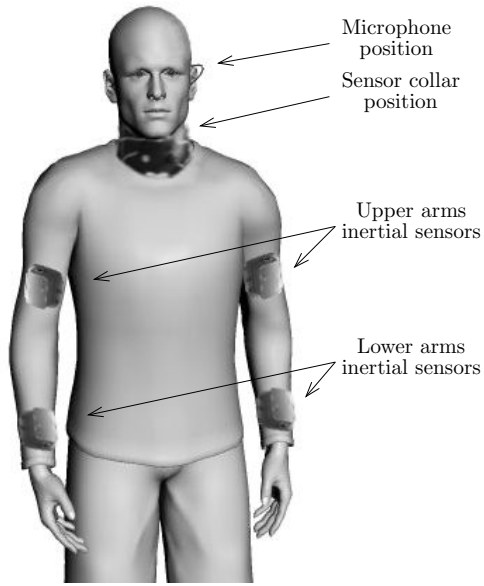


Figure 10.1. Schematic sensor positioning at the body.

$$E_{Cycle} \;=\; (\; E_M^{N_M},\; E_C^{N_C},\; E_S^{N_S}\;) \quad with\; N_M = 1, N_C \geq 1, N_S \geq 1 \tag{10.4}$$

We restricted our intake cycle model to consist of one initial movement event only, $N_M = 1$. This is useful in order to segment individual intake cycles and analyse the natural processing of these single "bites" in isolation. The food type estimation is facilitated by the abstraction, since the food item will not change during a cycle. Certain cycles event types may not be available in all intake cycles, e.g. there are usually no chew events for drinking activities.

10.2.2. Earley-Stolcke parsing algorithm

The aim of our event sequences analysis is to derive an event level segmentation that resembles the intake cycles and classify the food type in parallel. For this goal events are interpreted as terminal symbols of an Earley-Stolcke parser.

The parser processes symbols of the input stream sequentially by applying the defined PCFG. While processing, the parser keeps track of all possible derivations of the symbol sequence. With every new input symbol the number of possible derivations is increased as new alternatives appear or decreased when multiple solutions are resolved. For this purpose the parser keeps a set

of states for each position in the input stream. A state is described by the notation shown in Eq. 10.5.

$$i \ : \ {}_kX \rightarrow \lambda.\mu \ [\alpha, \gamma], \ with \ \ \lambda, \mu \in (NT \cup T)^* \tag{10.5}$$

The index i, $(i \geq 0)$ and the dot "." refers to the current position in the input stream, index k, $(k \leq i)$ indicates the begin of a sub-string given by the non-terminal X. The variables α and γ refer to forward- and inner probability respectively. The forward probability $\alpha_i({}_kX \rightarrow \lambda.\mu)$ is the summarised probability of all paths of length i that end at ${}_kX \rightarrow \lambda.\mu$. The inner probability γ is the summarised probability of all paths of length $i-k$ starting at $k :_k X \rightarrow .\lambda\mu$ and ending at $i :_k X \rightarrow \lambda.\mu$.

The Earley-Stolcke parsing algorithm consists of the states *Prediction*, *Scanning* and *Completion*. A brief summary of the algorithm operation is provided below, a more in-depth elaboration can be found in [10]. For every input symbol the states are processed and the probabilities α, γ are updated. In the prediction step all non-terminals are expanded as long as non-terminal symbols are available. In the scanning step a new input symbol is read and matched to a terminal. When a match was found, the position index i is incremented. All expansions that are not matched in this step are omitted from the current set of states. The completion step is the finalisation of the non-terminal derivation. All fully expanded non-terminals are added to the set of states. Prediction and completion steps can have loops due to cyclic expansions. These are resolved by the parsing concepts *left corner relation* and *unit production relation* [10].

A vital aspect for the PCFG-application in activity parsing is the handling of errors in the symbol sequence. Many related works expect that the input has a low error rate, e.g. Yamamoto et al. [11] and Moore et al. [8]. However the latter work provides a full framework to cope with multiple insertions, substitutions and deletions by hypothetically continuing parsing paths. It can be assumed that the complexity of the parsing algorithm increases significantly due to this complex error handling. Ivanov and Bobick [6] utilised grammar modifications and multivalued input vectors to address insertion and substitution errors. In this paper, we followed the approach of Ivanov and Bobick.

