Sparse Natural Gesture Spotting in Free Living to Monitor Drinking with Wrist-Worn Inertial Sensors

Giovanni Schiboni, Oliver Amft
Chair of Digital Health, FAU Erlangen-Nürnberg, Germany
(giovanni.schiboni, oliver.amft)@fau.de, www.cdh.med.fau.de

ABSTRACT
We present a spotting network composed of Gaussian Mixture Hidden Markov Models (GMM-HMMs) to detect sparse natural gestures in free living. The key technical features of our approach are (1) a method to mine non-gesture patterns that deals with the arbitrary data (Null Class), and (2) an optimisation based on multipopulation genetic programming to approximate spotting network's parameters across target and non-target models. We evaluate our GMM-HMMs spotting network in a novel free living dataset, including totally 35 days of annotated inertial sensor's recordings from seven participants. Drinking was chosen as target gesture. Our method reached an average F1-score of over 74% and clearly outperformed an HMM-based threshold model approach. The results suggest that our spotting network approach is viable for sparse natural pattern spotting.

ACM Classification Keywords
L.4.1.1 Human-centered computing: Mobile computing; F.6.4 Theory of computation: Genetic programming; F.5.3.4.1 Theory of computation: Random walks and Markov chains

Author Keywords
Wearable Devices; Automatic Dietary Monitoring.

INTRODUCTION
With gesture pattern spotting we describe the search and identification of a particular known signal shape in a continuous stream of arm-related motion sensor data. Spotting specific pattern events in arm motion can reflect manipulative gestures, relating to healthy behaviour. For example, drinking gestures could relate to the actual hydration level of an individual, as maintaining an adequate fluid intake is an important goal in dietary management [12]. Research has shown that naturally entrained manipulative gestures, such as lifting a container to take a sip, are harder to spot from sensor data than gestures designed to control devices, or to interact with devices, due to several challenges [7]. First, patterns of natural manipulative gestures are often more varying than the patterns of control or interaction gestures. The latter may be repeated and practiced until the intended effect is achieved, whereas the natural gestures are entrained from childhood, and vary by environment and specific object interactions. For example, drinking can be performed with various containers, which influence the gesture performance. Earlier work has even shown that containers could be distinguished from gesture patterns [2]. Second, the temporal distribution of natural gestures is often sparse, i.e., the relevant gestures are embedded in arbitrary data, the Null Class, with their summed event durations amounting to less than 1/100 of the total data. As a consequence, insertion errors (false positives) are very likely to occur. Third, the gestures are varying in duration, which raises the intra-class signal pattern variance. Previous investigations addressed natural gesture spotting in constrained settings, e.g., simulating free living conditions, or limiting the arbitrary data. Please see related work for details. Spotting sparse natural gestures as drinking, thus, remains an open challenge.
This work presents an online spotting approach for sparse gestures, designed for free living deployment. In particular, we investigate a spotting framework based on Gaussian Mixture Hidden Markov Models (GMM-HMMs) that describes gesture and non-gesture patterns in a multimodel network. We introduce a method to mine non-gesture models and a novel model optimisation method based on multipopulation genetic programming. The optimisation is key to tune sensitivity of GMM-HMMs. While we demonstrate the gesture spotting method in a free living validation for drink gestures, we believe that the method can be applied to other sparse natural gestures. In particular, the following contributions are made:

1. The GMM-HMM network architecture, the non-gesture mining approach and network optimisation method are presented. The mining intends to synthesise meaningful non-gesture models that help the network to reject the Null Class. In turn, the optimisation intends to derive model priors for each individual model in the multimodel network.

2. Spotting natural drink gestures, under realistic ratios of the relevant gesture vs. Null Class, was analysed from a daily life recording dataset. Seven participants wore inertial motion sensor units (IMUs), for about five days, from wakeup to bed time. We evaluate the spotting in personalised models.

3. We compare our spotting approach to the HMM threshold model to demonstrate the recognition challenge and the ad-

ISWC '18, October 8–12, 2018, Singapore, Singapore
© 2018 ACM. ISBN 978-1-4503-5967-2/18/10…$15.00
DOI: https://doi.org/10.1145/3267242.3267253

vantage of the GMM-HMMs network for sparsely occurring natural gestures.

RELATED WORK

Gestures spotting for dietary activity recognition has been studied in constrained, i.e., laboratory, or semi-constrained, e.g., limited time, or limited activity, environment in the last decade, e.g., see [7, 2, 13]. Although the aforementioned works presented promising results, the analysis of natural intake gestures is still an open problem. High pattern variability and varying environmental conditions could hardly be reproduced in a laboratory setting. As a consequence, directly transferring methods and models into free living results in a drastic performance drop.

Finding solutions for Null Class handling is among the major challenges in many fields where pattern spotting is considered. The Null Class problem is directly related to the sparsity of relevant gestures and the pattern similarity in the arbitrary data. In other words, the more arbitrary data, the lower will be the spotting performance results of the algorithm. Clearly, specific behaviour patterns in daily life are among the most challenging spotting tasks.

