

# An Intelligent Food-intake Monitoring System Using Wearable Sensors

Jindong Liu\*, Edward Johns\*, Louis Atallah\*, Claire Pettitt†, Benny Lo\*, Gary Frost† and Guang-Zhong Yang\*

\*The Hamlyn Centre, Imperial College London, United Kingdom

†The Nutrition and Dietetic Research Group, Faculty of Medicine, Imperial College London, United Kingdom

Email: {j.liu, ej09, l.atallah, c.pettitt, benny.lo, g.frost, gzy}@imperial.ac.uk

**Abstract**—The prevalence of obesity worldwide presents a great challenge to existing healthcare systems. There is a general need for pervasive monitoring of the dietary behaviour of those who are at risk of co-morbidities. Currently, however, there is no accurate method of assessing the nutritional intake of people in their home environment. Traditional methods require subjects to manually respond to questionnaires for analysis, which is subjective, prone to errors, and difficult to ensure consistency and compliance. In this paper, we present a wearable sensor platform that autonomously provides detailed information regarding a subject’s dietary habits. The sensor consists of a microphone and a camera and is worn discretely on the ear. Sound features are extracted in real-time and if a chewing activity is classified, the camera captures a video sequence for further analysis. From this sequence, a number of keyframes are extracted to represent important episodes during the course of a meal. Results show a high classification rate of chewing activities, and the visual log demonstrates a detailed overview of the subject’s food intake that is difficult to quantify from manually-acquired food records.

## I. INTRODUCTION

Obesity has become one of the main challenges facing the western healthcare systems and global economies. In 2009, the proportion of obese people in the USA adult population was 26.8% for women and 27.6% for men [18]. In the United Kingdom, 23.9% of women and 22% of men were recorded as being obese in the years 2008 to 2009 [8]. Globally, an estimated 500 million people are now classed as obese. Obesity is linked to many chronic diseases including diabetes, heart disease and cancer. Controlling this problem is an urgent yet challenging problem for the western countries, as well as some of the developing countries such as China and India. The cornerstone of modern public health policy is now to encourage a change in dietary behaviour.

Understanding an individual’s dietary behaviour, including food preference and consumption patterns, is one of the main steps in tackling the causes of obesity. Traditional methods for dietary assessments [19] require subjects to manually respond to information questionnaires. This relies on the users’ recall of their food-intake history and daily activities. The scope of questionnaires could span from 24 hours to several months, making accurate recall difficult. The method is time consuming, inaccurate and suffers from poor compliance. In a recent study, around 70% of users abandoned the long questionnaires before completion [6]. Furthermore, inaccurate reporting increases with body weight [15].

In addition to questionnaires and surveys, consumer electronic devices, such as PDAs and mobile phones, have been used to allow users to make immediate food-intake annotations. Bespoke devices such as wrist-worn bite counters are also used, where accelerometers are used to count how many bites that the user takes [16]. The number of bites is then used to estimate the caloric consumption. However, such estimation is highly dependent on the food and in many cases there is no information regarding the actual amount of food eaten. Some fitness monitoring devices such as Fitbit [7] can also track user activities and food consumption. However, users are still required to input the food information via online questionnaires. For these reasons, researchers are looking for more accurate methods requiring less user-involvement to assess general food-intake. Amft [1], for example, applied an ear-pad microphone to detect the chewing sound during food-intake and classify the sound from different types of food. However, the experiments show that it was hard to rely solely on acoustic properties to classify food in detail, although it works sufficiently well within a restricted set of food types. In real life situations, the accuracy of such methods is hampered by the background noise and other simultaneous sound sources. A solution to that could be the use of Electromyography sensors to detect the swallowing action [3]. However, such a method would require the user to wear a sensor collar around the neck, which is not convenient for daily monitoring. The above methods have several disadvantages and are difficult to apply to studies of large populations. Both paper-based and digital dietary questionnaires are dependent on user motivation, literacy, and self-awareness. Current automatic dietary monitoring system using accelerometers and microphones can reduce user involvement. However, such systems still cannot give a complete assessment of food-intake history. In this paper, we propose an intelligent food-intake monitoring system that can automatically detect eating activities. We combine an in-ear microphone with a miniature camera in a light-weight wearable headset. The sound from the microphones is classified in real-time into different eating activities and the camera takes snapshots of the food if a chewing activity is detected. The key images of food are then selected sequentially and a dietary assessment log is generated to reflect a user’s dietary behaviour.