10.2.3. Parsing of dietary activities

Since the relevant activities are very different regarding their activity event structure, e.g. eating and drinking consist of different events, each type of intake cycle was modelled with a dedicated PCFG. For each such PCFG we were interested in solving the scoring problem and determine, how well the current event sequence match the specific grammar.

Fig. 10.2 shows the parsing concept and the parser instantiations used in the evaluation. Events generated from one or multiple sensor pattern detectors are parsed by N parsers, where N equals the number of different PCFGs. Using a dedicated parsing instance for each grammar, provides a scalable solution that

can tolerate multiple differently structured nutrition activities. Eventually all parsing results are combined to a final decision based on the best matching sequence indicated by the parser forward probability.

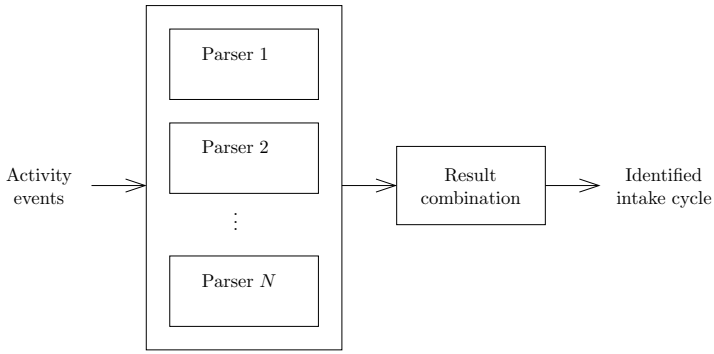


Figure 10.2. Concept of parsing using different PCFGs.

10.2.4. Probabilistic models for eating and drinking

For eating cycles we exploited the freedom of the parsing concept by modelling different foods and food categories by an individual grammar. Eq. 10.6 describes a generic rule for eating based on the intake cycle specification provided above (Eq. 10.4). Every cycle is described by an initiating movement symbol followed by non-terminal chew and swallow symbols[1]. The non-terminal symbols are expanded to a sequence of chew and swallow terminals based on the received chew and swallow events. The model is restricted to swallow terminals from chewed foods only, hence $E_{S,Chewed}$.

$$\begin{aligned} FOOD &\rightarrow E_M\ CHEW\ SWALLOW\ [1.0] \\ CHEW &\rightarrow E_C\ [0.1] \\ &\quad \mid E_C\ CHEW\ [0.9] \\ SWALLOW &\rightarrow E_{S,Chewed}\ [0.5] \\ &\quad \mid E_{S,Chewed}\ SWALLOW\ [0.5] \end{aligned} \tag{10.6}$$

The eating grammar shown above accounts for the number of occurrences of chew and swallow events (N_C, N_S) by the probabilities associated to each production rule. Typical food intake cycles contain multiple chew events, described by a high probability of one chew event followed by further chew events (0.9), while the derivation of a single chew event indicates the end of a chew sequence.

[1]Following the nomenclature in related works, non-terminal symbols are printed in upper case letters.

These probabilities have been chosen manually. Swallow events are modelled in this grammar as finalisation of the intake cycle occurring as one or multiple events.

Contrary to the eating cycle grammar, drinking requires less event types. Here, chewing is not involved in the cycle. Similar to the eating grammar, multiple swallowing events may occur. The movement is restricted to the drink gesture, $E_{M,Drink}$. For drinking a swallow terminal $E_{S,Fluid}$ (fluid bolus item) is required. The grammar is shown in Eq. 10.7.

$$\begin{aligned} DRINK &\rightarrow E_{M,Drink}\ SWALLOW\ [1.0] \\ SWALLOW &\rightarrow E_{S,Fluid}\ [0.5] \\ &\quad \mid E_{S,Fluid}\ SWALLOW\ [0.5] \end{aligned} \tag{10.7}$$

These grammar rules are applied and further discussed in the evaluation described in Section 10.4.