The use of a garbage model is widely applied in speech recognition applications, being used to model acoustic non-keyword patterns, e.g., see [14]. Other context recognition fields deploy garbage models too. Mannil et al. [9] described a framework to classify specific actions, and transitions among them, and to reject unwanted patterns. An early rejection of unwanted patterns was also implemented. Peng and Qian [10] presented an online full body gesture spotting from visual hull data. View-invariant features using multilinear analysis were extracted and used as input of a HMM-based spotting network. Moreover, they presented a method to mine specific non-gesture models from the training data, used as garbage model. Lee and Kim [8] were the first to introduce the concept of a HMM threshold model used to recognise interaction gestures in video data. A network was proposed comprising one or more HMM chains to model target classes, and an ergodic model to reject non-gesture patterns. The ergodic model was composed by the states of the target models, however their transition probabilities were tuned. In this work, we propose a multimodel approach to deal with the variety of Null Class patterns in a spotting network.

A common characteristic of the aforementioned frameworks is that their recognition performances are highly affected by the sensitivity of the network's components. Often, the model priors, or entry probabilities, are key to success or failure of the HMM network. Earlier works tackled the problem by using exhaustive search [8], through ad hoc methods [3], or analytically, via binary-mixed integer programming [6]. In this work, the model's entry probabilities are optimised with a multipopulation genetic algorithm.

METHODS

In order to spot gestures from the continuous stream of sensor data, i.e., accelerometer, gyroscope and magnetometer, we employed a HMM network. The HMM network has a two-fold purpose, being able to correctly detect drink gestures, represented as the target gesture model (TM), but also to reject other gestures and the Null Class, under the chosen spotting task, using non-target gesture models (NTMs).

GMM-HMM Training Procedure

In our framework, we employed GMM-HMMs due to their sequential modeling and inference capabilities with complex multivariate time-variant patterns. Our GMM-HMM is represented by the parameter set:

$$\lambda = \{A, B, \pi\}, \tag{1}$$

where A is the state transition probability matrix with size $S \times S$, B are the emission probabilities, π are the $S \times 1$ initial state probabilities, with number of states S. Each state is modeled as a Gaussian mixture. Our observations are continuous, multivariate random variables modelled as:

$$b_s(o) = \sum_{m=1}^{M} c_{\mathrm{s,m}} \mathcal{N}(o|\mu_{\mathrm{s,m}}, \Sigma_{\mathrm{s,m}}), \quad s = 1, \ldots, S, \tag{2}$$

where $\mu_{\mathrm{s,m}}$ and $\Sigma_{\mathrm{s,m}}$ are the mean and the covariance of the m^{th} mixture component from the s^{th} state, with $c_{\mathrm{s,m}}$ as the weight of the mixture, and M number of mixtures. The total number of emission parameters can be calculated as $S \times M \times size(\mu_{\mathrm{s,m}}, \Sigma_{\mathrm{s,m}})$. To reduce the overfitting risk, GMM covariance was constrained to diagonal matrices. It is well known that the Expectation-Maximization, used by the Baum-Welch algorithm, can only find local likelihood maxima that mainly depend on the arbitrary initialisations. Thus, successful training of the GMM-HMM parameters using Baum-Welch re-estimation depends on the initial model's probability assignment strategy.

Our GMM-HMM training involved the following steps.

Instance Resizing. Each training instance was resized to the same arbitrary length ψ. If the original length of the instance was smaller than ψ, we applied linear interpolation, otherwise downsampling. We chose ψ to be equal to the average length of all training instances. The resized instances were used only for clustering. In subsequent HMM training and analysis stages, the original, non-resized data were used. The procedure removed temporal variance of the data patterns, in order to reduce the in-cluster variance.

Instance Segmentation and Grouping. Each training instance was segmented in S equal length segments, as the number of states of the GMM-HMM. Subsequently, temporally correspondent segments were grouped together.

K-Means Clustering. A K-means clustering per segment group was performed.

Gaussian Mixture Distributions Initialisation. Each GMM mean parameter $\mu_{\mathrm{s,m}}$ was initialised with a cluster's center value, and the covariance matrix $\Sigma_{\mathrm{s,m}}$ was initialised with an identity matrix.

Baum-Welch Training. GMM-HMMs were trained as a *left-right* model using the original, non-resized instances. All models were defined as *absorbing* chains, i.e., the terminal state's self transition was set to one.

Non-Gestures Modelling

We designed a procedure to mine NTM data instances from the training dataset. Apart from the mined NTMs, we also

Number of TMs ($\|\Gamma^+\|$)	1	Number of model states (S)	7
Number of NTMs ($\|\Gamma^-\|$)	43	Number of Gaussian mixtures (G)	4
Number of gesture-based NTMs	3	Number of features (F)	9
Number of non-gesture NTMs	40	Resize length for clustering	500
Cross-validation folds (CV)	5	Number of selected NTMs for training (Q)	200
Number of target instances per part. (P)	see Tab.3		

Table 1: Hyperparameters used for the GMM-HMM network.

incorporated selected counterexample gesture-based NTMs, as described further below. Our extraction and training procedure involved the following steps.