The novelty of our method can be summarised as: (i) developing a noise-resilient sound activity detection method



Fig. 1: Left: a subject wearing the food-intake sensor during lunch. Right: the profile of the sensor.

suitable for daily use, (ii) introducing food images into the dietary assessment to improve the assessment accuracy; (iii) selecting key images automatically to minimise the size of the food-intake assessment log.

In the remainder of the paper, we will introduce the system structure in Section II, together with presenting the sound activity detection and key image selection methods. Preliminary experimental results are shown in Section III, where our method shows feasibility in realistic situations and robustness to environmental noise. Conclusions are given in Section IV.

## II. SYSTEM MODEL OF FOOD-INTAKE MONITORING SYSTEM

In general, food-intake includes four sub-activities: (i) food preparation and ingestion, such as cutting food on a plate; (ii) food breakdown in the mouth (chewing); (iii) bolus transport (swallowing and oesophageal movement); and (iv) gastric activity (stomach movement). In this paper, we will only focus on the monitoring of the first two activities using a wearable sensor. The food-intake monitoring sensor shown in figure 1 was designed to have a miniature camera (standard mobile phone camera) and a microphone (Sony ECM TL3). Sound is recorded at a sample rate of 44.1 kHz, and images are captured at a resolution of 640 by 480 at 30 frames per second.

The sensor is wireless and light-weight making it easy to monitor long-term daily activities with comfort akin to that of a hearing aid. When a subject wears the sensor, the camera is directed towards the table to take images of the food container. The microphone is placed just outside the ear canal to measure sound propagation. The reason for using the ear as a location for the sensor is threefold: (i) because of its stability despite motion, and the possibility of integrating accelerometers which can enable accurate detection of activity and energy expenditure [4]; (ii) acoustic information from chewing is less susceptible to environment noise because the chewing sound wave is transmitted through the skull [2]; (iii) the camera has a similar viewing angle to the subject’s eyes

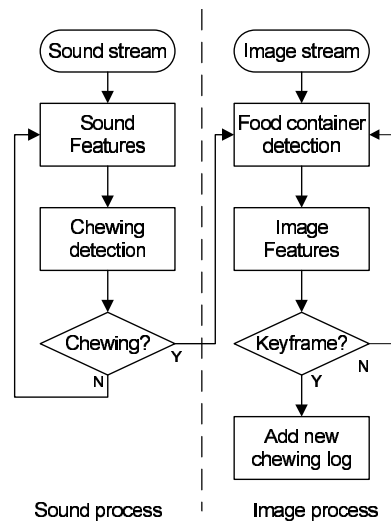


Fig. 2: Schematic system structure.

therefore it enables the capture of more realistic images of the food.

Figure 2 shows the schematic structure of the proposed system. It mainly includes two processes: (i) the process of detection of chewing activity from sound, and (ii) the process of “keyframe” detection from images for the visual log. Chewing activity detection runs continuously and keyframe detection is only activated once triggered by a successful detection of chewing activity. For chewing detection, the sound stream from the microphone is analysed using overlapping time windows and features are extracted from the sound. A neural network classifier using extreme learning machines [12] is then applied to classify the sound into one of four basic activities: speech, chewing, drinking and others. It is worth noting that chewing activity is recognised from both the chewing sound and that of preparation such as cutting using a knife. The details of the chewing activity detection method are explained further in section II-A. As soon as a chewing activity is detected, the keyframe detection algorithm is triggered and continues until the chewing activity is not detected for a prolonged period of time. Due to the highly deformable nature of food and the cluttered background, it is very difficult to recognise food from images directly. Instead, we assume that the food is contained in a circular container such as a plate which is relatively easier to recognise. Therefore, if a container is detected, we can assume there is food in the container because the visual processing is triggered only when a chewing activity is detected. Consequentially, a new food-intake log is written to a file (see Figure 3) which includes a series of food snapshots and time-stamps to outline the consumption history. The description of the image processing is given in Section II-B.