10.3. Evaluation procedure

10.3.1. Evaluation data set

In order to analyse the performance of our parsing approach we recorded a data set of eating and drinking activities using the sensors as described in Section 10.2 above. The sensors were positioned as shown in Fig. 10.1. While the test user was eating different food products an observer annotated the recordings online. In a post-processing step this annotation was manually refined by reviewing the waveforms to obtain sections that reflect the boundaries of the described event types. The annotation information for every event was then used as input for the parsing evaluation.

Tab. 10.1 summarises the recorded foods. In total 3799 events were recorded and annotated from eating and drinking of one test user consuming 11 foods in 162 intake cycles.

10.3.2. Performance analysis

Since there is no automatic algorithm training step involved in the applied parsing approach, we did not partition the data into training and testing set. Instead, we used the entire data set to test the parsing and the grammars.

To analyse performance, we utilised the metrics *Precision* and *Recall*, commonly used for algorithm evaluation in information retrieval applications. These metrics are derived as follows:

$$Recall = \frac{\text{Recognised intake cycles}}{\text{Relevant intake cycles}} \tag{10.8}$$

Table 10.1. Description of the recorded food data set.

Food item	Description
Drink	Drinking from a glass. Drinking does not involve chewing.
Cornbar, Biscuit, Peanuts, Potato chips	Eating the food items using the hand to bring the food to the mouth. The foods are of dry texture during chewing.
Lasagna	Eating lasagna using fork and knife. The cooked food is of soft texture. The swallow bolus is of variable consistency.
Lettuce	Eating using fork and knife. The food is of wet texture. The swallow bolus is of variable consistency.
Bread	Eating bread using the hand to bring the food to the mouth. The food is of soft texture during chewing.
Soup	Eating a soup from a bowl using a spoon. This food item provides no chewing events.
Apple	Eating an apple using the hand to bring the food to the mouth. The food is of wet texture. The swallow bolus is of variable consistency.
Yogurt	Eating from a mug using a spoon. This food item provides no chewing events. The swallow bolus is of variable consistency.

$$Precision = \frac{\text{Recognised intake cycles}}{\text{Retrieved intake cycles}} \tag{10.9}$$

Relevant intake cycles corresponds to the annotated number of actually conducted intake cycle instances. *Retrieved intake cycles* represents the number of cycles returned by the parsing algorithm. Finally, *Recognised intake cycles* refers to the correctly returned number of cycles. Both metrics are defined for the value range $[0, 1]$. A recall value close to one indicates a good sensitivity of a method to return relevant intake cycles, while a precision value close to one indicates that the method does return few insertion errors.

10.4. Results

In the first analysis step we aimed at exploring the sequential properties of the intake cycles and feasibility of the grammar models. For this purpose we applied the simple eating and drinking grammars as defined in Eq. 10.6, 10.7 to the individual foods. For movement and swallow events the abstract event instances were used as described in Tab. 10.1. For chew events we assumed in this step that every food can be modelled by a food-specific symbol. Fig. 10.3 shows the achieved parsing performances using the metrics precision and recall.

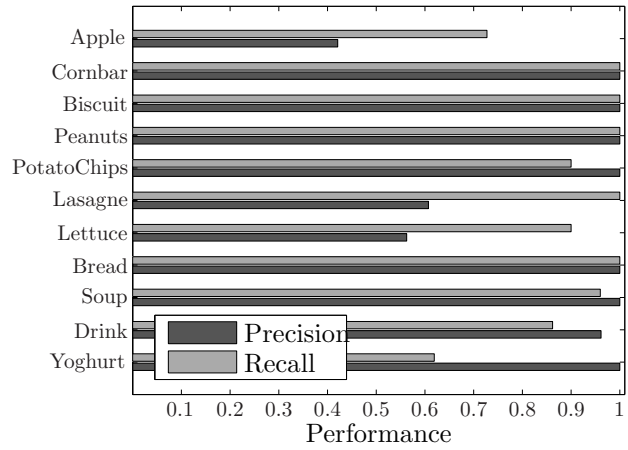


Figure 10.3. Performance chart for the intake cycle detection of the simple food grammars shown in Eq. 10.6, 10.7. For precision and recall, best performance is found towards high values.

These performance values show that the simple model is not a feasible solution for all food types. For several food items many insertion errors were retrieved, indicated by the low precision value at ∼0.6 or below. The used grammar requires a strict sequence of chew and swallow events while many foods contain alternating chew and swallow events, e.g. apple and lasagna. These food items contain more fluid than dry foods, e.g. peanuts, that lead to increased swallow rates. Moreover, the intermediate swallows are an additional food-specific feature that could be explored.