Segmentation. Segments of the continuous stream of data were collected using an energy-based segmentation whilst leaving out gesture instances. The motion energy was calculated from the angular rate ω, output of the gyroscope sensor as:

$$\theta_t = |\omega_{x,t}| + |\omega_{y,t}| + |\omega_{z,t}|, \tag{3}$$

where θ is the energy of the rotational motion at time t.
A running average, with a window size of $0.5s$, was applied to θ, and, subsequently, a valley detection was performed with a minimum valley distance of 2 s. The identified points were used as starting and ending points of NTM instances.

Instance Resizing. The same procedure described in the training procedure to resize instances, see Sec. 3.1, was applied here. The resized instances were used only for clustering. In subsequent HMM training and analysis stages, the original, non-resized data were used.

Feature Extraction. Each training instance was segmented in a number of equal length segments. Mean, standard deviation, variance, and skewness, were calculated for all instance's segments. A feature vector per instance was collected.

Principal Component Analysis. We applied *principal component analysis* to reduce the instance feature space, ensuring that the subsequent K-means could find meaningful clusters.

K-Means Clustering. A K-means clustering was performed with the aim to derive K NTMs.

Training Samples Selection. Using Euclidean distance, we selected the closest instances to the cluster centers, grouping a number of NTM training samples for each class. The number of selected instance is indicated in Tab. 1.

GMM-HMM Training Procedure. Each class was processed via the training procedure explained in Sec. 3.1.

Network Construction and Optimisation

A major challenge for building an effective spotting network is to define the models' sensitivity encoded by the entry probabilities. Tuning of the entry probabilities affects the posterior probabilities of the models. In this section, we detail the entry probabilities' relevance and how the values were optimised.
Our network is composed of parallel TMs and NTMs, connected by a non-emitting starting and ending states, *meta-start* and *meta-end* respectively, see Fig. 1. We denote the gesture TM as γ^+, the set of TM's parameters as λ_{γ^+}, and the set of TMs as Γ^+, in our application $|\Gamma^+| = 1$. We also denote a gesture NTM as γ^-, the set of NTM's parameters as λ_{γ^-}, and the set of gesture NTMs as Γ^-, in our application $|\Gamma^-| = 43$. The set Γ^- comprises gesture-based models, as different kind of food intake that we conveniently extracted from the dataset,

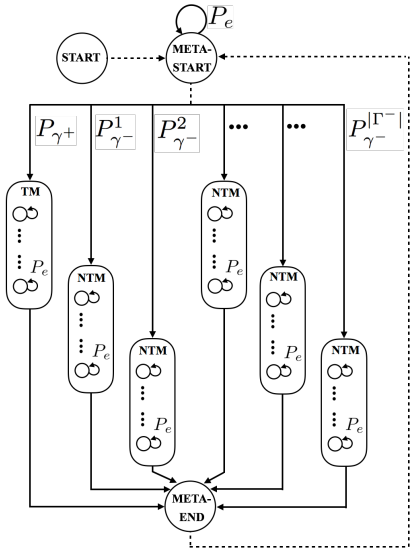

Figure 1: GMM-HMM network with a non-gesture multi-model mining for on line spotting and rejection of temporal patterns. Our multimodel approach mines $|\Gamma^-|$ non-gesture models from the Null Class data and aligns them using a genetic algorithm optimisation.

and non-gesture patterns mined by following the procedure described in Sec. 3.2. The network is optimised by finding the entry probabilities $P_{\gamma^+}\ \forall\gamma^+ \in \Gamma^+$, and $P_{\gamma^-}\forall\gamma^- \in \Gamma^-$ such that, for a given target observation sequence O_T:

$$P_{\gamma^+}P(O_T|\lambda_{\gamma^+}) > P_{\gamma^-}P(O_T|\lambda_{\gamma^-})\ \forall\gamma^- \in \Gamma^-. \tag{4}$$

Also, for a given non-target observation sequence O_{NT}:

$$\exists\ \gamma^* \in \Gamma^-\ |\ P_{\gamma^+}P(O_{NT}|\lambda_{\gamma^+}) < P_{\gamma^*}P(O_{NT}|\lambda_{\gamma^*}), \tag{5}$$

with the constraint:

$$P_e + \sum_{\gamma^+\in\Gamma^+} P_{\gamma^+} + \sum_{\gamma^-\in\Gamma^-} P_{\gamma^-} = 1, \tag{6}$$

with P_e being the self transition probability of the models' terminal states, see Fig. 1. Since our GMM-HMM models resembled an absorbing Markov chain, their final state's self transition was $P_e = 1$. Once the models were incorporated into the spotting network, the self transition was changed to $P_e \approx 1$ in order to permit transitions towards the meta-start state. The constraint in Eq. (6) guarantees consistency of the transition matrix to model transitions to and from the meta-start state.
The condition in Eq. (4) guarantees that the TM has a higher likelihood than all NTMs, given the target observation sequence O_T. The condition in Eq. (5) guarantees that a non-target model is associated with a non-target observation sequence O_{NT}. The probabilities in Eq. (4, 5) are computed from annotated isolated instances of the training set, i.e., $P(\bullet|\lambda_{\gamma^+})$, and from the NTMs mined with segmentation procedure, i.e., $P(\bullet|\lambda_{\gamma^-})$. Each sequence O_T, and O_{NT} represents a constraint on Eq. 4 and Eq. 5, respectively. Determining the values $P_{\gamma^+}\ \forall\gamma^+$ and $P_{\gamma^-}\ \forall\gamma^-$, which satisfy as many of the constraints