### A. Sound Feature Extraction and Recognition

Sound in real life could result from a variety of sources such as speech and ambient noise. One typical recording is shown in Figure 4. It is evident that it is difficult to identify chewing







Snapshot	Time	Consumption (%)
	18.33	0
	18.35	20
	18.37	40
	18.39	60
	18.41	80
	18.45	100

Fig. 3: An example of food intake log

activities from this and impossible to determine the identity of the food type from the sound alone. In Amft's work [1], food types were grouped into three clusters: wet loud and soft quiet. However, the experiments were conducted in a quiet, controlled environment rather than real life scenarios. In our research, we introduce vision on a single wearable sensor which allows for a simplified sound processing system. We use sound to classify only four activity types: speech, eating, drinking and others in order to ensure a robust classifier.

1) *Sound Feature and Feature selection*: From continuous sound, we calculate three types of sound features for every three second sound interval  $x(t)$ . These features include (i) energy features, such as energy entropy  $EE(i)$ , and short time energy  $STE(i)$ ; (ii) spectral features, such as spectral roll-off  $SRO(i)$ , spectral centroid  $SC(i)$ , spectral flux  $SF(i)$  and spectral average of sub-bands  $SA(i, j)$ ; and (iii) temporal features, e.g. zero crossing rate  $ZCR(i)$  and peak gap between two neighbored local maximal energy peaks  $PG(i)$ , where  $i$  is the time index and  $j$  is the frequency sub-band index. Both energy features and spectral features are calculated on every 500 ms non-overlapped time windows within  $x(t)$ . Statistical factors are calculated over all three types of features such as maximum  $\max(\cdot)$ , minimum  $\min(\cdot)$ , standard derivation  $\text{std}(\cdot)$  and mean value  $\text{mean}(\cdot)$  etc. There are in total 76 features for every 500 ms and 456 feature over a typical 3s  $x(t)$ . In order to speed up the processing for real-time situations, we applied the following three approaches:

- Selective processing on local energy peaks. As the chewing sound is accompanied by a biting sound between teeth, which is typically manifested as a peak with

short time energy  $STE(i)$ . Therefore, we can selectively calculate the feature around the local peaks to reduce the calculation time.

- Salient feature selection to reduce the number of features. In this work, we followed the approach in [4] and compared three types of feature selection algorithms. These were: Relief, Simba (margin based feature selection) [10] and mRMR [14]. More information on these algorithms is included in [4]. The three algorithms were used to compare feature saliency and the following features showed to be the most discriminant between classes:  $\text{mean}(EE(i))$ ,  $\max(ZCR(i))$ ,  $\text{mean}(ZCR(i))$ ,  $\text{mean}(SC(i))$ ,  $\min(SF(i))$ ,  $\max(SF(i))$ ,  $\text{std}(SF(i))$ ,  $\text{mean}(SF(i))$ ,  $\max(STE(i))$ ,  $\text{mean}(SA(i, j))$  and  $\text{std}(SA(i, j))$ . The total number of features is reduced from 76 to 27.
- Extreme Learning Machines for quick learning and classification. An Extreme Learning Machines (ELM) randomly generates all the hidden-node parameters of generalized Single-hidden Layer Feed-forward Networks (SLFNs) and analytically determines the output weights of SLFNs [11]. Compared to the popular Back-Propagation (BP) Algorithm and Support Vector Machine (SVM) / Least Square SVM (LS-SVM) classifiers, ELM has several advantages, such as faster learning and greater generalization. We will briefly present the application of ELM in the following section.