In the following step the food model was refined for non-dry foods to incorporate the typical intermediate swallowing activity. Eq. 10.10 shows the updated grammar.

$$
\begin{aligned}
FOOD &\rightarrow E_M\ MAST^{+}\ [1.0] \\
MAST &\rightarrow CHEW\ SWALLOW\ CHEW\ [0.2] \\
&\quad \mid CHEW\ SWALLOW\ MAST\ [0.8] \\
CHEW &\rightarrow E_C\ [0.1] \\
&\quad \mid E_C\ CHEW\ [0.9] \\
SWALLOW &\rightarrow E_{S,Chewed}\ [0.5] \\
&\quad \mid E_{S,Chewed}\ SWALLOW\ [0.5]
\end{aligned}
\tag{10.10}
$$

$^{+}MAST. =$ Mastication

Using this model, we repeated the analysis of step 1. Fig. 10.4 shows the

parsing performances for this analysis using precision and recall. A clear improvement for food items containing fluid was achieved, e.g. for lasagna the precision increased from ∼0.6 to 1 indicating that no insertion errors were retrieved when parsing the data set with this grammar.

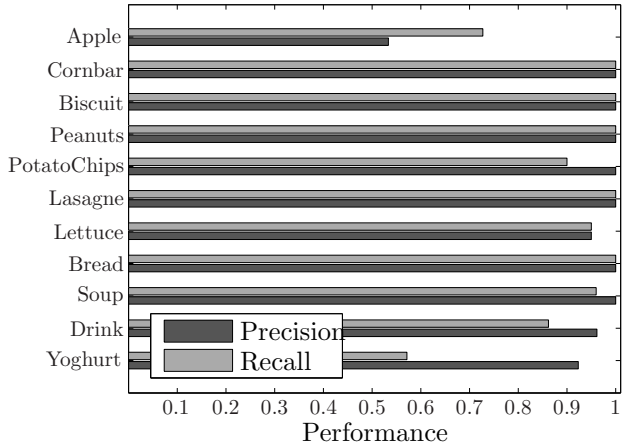


Figure 10.4. Performance chart for the intake cycle detection of the refined non-dry food grammars shown in Eq. 10.10. For precision and recall, best performance is found towards high values.

In a further step we analysed the performances of intake cycles grouped into food categories. We defined the groups based on the similarity of food texture, movement and swallow type. The group "Dry" contained bar, biscuit, peanuts and chips. Yogurt and soup were grouped into "Spoon" since no difference in the activity event sequences was expected for the food items: movement and swallows are similar for both foods and both are not chewed. Fig. 10.5 shows the precision and recall results for the "Dry" and "Spoon" food groups in comparison with the remaining foods.

The very good performance for the group "Dry" indicates that all food items in this group are similar in their event sequence structure. However the new "Spoon" group suffered from high deletion errors, indicated by the low recall. This is mainly due to the weak matching of the grammar on yogurt intake, since yogurt consisted of highly fluctuating number of swallows.

10.5. Conclusion

We presented an approach to detect dietary intake cycles from on-body activity event sequences. The event sequences were modelled using probabilistic grammars. The approach was evaluated with sensor data annotations and the algorithm performance was derived for detecting intake cycles.

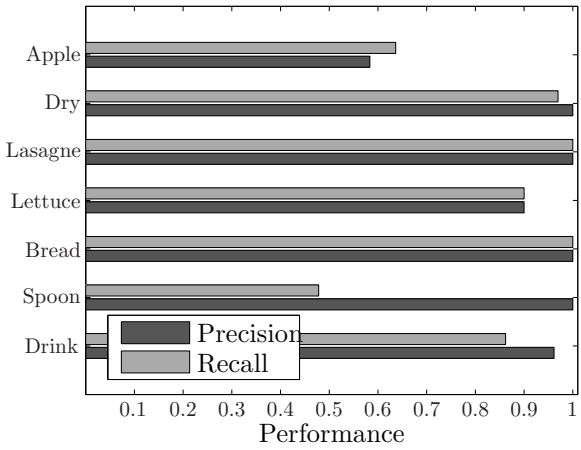


Figure 10.5. Performance chart for detection of intake cycles of "Dry" and "Spoon" groups in comparison with the remaining foods. For precision and recall, best performance is found towards high values.