as possible, corresponds to correctly classify relevant and irrelevant observation patterns.
The problem to find the entry probabilities $P_{\gamma^+}\ \forall\gamma^+$ and $P_{\gamma^-}\ \forall\gamma^-$ is NP-hard and cannot be handled by brute force search. Thus, a distributed multipopulation genetic programming technique [4] was used.
We developed a genetic algorithm that produced through evolutionary generations a set of candidate solutions. A solution was interpreted as a vector of entry probability values. As *fitness function*, we employed the weighted accuracy of the isolated classification task between TMs and NTMs. Each candidate solution was identified as an *individual* being part of a *population*. The populations were distributed on separate islands and some individuals were allowed to migrate between islands with a certain frequency. The *island model* was chosen to explore as many search areas as possible and maintaining diversity among populations through a population migration.

Our multipopulation genetic programming procedure in-

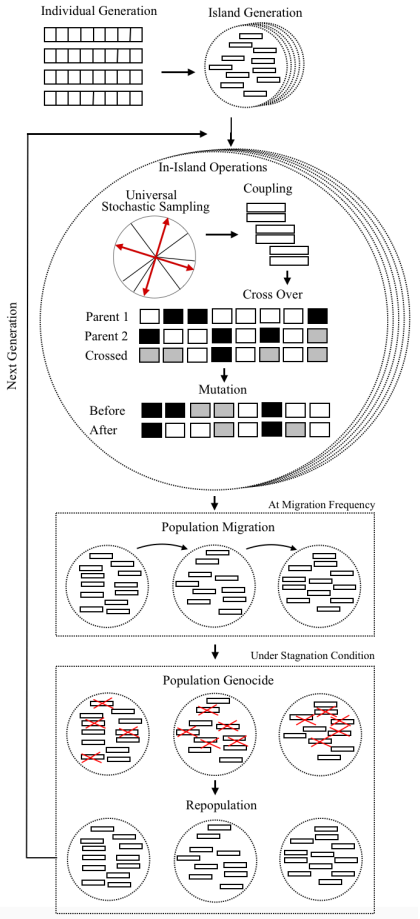

Figure 2: Multipopulation genetic programming procedure. In-island individuals evolved through generations via a stochastic coupling procedure. Cross-over and mutation operations were applied. Population migrations were allowed at a certain frequency. Under stagnation condition, a fraction of population was eliminated and substituted with new individuals.

volved the following steps, see Fig. 2.

Individual Generation. Each individual was encoded as a set of real numbers representing the entire set of entry probabilities:

$$\bigcup_{\gamma\in\Gamma^+\cup\Gamma^-} P_\gamma. \tag{7}$$

The individual generation was constrained by (6).
Island Generation. All individuals were grouped in populations and separated in different islands. Our island model was homogeneous, since the parameters and the genetic material were identical in all the subpopulations.
In each island, the following operations were performed:
Universal Stochastic Sampling. The aim of the operation was to choose parents to breed in order to create new genetic material for the next generations. This technique for choosing potential useful solution was used because it exhibits no bias and minimal spread. A single random value was used to sample all the solutions, by sampling them at evenly spaced intervals, that were proportional to the individuals' fitnesses.
Coupling. Selected individuals were randomly coupled among themselves. We forced each individual to be able to couple with just one partner per generation.
Cross Over. Since our individuals were decoded as real numbers, the cross over operation was interpreted as the mean between correspondent elements of the individual. The cross over operation's execution on each element of the individual depended on a tunable cross over rate parameter.
Mutation. The mutation operation was interpreted as substituting an element of the individual with a new random element. The mutation operation's execution on each element of the individual depended on a tunable mutation rate parameter.
The following sporadic operations were also applied to all islands:
Individual Substitution. Along the generations, some worst individuals were substituted by new random individuals.
Migrating Individuals. A fraction of each population at a certain migration frequency was set to move between islands. The communication topology among islands was defined as random. Individuals in the departure and destination islands were ranked in terms of fitness value. A number of best individuals from the departure island were selected to migrate, after being mutated. A number of worst individuals from the destination island were selected to be substituted.
Population Genocide and Repopulation. Along the evolutionary process, the average fitness value among islands may reach local maxima points. If no improvement in the average fitness was recorded for a certain number of generations, i.e., stagnation threshold, a stagnation condition was declared. As a consequence, a random elimination, i.e., population genocide, of a number, i.e., genocide ratio, of the worst individuals, was performed, followed by a repopulation via random generation of new individuals.

On Line Decoding

The Viterbi algorithm was used to estimate the joint probability of an observation sequence along the path of state transitions in the network. The optimality criterion was used to maximise probability of a state sequence. At each time instant t an observation sample o_t, i.e., a multivariate motion feature, was processed.