2) *Extreme Learning Machine for classification*: For  $N$  arbitrary distinct samples  $(\mathbf{x}_i, \mathbf{t}_i)$ , where  $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in \mathbf{R}^n$  is the input, the sound feature vector in our case, and  $\mathbf{t}_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in \mathbf{R}^m$  is the desired output, the targeted sound type ( $m=4$ ) in our case, the output of a single layer ELM with  $\tilde{N}$  hidden neurons and a non-linear kernel function  $g(x)$  is modeled as:

$$\sum_{i=1}^{\tilde{N}} \beta_i g(\mathbf{w}_i \cdot \mathbf{x}_j + b_i) = \mathbf{o}_j, j = 1, \dots, N, \quad (1)$$

where  $\mathbf{w}_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$  is the weight vector connecting the  $i$ th hidden neuron and the input  $\mathbf{x}_j$ ,  $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$  is the weight vector between the  $i$ th hidden neuron and the output neurons.  $b_i$  is the bias of the  $i$ th hidden neuron.  $\cdot$  indicates the dot product. The main advantage of the ELM over classical artificial neural networks is that  $\mathbf{w}_i$  and  $b_i$  are randomly generated rather than iteratively learned. Therefore, the training time is greatly deduced. Given a set of input  $\mathbf{x}_i$  and the corresponding desired output  $\mathbf{t}_i$ , the training target is now to find  $\beta_i$  in Equation 1 such that  $\sum_{j=1}^{\tilde{N}} \|o_j - t_j\| \rightarrow 0$ . According to the work in [11], the unique smallest norm least-squares solution is:

$$\hat{\beta} = \mathbf{H}^\dagger \mathbf{T} \quad (2)$$

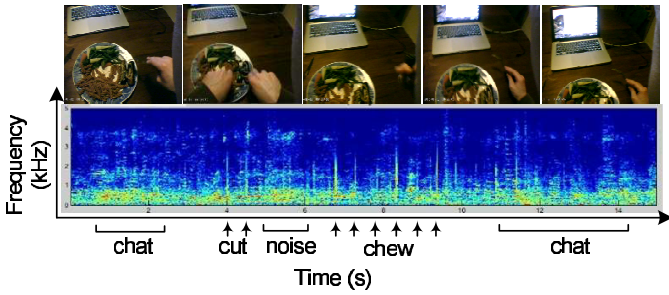


Fig. 4: An example of 15s-long input signals.

where  $\mathbf{H}^\dagger$  is the Moore-Penrose generalized inverse of hidden layer output matrix  $\mathbf{H}$  [17].

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_1 + b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}} \quad (3)$$

$\mathbf{T}$  is the matrix of the desired output,

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ \vdots \\ \mathbf{t}_N^T \end{bmatrix}_{N \times m} \quad (4)$$

and  $\hat{\beta}$  is the approximation of the output weight  $\beta$ ,

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_N^T \end{bmatrix}_{N \times m} \quad (5)$$

. In details,  $\mathbf{H}^\dagger$  is calculated as:

$$\mathbf{H}^\dagger = ((\mathbf{I}_N/C) + \mathbf{H}\mathbf{H}^T)^{-1} \quad (6)$$

, where  $C$  is customizable regularisation factor and  $\mathbf{I}_N$  is  $N$  by  $N$  identity matrix. Once we calculate  $\hat{\beta}$ , a classifier is ready for use by replacing  $\beta$  in Equation 1 with  $\hat{\beta}$ .