We analysed different variants of the grammars, starting with simple and strict sequencing rules. The analysis however showed, that these rules were not capable to catch intermediate swallows in certain food cycles. Hence, we adapted the grammars to better accommodate the observed sequences. With the refined rules the detection rates of non-dry foods improved clearly. This analysis addressed the basic intake cycle modelling on individual foods. In order to handle multiple food items a further abstraction from individual foods was needed. For this purpose the food items were grouped by similar texture and intake characteristics. We analysed the feasibility of using one grammar for the detection in each food group.

Overall detection rates of ∼80% were achieved for precision and recall in the food category analysis. This indicates that the intake cycle modelling using probabilistic grammars is a feasible solution. The evaluation was performed with event data acquired from one subject only. However we expect that the approach is scalable to multiple users since no automatic model training was used that would fit the model to the event data. Hence the grammar models applied in this work were rather tuned for food features than for the test user.

10.6. Future work

We plan to further analyse the PCFG approach for detecting dietary intake activities. While the general feasibility of probabilistic grammars was shown in this work, the interconnection with event detection methods will be investi-

gated. Moreover we intend to evaluate the method on further food items and test users.

Bibliography

[1] O. Amft, H. Junker, and G. Tröster. Detection of eating and drinking arm gestures using inertial body-worn sensors. In B. Rhodes and K. Mase, ed., *ISWC 2005: IEEE Proceedings of the Ninth International Symposium on Wearable Computers.*, pp. 160–163. IEEE Press, October 2005. doi:10.1109/iswc.2005.17.

[2] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of chewing sounds for dietary monitoring. In M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, ed., *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing*, vol. 3660 of *Lecture Notes in Computer Science*, pp. 56–72. Springer Berlin, Heidelberg, September 2005. doi:10.1007/11551201_4.

[3] O. Amft and G. Tröster. Methods for detection and classification of normal swallowing from muscle activation and sound. In E. Aarts, R. Kohno, P. Lukowicz, and J. C. Trainini, ed., *PHC 2006: Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–10. ICST, IEEE digital library, November 2006. doi:10.1109/pcthealth.2006.361624.

[4] Y. Du, F. Chen, W. Xu, and Y. Li. Recognizing interaction activities using dynamic bayesian network. In *ICPR 2006: Proceedings of the 18th International Conference on Pattern Recognition*, vol. 1, pp. 618–621, Aug 2006. doi:10.1109/icpr.2006.977.

[5] J. C. Earley. An efficient context-free parsing algorithm. *ACM*, 13(2):94–102, Feb. 1970.

[6] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE T Pattern Anal*, 22(8):852–872, Aug 2000.

[7] K. Kitani, Y. Sato, and A. Sugimoto. Deleted interpolation using a hierarchical bayesian grammar network for recognizing human activity. In *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 239–246, Oct 2005. doi:10.1109/vspets.2005.1570921.

[8] D. Moore and I. Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *AAAI 2002: Proceedings of the American Association for Artificial Intelligence*, pp. 770–776, 2002.

[9] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput Vis Image Und*, 96(2):163–180, Nov. 2004. ISSN 1077-3142. doi:10.1016/j.cviu.2004.02.004. Special issue on event detection in video.

[10] A. Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *CompLingu*, 21(2):165–201, 1995.

[11] M. Yamamoto, H. Mitomi, F. Fujiwara, and T. Sato. Bayesian classification of task-oriented actions based on stochastic context-free grammar. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pp. 317 – 323. IEEE, April 2006. doi:10.1109/fgr.2006.28.

