Probabilistic parsing of dietary activity events

Oliver Amft, Martin Kusserow and Gerhard Tröster
Wearable Computing Lab, ETH Zurich, Switzerland
email: amft@ife.ee.ethz.ch

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.**

Keywords—**Automatic dietary monitoring, nutrition, intake cycle, eating detection, PCFG, Earley-Stolcke parsing.**

I. 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 foodstuff 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.

A. 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.

B. 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.

C. 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. 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.

$$X \to \lambda,\ with\ X \in NT,\ \lambda \in (NT \cup T)^* \qquad (1)$$

A context-free grammar is extended to a PCFG by assigning a probability P to each production rule. This principle is shown in Eq. 2.

$$X \to \lambda \quad [P] \qquad (2)$$

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. 3.

$$\forall i : \quad \sum_j P\left(X_i \to \lambda_j\right) = 1 \qquad (3)$$

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 [4]. This algorithm was extended to probabilistic processing by Stolcke [5].

Further in this section, related works for activity sequence modelling and activity parsing are discussed. Section II describes our detection approach in the three on-body sensing domains and introduces the activity event parsing method. In Section III a experimental procedure is sketched to acquire and analyse evaluation data. Section IV reports the achieved performance of the event parsing approach. Finally Section V summarises the work and Section VI provides an outlook on future research.

D. 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 [6]. Bayesian networks are by far the most flexible framework for reasoning and have been applied to the recognition of human activities, e.g. [7]. Moreover combinations with other reasoning approaches, such as PCFGs have been attempted, e.g. [8].

Research work relying on PCFGs for activity recognition have been presented in the domain of image recognition mainly, e.g. Moore and Essa [9], Ivanov and Bobick [10] 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.

II. 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.

A. 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. 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 III.

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

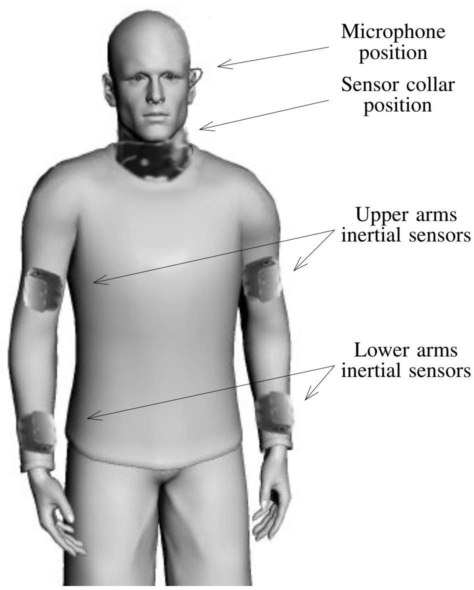

Fig. 1. Schematic sensor positioning at the body.

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. 4.

$$\begin{aligned} E_{Cycle} &= (\ E_M^{N_M}, \ E_C^{N_C}, \ E_S^{N_S} \ ) \\ & \quad with \ N_M = 1, N_C \geq 1, N_S \geq 1 \end{aligned} \tag{4}$$

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.

B. 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. 5.

$$i \ : \ {}_kX \rightarrow \lambda.\mu \ [\alpha, \gamma], \ with \ \ \lambda, \mu \in (NT \cup T)^* \tag{5}$$

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 [5]. 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* [5].

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. [9]. 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 [10] utilised grammar modifications and multivalued input vectors to address insertion and substitution errors. In this paper, we followed the approach of Ivanov and Bobick.

C. 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. 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.

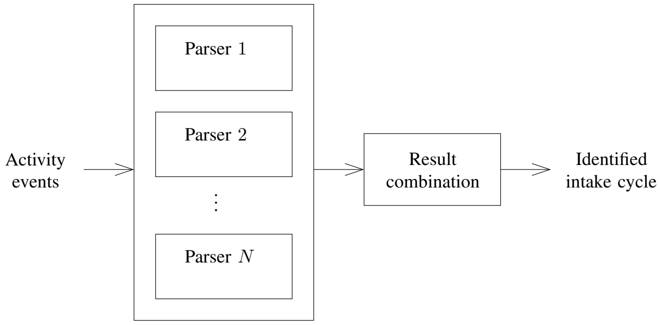

Fig. 2. Concept of parsing using different PCFGs.