Population size (PS)	1000	Migration frequency	$\frac{1}{10 \text{ generations}}$
Mutation rate (MR)	0.6	Migrating population size	200
Cross over rate (CR)	0.6	Communication topology	Random
Number of Generation (NG)	100	Stagnation threshold	20 generations
Islands' number	40	Genocide ratio	0.5

Table 2: Hyperparameters of the multipopulation genetic programming optimisation.

Once defined $\Gamma = \Gamma^+ \cup \Gamma^-$, we call S_s^γ the s^{th} state belonging to model γ, with $\gamma \in \Gamma$, $\mathscr{S}$ the set of all the states of the network, σ_t the hidden state at time t, and $\mathbf{O}_t$ as the observation sequence up to time t.
The joint probability of $\mathbf{O}_t$ and the optimal state path up to $\sigma_t = S_s$ is:

$$\delta_t(S_s) = \max_{\sigma_1,\ldots,\sigma_{t-1}} p(\mathbf{O}_t, \sigma_1, \ldots, \sigma_{t-1}, \sigma_t = S_s|\lambda). \quad (8)$$

At each time instant t, $\delta_t(S_s)$ was processed based on the value $\delta_{t-1}(S_s)$, $\forall S \in \mathscr{S}$.
The goal was to detect the start and endpoints of target gestures embedded in the stream of data. By monitoring the temporal pattern of the individual models' likelihood, we were able to detect when a gesture was executed. At time instant t, if the end probability of the model γ^+ was the largest among all the models, the observation o_t was considered to be an endpoint candidate of a gesture g. Analytically, the endpoint implies the following two conditions:

$$g = \arg\max_{\gamma \in \Gamma} \delta_t(S_e^\gamma) \text{ and } \gamma \in \Gamma^+. \quad (9)$$

When a set of consecutive endpoints were detected, the starting time was found by backtracking along the most probable paths and choosing the one with the highest likelihood.

Feature Selection

The IMU provided nine dimensional continuous sensor data, i.e., three dimensional accelerometer, gyroscope, and magnetometer signal. We applied findings from Günter and Bunke [5] in order to implement a forward selection-like algorithm to select the best continuous feature dimensions. Only gesture models were used in the procedure for feature selection, i.e., drink and eating gestures (fork, spoon and hand). A classification scheme was adopted by using weighted accuracy as performance metrics. As result, we observed that best performance were obtained when all features were employed in the HMM modelling.

Performance Comparison

Performance comparisons were made with the HMM-based threshold model [8], usually used in gesture recognition for human-machine interaction. The TM and gesture-based NTMs were kept the same for the threshold model spotter. The threshold model was composed of TM and gesture-based NTM states. The threshold model was ergodic, and emission probabilities and state self transitions were left unchanged. The entry probabilities of the model were found through a brute-force search.

Daily Life Activity Data Collection

Seven healthy volunteers (2 females, 5 males) aged between 20 and 40 years with a normal BMI were recruited among the universities' students and lecturers, not including the authors. After orally explaining the study procedures and the objective, and confirming that participants understood the description, they could provide consent in written form to participate in the recordings. Participants were informed that they could exit from the study at any time, without disadvantage or providing reasons, in which case their data would be deleted.
Participants wore a Shimmer IMU attached with a stretchable band at the wrist of their dominant arm during regular daily routine. Participants received the sensors on the day before the recording and were asked to attach the sensor in the morning after waking up. The sensor was worn throughout the day, until bedtime and only detached during activities that could immerse the sensor in water (e.g. swimming) or that could break the sensor (e.g. weight lifting). Each participant wore the sensor for approximately five consecutive days. Participants were asked to follow their usual daily activities and to go to the lab during the afternoon of each recording day to exchange the sensor with a recharged one and receive another sensor for the next day. Sensors were configured to store the measurements in the internal flash memory. When sensors arrived at the lab, they were immediately read out.

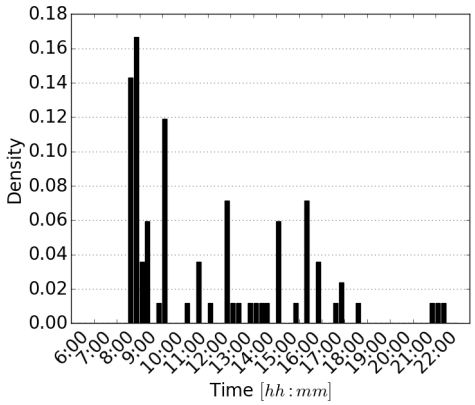

Figure 3: Example distribution of drink gestures over five recording days of one participant P1 based on self annotations and post-recording label revision .