Glossary

Terms and definitions

Activity event	Application-specific non-repetitive pattern in continuous sensor data.
Anorexia nervosa	Eating disorder, obsessive fear of gaining weight.
Binge eating	Eating disorder, periodically uncontrolled food consumption.
Bulimia nervosa	Eating disorder, recurrent binge eating followed by extreme compensatory behaviour.
Calorimetry	Measurement of heat as a result of physical reactions.
Chewing cycle	One period of mandible closing and reopening aiming at food breakdown between teeth. See *Chewing event*.
Chewing event	Used to reference the closing phase of a chewing cycle.
Chewing sequence	All consecutive chewing cycles used to consume one food piece from ingestion into the mouth until final swallow. (Intermediate swallowing may occur between chewing cycles.)
Composite activity	Activity combined from one or more sub-activities in a activity hierarchy.
Crash dieting	Diet, extreme in its deprivations targeting rapid weight loss.
Cricoid cartilage	Cartilage ring around trachea, caudal of the hyoid bone and thyroid cartilage.
Deletion error	False negative; missed activity event of an recognition algorithm.

Digestion	Mechanical and chemical breakdown of food for absorption in the gastro-intestinal tract.
Ear occlusion	In audiology and hearing: Extend of ear canal insulation for air-conducted sound transmission with environment.
Embedding data	Sensor data enclosing relevant temporal data sections (events) in the application's scope. Embedding data does not represent relevant sections itself.
Oesophagus	Organ (muscular tube) transporting a food bolus from the pharynx to the stomach.
Gastro-intestinal tract	System of organs that decompose food and absorb nutrients (digestive tract, gastrointestinal tract, GI tract).
Insertion error	False positive; incorrectly retrieved activity event of a recognition algorithm.
Intake cycle	Composite activity, used to describe food consumption, consisting of intake gesture, chewing, and all swallowing events. The chewing sequence is a subset of the intake cycle.
Intake gesture	Movement of arms and torso intended at food intake.
Malnutrition	Improper or insufficient diet (nutrients) to maintain normal body functions, caused by under- or over-nutrition.
Metabolic rate	Speed of energy conversion in cells.
Metabolism	Collection of chemical reactions to convert energy in living cells.
NULL class	See *Embedding data*.
Nutrients	Food substances required for metabolism, including proteins, fats, carbohydrates, vitamins, dietary minerals, and water.
Occlusion	See *Ear occlusion*.
Orthorexia nervosa	Eating disorder, fixation on food considered healthful.

Pharyngeal movement	Swallowing manoeuvre aiming at food transport form the mouth into the oesophagus without contaminating the airway.
Precision	Performance measure indicating the number of *insertions*, defined as the number of *recognised activity events* divided by the the number of *retrieved activity events*.
Recall	Performance measure indicating the number of *deletions*, defined as the number of *recognised activity events* divided by the the number of *relevant activity events*.
Recognised activity event	Activity event that was retrieved and counted as correct, according to the performance evaluation procedure.
Relevant activity event	True activity event, existing in the dataset and annotated as ground truth.
Retrieved activity event	Activity event that was reported by the recognition procedure. Subsequent performance evaluation accounts retrieved events as correct (see *Recognised activity event*) or error (see *Insertion error*)
Thyroid cartilage	Largest cartilage of the laryngeal skeleton. The laryngeal prominence (Adam's apple) is a formed by the angle of this cartilage.

Abbreviations

ADM	Automatic dietary monitoring
AGREE	Event fusion by agreement in multiple sources
a.u.	arbitrary unit
BMI	Body mass index
C4.5	Decision tree classification algorithm
CEP	Cepstral coefficients
CL	Intake gesture "Cutlery" (using fork and knife)
COMP	Event fusion by comparison of multiple sources
CS	Classification stage
dB	Decibel
DK	Intake gesture "Drink"
ECG	Electrocardiogram
EGG	Electrogastrography
EMG	Electromyography
FFT	Fast fourier transform
FN	False negatives
FP	False positives
FSS	Feature similarity search
FSR	Force sensitive resistors
GI	Gastro-intestinal
HCI	Human computer interaction
HD	Intake gesture "Hand" (using hand only)
HMM	hidden Markov model
Hz	Hertz
IH-EMG	Infra-hyoid electromyography
IMU	Inertial measurement unit
KNN	k-nearest neighbour classifier
LDA	Linear discriminant analysis (also known as Fisher discriminant analysis)
LR	Event fusion by logistic regression
Mic	Microphone
N/A	not applicable
NB	Naïve Bayes classifier
n.d.	not defined
NSGA	Non-dominated sorting genetic algorithm