### B. Keyframe Detection

Keyframe detection from image sequences has been widely addressed in fields such as medical imaging [9], robot navigation [20] and lifestyle logging [5]. Typically, local features representing texture corners are first extracted from each image. Then, an intra-image distance function based upon feature correspondences is used to compute pairs of frames over which there has been a significant change in appearance. Whilst this method provides good results when the entire content of an image is of importance, such as in robot navigation, in our case we are only interested in the area of the image representing the plate. Applying standard keyframe detection to our image sequences would trigger keyframes every time the subject looks away from the plate, due to rapid changes in scene content from the introduction of the background scene, rather than due to the more subtle changes from the food consumption. Furthermore, keyframe detection using local features is highly sensitive to the texture of the image and

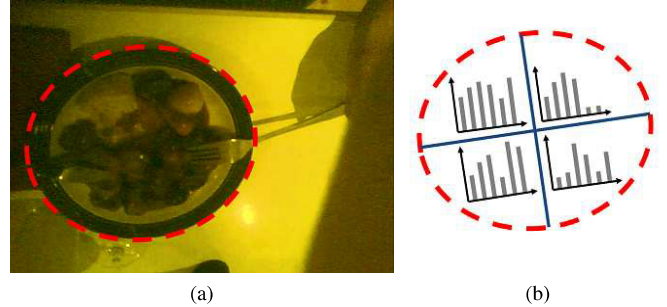


Fig. 5: Method for comparing food content between images. In (a), an ellipse is detected, representing the plate. Then in (b), four colour histograms are computed, one for each quadrant of the ellipse. Adjacent images are then compared by computing histogram distances.

works best when keypoints are distributed uniformly across the image. In our case, the highly variable appearance of different food types, together with the deformable nature of food over the course of a meal, result in unstable distributions of keypoints that do not directly reflect the removal of food from the plate.

As such, we propose an alternative solution by first locating the plate in an image, and then using the colour content within the plate, rather than texture, to detect the removal of food. Given a chewing activity is detected by the wearable sensor, a video sequence is initiated by the sensor's camera. Each image is first preprocessed to search for ellipses [13], and any ellipse with dimensions that fall within a heuristically-defined range is considered the plate of interest. Then, the ellipse is divided into quadrants, and a normalised colour histogram is computed for each quadrant. The C-color space [6] is used owing to its stability over illumination changes (shadows are often found due to interference from the subject's body), and 1000 bins (10 for each colour channel) are used for the histogram. The change in colour content of two images is then computed by summing, across all quadrants, the Euclidean distance between respective colour histograms. Figure 5 demonstrates the ellipse detection and colour histogram extraction for a sample image. For compiling a visual log of food intake, it is necessary to specify the granularity of the log, i.e., the amount of food consumed before a new keyframe is extracted. In practice, this would be determined by the particular system, and reflects a compromise between detail and brevity of the log. One approach would be to specify a threshold in the colour histogram distance between consecutive images, whereby exceeding this threshold indicates sufficient removal of food for a keyframe to be extracted. However, this parameter would require parameter tuning for each meal based on factors such as the food types and illumination conditions. We propose a more generalisable approach whereby the single parameter  $n$  is chosen, representing the number of keyframes to be recorded in the log. This allows a consistency across all food logs that does not require any prior knowledge about

TABLE I: Comparison between ELM and MLF

	ELM		MLF	
	N=500	N=2000	N=500	N=2000
Train Time (s)	0.26	7.97	15	105
True Positive (%)	71.6	70.5	64.7	68.1
False Positive (%)	28.4	29.5	35.3	31.9

conditions. We select keyframes by summing the inter-image colour histogram distances across the entire image sequence, and extracting images representing the  $(100 \times \frac{k}{n})^{th}$  percentiles of this summation for  $k = 0 \dots n$ . In this way, the removal of food from the plate is reflected equally across the full set of keyframes, whilst also focusing on episodes when food has been removed from or significantly disturbed upon the plate.

### III. EXPERIMENTAL RESULTS

Experiments were implemented in a real world rather than controlled environments. Six subjects were invited to wear our device to record videos during lunch lasted for up to 30 minutes. The recording was taken in an university staff restaurant with a capacity of 300 people (check exact size). The only restriction to the current experiments is that subject uses a circular plate or bowl to contain the food. The recordings were annotated manually and then used to train and test an ELM classifier. We use 60% of annotated data for training and 40% for testing. The sound is recorded at 44100 Hz and the video sequence is recorded by 30 frames per second.