D. 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. 6 describes a generic rule for eating based on the intake cycle specification provided above (Eq. 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 $E_{S,Chewed}$.

$$\begin{aligned} FOOD &\rightarrow E_M\ CHEW\ SWALLOW\ [1.0] \\ CHEW &\rightarrow E_C\ [0.1] \\ &\mid E_C\ CHEW\ [0.9] \\ SWALLOW &\rightarrow E_{S,Chewed}\ [0.5] \\ &\mid E_{S,Chewed}\ SWALLOW\ [0.5] \end{aligned} \tag{6}$$

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. 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. 7.

$$\begin{aligned} DRINK &\rightarrow E_{M,Drink}\ SWALLOW\ [1.0] \\ SWALLOW &\rightarrow E_{S,Fluid}\ [0.5] \\ &\mid E_{S,Fluid}\ SWALLOW\ [0.5] \end{aligned} \tag{7}$$

[1]Following the nomenclature in related works, non-terminal symbols are printed in upper case letters.

These grammar rules are applied and further discussed in the evaluation described in Section IV.

III. Evaluation procedure

A. 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 II above. The sensors were positioned as shown in Fig. 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. I summarises the recorded foods. In total 3799 events were recorded and annotated from eating and drinking of one test user consuming 11 foodstuffs in 162 intake cycles.

TABLE I
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.
Lasagne	Eating lasagne 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.
Yoghurt	Eating from a mug using a spoon. This food item provides no chewing events. The swallow bolus is of variable consistency.

B. 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:

$$Recall = \frac{\text{Recognised intake cycles}}{\text{Relevant intake cycles}} \tag{8}$$

$$Precision = \frac{\text{Recognised intake cycles}}{\text{Retrieved intake cycles}} \tag{9}$$

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.

IV. 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. 6, 7 to the individual foods. For movement and swallow events the abstract event instances were used as described in Tab. I. For chew events we assumed in this step that every food can be modelled by a food-specific symbol. Fig. 3 shows the achieved parsing performances using the metrics precision and recall.

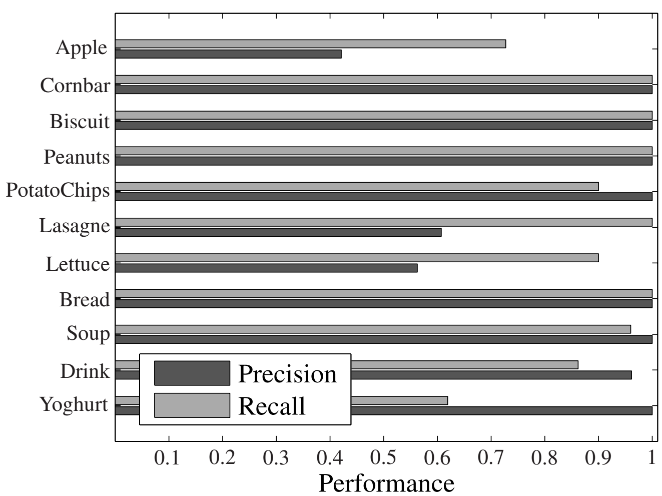

Fig. 3. Performance chart for the intake cycle detection of the simple food grammars shown in Eq. 6, 7. For precision and recall, best performance is found towards high values.

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 lasagne. 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 shows the updated grammar.

$$\begin{aligned}
FOOD \quad & \rightarrow E_M\ MAST^{+}\ [1.0] \\
MAST & \rightarrow CHEW\ SWALLOW\ CHEW [0.2] \\
& \mid CHEW\ SWALLOW\ MAST\ [0.8] \\
CHEW & \rightarrow E_C\ [0.1] \\
& \mid E_C\ CHEW\ [0.9] \\
SWALLOW & \rightarrow E_{S,Chewed}\ [0.5] \\
& \mid E_{S,Chewed}\ SWALLOW\ [0.5]
\end{aligned} \tag{10}$$

$^{+}$*MAST.* = Mastication

Using this model, we repeated the analysis of step 1. Fig. 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 lasagne the precision increased from ∼0.6 to 1 indicating that no insertion errors were retrieved when parsing the data set with this grammar.