To interpret the sensor data, the following procedure was used. Participants where asked to fill in a paper diary time sheet that allowed them to mark activities (drinking, eating types: fork, spoon, hand, and daily routines) with lines. The time sheet had a two-minute resolution. In a post-processing step, the participants' annotations were reviewed together with the sensor data. Based on the sensor data patterns (primarily acceleration and gyroscope waveforms), labels marking the start and end of each drinking and eating gesture were created by a study manager, who was experienced to interpret the signal patterns. All data and labels were reviewed and refined by a second study manager, who was similarly trained to interpret the signal patterns. We estimated the average label start/end resolution to be ± 0.2 s.
While the participants annotated drinking occasions, labels

	P1	P2	P3	P4	P5	P6	P7
Number of instances	84	154	139	164	279	129	135
Data Size [h]	57.1	56.8	48.0	59.3	39.8	67.4	36.9
Gesture Data Size [h]	0.12	0.19	0.14	0.23	0.27	0.13	0.16
Event Rate [1/h]	1.4	2.7	2.9	2.7	7.0	1.9	3.6

Table 3: Number of drink gestures instances, duration values, and event rates over five recording days. Each column refers to one participant $[P_1, \ldots, P_6]$.

GMM-HMM	
Model Training	$\mathcal{O}(S^2 \cdot T)$
Prediction per sample	$\mathcal{O}((B+1)S)$
Observation Likelihood (B)	$\mathcal{O}(G \cdot F)$
Optimisation	
Genetic Algorithm	$\mathcal{O}$(PS * NG $\mathcal{O}$(Fitness) * (CR * $\mathcal{O}$(Crossover) + MR* $\mathcal{O}$(Mutation)))
Fitness	$\mathcal{O}$((Q * (CV-1))+P)
Crossover/Mutation	$\mathcal{O}(\lvert\Gamma^+\rvert + \lvert\Gamma^-\rvert)$

Table 4: Time complexity of the main components of our framework. Symbols refer to Tab. 1 and Tab. 2.

were created for each gesture instance for the spotting analysis. Since the study managers searched the entire data recordings, labels were defined within and besides the participant annotations. The annotations nevertheless provided important structural markers to understand the daily routine and interpret the sensor data, e.g., riding a bicycle or running. The annotation procedure was kept minimally invasive for the participants to retain their natural behaviour. As a consequence, gesture labels could have been missed, which would show in the spotting performance as insertions and reduced precision. The average monitoring duration per participant was 49.2 hours, with an average 0.53 hours spent in food intake-related gestures. See Tab. 3 for participant-related recording durations. In Fig. 3, the temporal density distribution of drink gestures is depicted for a typical participant.

Spotting Evaluation

Personalised GMM-HMM models were created, repeating the training procedure for each participant. A 5-fold cross-validation was applied to evaluate spotting performance, where each fold corresponded approximately to one day. An adjustment of the bounds of each fold was made to balance the target gesture instance count across folds. In Tab. 1, empirically found hyperparameters are listed. Drink gestures were used for the TM. Besides the mined NTMs, the labeled food intake gestures (fork, spoon, hand) were used to derive three gesture-based NTMs from the training dataset, as described above. We chose these gestures as additional NTMs, to specifically avoid false positives due to the similarity of drink and eating gestures.

Retrieval Perfomance

Precision, recall and F1 score were chosen to analyse the retrieval performance. A measure of retrieval generalisation is the event rate, which has been introduced to assess sparse instance scenarios [1]. The event rate measures the sparsity of the relevant instances in the dataset. We ran the recognition's evaluation four times per participant per fold. Each time a larger portion of data was processed. Initially, at an event rate of 303.3/h, only data related to eating episodes were considered, see Tab.6. In subsequent iterations, an increasing share of the Null Class was included in the spotting task, resulting in reducing event rates down to 3.2/h.

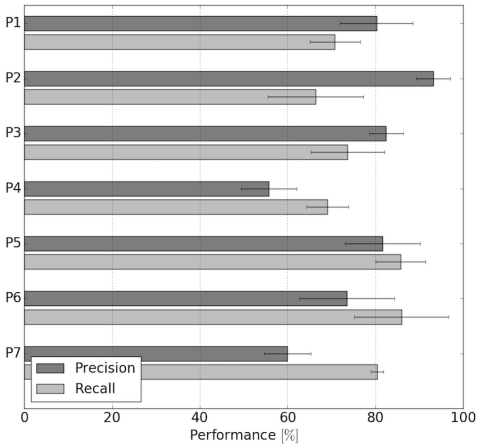

Figure 4: Average gesture spotting retrieval performance per participant $[P_1, \ldots, P_7]$ with 5 fold cross-validation.

RESULTS

In Fig. 4, retrieval performance of all participants is reported. Average precision was 75.2%, with a variation by participant ranging from 55.7% for P4 to 93.2% for P2. Average recall was 76.1% ranging from 66.4% for P2 to 85.9% for P6. In Fig. 6, the comparison of our method and the HMM threshold model is shown. The threshold model's average precision was 28%, with a minimum of 11%, for P4, and a maximum of 34.8%, for P2. Average precision difference between our method and threshold model was of 47.2%. The threshold model's average recall was 49.1%, with a minimum of 15.5%, for P4, and a maximum of 66.9%, for P2. Average recall difference between our method and threshold model was of 26.9%. Tab. 5 finally summarises the spotting results for the two methods. The average number of insertions per partic-

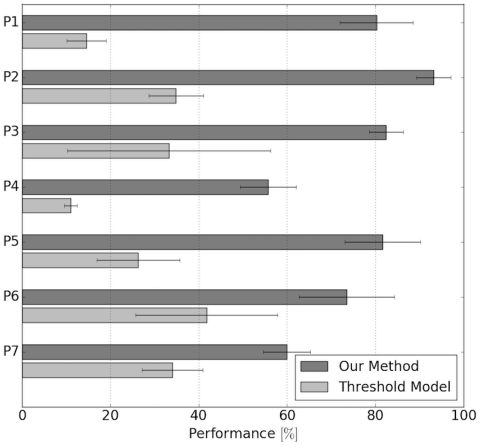

Figure 5: Comparison of average gesture spotting precision between our method and the threshold model approach per participant $[P_1, \ldots, P_7]$.