NT	Non-terminal
PCFG	Probabilistic context-free grammar
PDA	Personal data assistant
PR	Precision-recall
PS	Preselection stage
RFID	Radio frequency identification
RMS	Root mean square
ROC	Receiver Operator Characteristics
SEMG	Surface electromyography
SP	Intake gesture “Spoon”
SM-EMG	Submental electromyography
SWAB	Sliding window and bottom-up
TEF	Thermic effect of food intake
TN	True negatives
TP	True positives
UEM	Universal eating monitor
WHO	World Health Organisation

Acknowledgements

I am sincerely thankful to my advisor Prof. Gerhard Tröster, who accepted my research proposal on diet monitoring and encouraged me in my work. This work would not have been possible without the facilities available at the Wearable Computing Lab. at ETH which he leads.

I am particularly grateful to Stefan Ramseier and Alexander Fach of ABB Switzerland, who encouraged me to follow my interests in research at ETH.

My gratitude goes to Prof. Wolfgang Langhans and Serge Reichlin for inspiring my work and eventually serving as co-examiners of the thesis. I owe special thanks to Prof. Paul Lukowicz, for his steady inspiration and support. I thank Paolo Colombani of ETH Nutrition Biology group and Prof. Nori Geary of ETH Physiology and Behaviour group for discussions and their support of my diet monitoring studies. I am very grateful to all participants for their time spent in various eating studies. Furthermore, I like to thank all the enthusiastic students who contributed with their semester, bachelor, and master theses to the success of this work.

I wish to thank all colleagues at the Wearable Computing Lab., who in some way influenced my research: Holger Harms, Martin Kusserow, Clemens Lombriser, Corinne Mattmann and Thomas Stiefmeier all deserve credit. Moreover, my gratitude goes to several former colleagues, Urs Anliker, Holger Junker, Ivo Locher, Julian Randall, Mathias Stäger, Marc von Waldkirch and Jamie Ward for their support when I started at the lab.

The various projects that I have worked on at ETH would not have been possible without administrative support of Ruth Zähringer and technical help of Urs Egger. I cordially thank them for that.

I am grateful to all members and collaborators of the QBIC project for their effort to make QBIC a success and free my time for this thesis.

I thank David Bannach and Kai Kunze of University Passau for our refreshing collaboration projects that allowed me to test-drive some recognition algorithms that I originally implemented for diet monitoring.

Within the MyHeart project I thank Roland Gasser of FHNW, as well as Harald Reiter, Elke Naujokat, and Robert Pinter of Philips Research, who worked with me on weight coaching concepts. I sincerely hope to advance these concepts in the future.

I am deeply grateful to Corina Schuster for her listening and advice, and for encouraging me to strive for my goals and joy of research.

I acknowledge the financial support of the EU MyHeart project and the Swiss State Secretariat for Education and Research.

Zürich, June 2008 Oliver Amft

Curriculum Vitae

Personal information

Oliver Amft

Born, 16 August 1975, Dresden, Germany
Citizen of Germany

Education

2004–2008	Ph.D. studies (Dr. sc. ETH) in information technology and electrical engineering at ETH Zurich, Switzerland.
1994–1999	M.Sc. studies (Dipl.-Ing.) in information technology and electrical engineering at Chemnitz University of Technology, Germany.
1990–1994	Abitur, Gottfried-Leibniz-Gymnasium, Chemnitz, Germany.

Work experience

2004–present	Research staff member at Wearable Computing Lab, Electronics Laboratory, ETH Zurich, Switzerland.
2004–present	Technology consultant, Utility Communication, ABB Switzerland Ltd., Baden, Switzerland.
2001–2004	R&D Project Manager and Senior Development Engineer, Utility Communication, ABB Switzerland Ltd., Baden, Switzerland.
2000–2001	R&D Development Engineer, ABB Power Automation Ltd., Turgi, Switzerland.
1999	Research internship, Envia AG, Chemnitz, Germany. (6 months)
1999	IT services internship, Siemens Business Services, Chemnitz, Germany. (6 months)
1998–1999	R&D internship, Network IC group, Infineon Technologies, San Jose, USA. (6 months)
1994–1998	Werkstudent (sales, engineering, services), Siemens AG, Chemnitz/Leipzig, Germany.