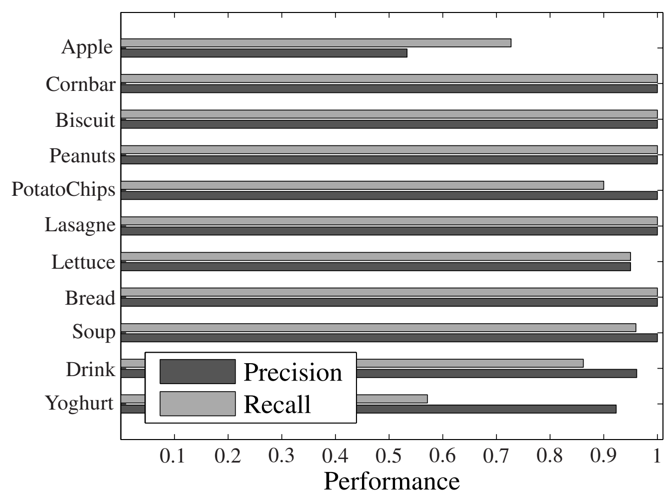

Fig. 4. Performance chart for the intake cycle detection of the refined non-dry food grammars shown in Eq. 10. For precision and recall, best performance is found towards high values.

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. Yoghurt 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. 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

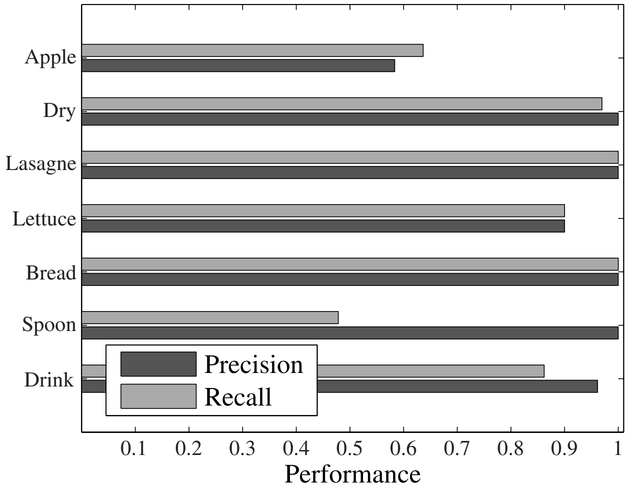

Fig. 5. Performance chart for detection of intake cycles of "Dry" and "Spoon" groups in comparison with the remaining foods. For precision and recall, best performance is found towards high values.

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 yoghurt intake, since yoghurt consisted of highly fluctuating number of swallows.

V. 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.

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 foodstuffs 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 $\sim$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.

VI. 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 investigated. Moreover we intend to evaluate the method on further food items and test users.

Acknowledgement

The authors are grateful to Dr. Yuri Ivanov for providing his Earley-Stolcke parser implementation for reference. The parser was re-implemented for this project. The work was supported by the Swiss State Secretariat for Education and Research (SER).

References

[1] O. Amft, H. Junker, and G. Tröster, "Detection of eating and drinking arm gestures using inertial body-worn sensors," in *ISWC2005: IEEE Proceedings of the Ninth International Symposium on Wearable Computers.*, October 2005, pp. 160–163.

[2] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster, "Analysis of chewing sounds for dietary monitoring," in *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing.*, M. Beigl, S. Intille, and J. Rekimoto, Eds. Springer: Lecture Notes in Computer Science, September 2005, pp. 56–72.

[3] O. Amft and G. Tröster, "Methods for detection and classification of normal swallowing from muscle activation and sound," in *Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare*. ICST, November 2006.

[4] J. C. Earley, "An efficient context-free parsing algorithm," *Communications of the ACM*, vol. 13, no. 2, pp. 94–102, Feb. 1970.

[5] A. Stolcke, "An efficient probabilistic context-free parsing algorithm that computes prefix probabilities," *Computational Linguistics*, vol. 21, no. 2, pp. 165–201, 1995.

[6] N. Oliver, A. Garg, and E. Horvitz, "Layered representations for learning and inferring office activity from multiple sensory channels," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 163–180, Nov. 2004, special issue on event detection in video.

[7] 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, Aug 2006, pp. 618–621.

[8] 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*, Oct 2005, pp. 239–246.

[9] 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*, 2002, pp. 770–776.

[10] Y. Ivanov and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852–872, Aug 2000.

[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*. IEEE, April 2006, pp. 317 – 323.