	F1-Score	
	Mean Value [%]	Standard Deviation [%]
Threshold Model	32.0	8.7
Our Method	74.4	4.5

Table 5: Overall retrieval performance of the two methods.

ipant was 42 and the average number of deletions was 35. Considering the multiday dataset per participant, the average performance degradation due to event rate variation was 7.8% with a standard deviation of 4.9%, see Tab.6.

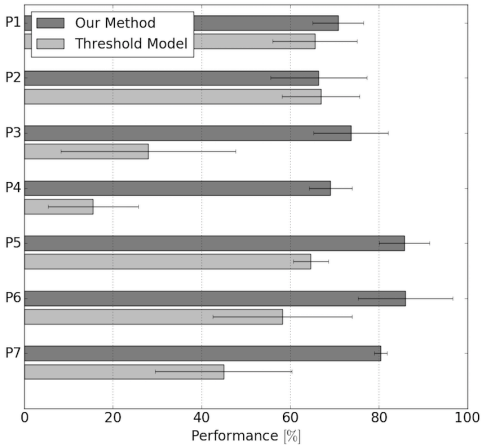

Figure 6: Comparison of average gesture spotting recall between our method and the threshold model approach per participant $[P_1, \ldots, P_7]$.

DISCUSSION

The results showed that a personalised spotting and rejection of non-relevant gestures is feasible in the free living setting. Two key challenges had to be addressed. (1) The variability of the natural target gesture patterns were captured by the TM in our GMM-HMM network. (2) The substantial Null Class due to the multiday dataset needed to be handled in a set of NTMs to reject observations that did not resemble a target gesture pattern. Our NTM mining found patterns suitable to effectively reject Null Class data. We consider that gesture-based NTMs (fork, spoon, hand gestures), which were particularly similar to the TM, helped us to incorporate domain knowledge. When we view the spotting task as an information retrieval problem, the gesture-based NTMs correspond to search with exclusion patterns, i.e. search for the TM, but not these gesture-based NTMs.

The multipopulation genetic HMM parameter optimisation has been essential to balance insertions and deletions of the spotting network. The retrieval performances reached an adequate level for a practical implementation. Participants P4 and P7 performed worse than the others. P4 and P7 showed the largest gap in performance when varying the dataset size. Since all TM gesture instances appeared in all dataset sizes

		F1-Score [%]						
Data Size [h]	Ev. Rate [1/h]	P1	P2	P3	P4	P5	P6	P7
0.52	303.3	79.1	79.7	81.6	78.0	89.5	88.8	81.6
26.0	6.4	76.8	78.9	79.3	71.8	86.5	84.3	74.9
39.1	4.2	75.9	78.2	78.2	66.9	85.1	81.6	71.3
52.1	3.2	75.1	77.7	77.5	61.7	83.7	79.2	68.5

Table 6: Retrieval performance for varying dataset sizes and average event rates, illustrating the method generalisation. See main text for details.

of Tab. 6, the similar performance across all participants observed for the 0.52 h dataset size indicates that the TM could be correctly retrieved. The F1 drop of more than 13% for P4 and P7 when increasing the dataset size may be due to target gesture instances missing labels or insufficient Null Class rejection of the model.

When comparing our method's results with a threshold model, the latter showed to be inadequate for spotting sparse natural gestures from continuous features. Further investigations are required to clarify all reasons for the performance gap between the methods. Nevertheless, some interpretation can be done already.

The threshold model's principle is based on the fact that a gesture is modelled as sequential progression of segmental patterns, i.e., states, and the ergodic structure makes the arbitrary combination of such segments match with arbitrary data. A threshold model is built by connecting the target patterns' states in an ergodic model, keeping their emission probability and self transition unchanged. Thus, the threshold model rejection mechanism is based on the discriminative power of the transition probabilities' dynamic range. A well known intrinsic problem of HMMs is the imbalance between the dynamic ranges of the transition and emission probabilities. Exposed by Rabiner and Huang [11], the phenomenon reveals the weakness of the transition probabilities's discriminative power. Our setup imposes the use of continuous features derived from raw sensor signals. As a consequence, a typical natural intake gestures pattern is represented by a large number of samples, i.e., gesture length range between 100 and 1500 samples. The need to model the temporal structure of signals with high sampling frequency and large gesture pattern variance is not satisfied by the exponential temporal distribution that characterises the standard HMM transitions. For long observation sequences, the exponential temporal distribution implies that the self transition a_{ii} for state s is asymptotically close to 1. The equiprobability of all transitions exacerbates the lack of discriminative power of the transition probabilities in path decoding. In fact, the discrimination among models is delegated to the emission probabilities. The loss of discriminative power among the target models and the threshold models could be due to the equiprobability of both transition and emission probabilities. In our method, diversity among target and rejection emission parameters is guaranteed by the independent training procedure.

In terms of training time, our method is significantly more complex than a threshold model. In Tab. 4, the time complexities for the framework's components are listed. The training effectiveness mainly depends on the number of generations

and the population size of the genetic algorithm. We fixed the number of generations to 100 to balance between computational time and performance. Longer training time could yield better results. The multipopulation genetic optimisation is well suited to be performed in parallel computation, thus drastically reducing computational time. In terms of prediction time, our method's complexity is comparable to the threshold model, being proportional to the network's number of states. In future work, we plan to study the effect of varying the hyperparameters of the framework as the number of generations, and the number of NTMs and TMs.

Some aspects that could alter the retrieval performances should be discussed. Firstly, we included specific annotated non-target gestures, i.e., food intake, in the HMM network, which likely improved the spotting performance. Nevertheless, the gesture-based NTM illustrate the features of our GMM-HMMs network approach. Secondly, we did not specifically model container type differences, which may have reduced performance and partly justify the need for personalised models. The personalised models additionally captured the variation in drinking styles, and personal environments of the participants. Thirdly, while sensor rotation and shifting at the wrist could have occurred, it did not reduce performance profoundly. We consider that the wrist position was well known to the participants from wearing other accessories, e.g. watches. Probably participants maintained the sensor strap in a comfortable wearing position as done with a watch.

Our work was intended to elaborate a spotting method that suits for sparse natural gesture patterns in free living data, thus, we limited our investigation to spotting gestures at the dominant arm.

CONCLUSION

In this work, we presented a GMM-HMM spotting network including a non-target gesture mining method and a genetic optimisation of the HMM probabilities, intended to find sparse natural gestures in free living data. Our method was evaluated in a daily life dataset collected in free living including totally 35 days of annotated recording of seven participants. Comparisons were done with a threshold model approach for gesture recognition. Our approach yielded an average F1-score of over 74%. Comparisons with a HMM threshold model confirmed the benefit of our approach for gesture spotting and illustrated the challenge that sparse gesture in free living data represent. We consider the methodology presented in this work a step towards sensor-based natural gesture spotting in free living. Moreover, the present work provides a method to fluid intake monitoring and could potentially help users to maintain healthy intake patterns.

REFERENCES

1. Oliver Amft. 2010. Adaptive activity spotting based on event rates. In *Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC), 2010 IEEE International Conference on*. IEEE, 169–176.
2. Oliver Amft, David Bannach, Gerald Pirkl, Matthias Kreil, and Paul Lukowicz. 2010. Towards wearable sensing-based assessment of fluid intake. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*. IEEE, 298–303.
3. Mahmoud Elmezain, Ayoub Al-Hamadi, and Bernd Michaelis. 2009. Hand trajectory-based gesture spotting and recognition using HMM. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 3577–3580.
4. Francisco Fernández, Marco Tomassini, and Leonardo Vanneschi. 2003. An empirical study of multipopulation genetic programming. *Genetic Programming and Evolvable Machines* 4, 1 (2003), 21–51.
5. Simon Günter and Horst Bunke. 2003. Fast feature selection in an HMM-based multiple classifier system for handwriting recognition. In *Joint Pattern Recognition Symposium*. Springer, 289–296.
6. Mithun George Jacob and Juan Pablo Wachs. 2014. Context-based hand gesture recognition for the operating room. *Pattern Recognition Letters* 36 (2014), 196–203.
7. Holger Junker, Oliver Amft, Paul Lukowicz, and Gerhard Tröster. 2008. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recognition* 41, 6 (2008), 2010–2024.
8. Hyeon-Kyu Lee and Jin-Hyung Kim. 1999. An HMM-based threshold model approach for gesture recognition. *IEEE Transactions on pattern analysis and machine intelligence* 21, 10 (1999), 961–973.
9. Jerry Mannil, Mohammad-Mahdi Bidmeshki, and Roozbeh Jafari. 2011. Rejection of irrelevant human actions in real-time hidden Markov model based recognition systems for wearable computers. In *Proceedings of the 2nd Conference on Wireless Health*. ACM, 8.
10. Bo Peng and Gang Qian. 2011. Online gesture spotting from visual hull data. *IEEE transactions on pattern analysis and machine intelligence* 33, 6 (2011), 1175–1188.
11. LR Rabiner and BH Juang. 1992. Hidden Markov models for speech recognition - Strengths and limitations. In *Speech Recognition and Understanding*. Springer, 3–29.
12. Michael N Sawka, Samuel N Cheuvront, and Robert Carter. 2005. Human water needs. *Nutrition reviews* 63, s1 (2005).
13. Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1029–1040.
14. Shilei Zhang, Zhiwei Shuang, Qin Shi, and Yong Qin. 2010. Improved mandarin keyword spotting using confusion garbage model. In *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 3700–3703.