#### A. Classification Using Sound

We first evaluate the efficiency of ELM and compared it to a classical multilayer feedforward neural network (MLF) which applies Back-Propagation (BP) Algorithm to tune the parameters in its hidden layers. We set the number of hidden neurons  $\tilde{N}$  identical in both ELM and MLF to compare the training time and recognition rate. The non-linear kernel function  $g(x)$  in Equation 1 is set as the sigmoid function and the value of  $C$  in ELM's equation 6 is arbitrarily set to 20. (We will show later that ELM's performance is not sensitive to the setting of  $C$  and  $\tilde{N}$ ). The comparison results are shown in Table I. The ELM out-performs the MLF on both recognition rate and training time. It is notable that the average recognition rate in table I is not particularly high, but this is to be explained later.

We then compared the performance of our classifier with and without feature selection. The feature selection procedure reduced the number of feature to calculate from 76 to 27. The results are shown in Table II, indicating that the computational time of sound features is cut by over 4 times, and the recognition rate is slightly improved after the feature selection.

To test the robustness of our method to the background noise, we recorded a data set (6 subjects, labeled as Set A) in a relatively quiet environment in contrast to the real life data (labeled as Set B) in the catering restaurant. The results are shown in Table III. This shows that the performance of ELM is reduced by about 9% even though the level of

TABLE II: The time saving of feature selection

	Non-Feature Selection		Feature Selection	
	N=500	N=2000	N=500	N=2000
Feature Calc Time (ms)	542	542	122	122
Train Time (s)	0.35	6.20	0.26	7.97
True Positive (%)	66.7	68.6	71.6	70.5

TABLE III: The robustness of ELM over background noise

	Background Noise (dB)	True Positive (%)	
		ELM	MLP
Data Set A	22	80.4	84.0
Data Set B	51	71.6	64.7

background noise is increased by over 30 dB. In contrast, the MLP method's performance decreased by almost 20%.

In many practical applications, choosing a model's parameters is done in an empirical manner. However, the performance of the model is likely to be sensitive to the chosen parameters. To study the sensitivity of the ELM classifier, we tested the model over a range of  $C$  and  $\tilde{N}$ . See Figure 6. This indicates that the ELM's performance is very stable once the  $C$  is larger than 1 and  $\tilde{N}$  larger than 50.

To investigate the reason for the achieved recognition rate, we studied the confusion matrix of ELM classifier among four classes. See Table IV. It is notable that the class of drinking has a very low recognition rate. This is because the drinking sound is much quieter than the sound in the other classes and is suppressed by the loud background noise.

#### B. Key Frame Detection and Food Intake Log

From the video captured by the wearable camera, the method in Section II-B was then used to compute keyframes for the food intake log. Figure 7 shows examples of the log with two values of  $n$ , demonstrating two logs of different granularity levels. In this example, the logs show a balanced consumption of food across the course of the meal, but with patients exhibiting unusual eating habits the log will provide feedback on factors such as the order in which food types are eaten, the rate of consumption, and the overall proportion of a meal that is eventually consumed.

The qualitative and subjective nature of keyframe interpretation provides a challenge in evaluating our keyframe detection method. For the most practical evaluation, sets of keyframes would be analysed by doctors or nutrition experts to determine the relative usefulness of different sets in analysing a patient's dietary habits. However, this is an intensive procedure and remains susceptible to bias and human error unless a significantly, perhaps prohibitively so, large set of test subjects

TABLE IV: Confusion Matrix of ELM classifier

	Target (%)			
	Eating	Drinking	Speech	Others
Eating	82.51	0	0.13	17.36
Drinking	28.57	24.18	1.10	46.15
Speech	4.22	0	81.93	13.6
Others	13.07	0	1.32	85.61

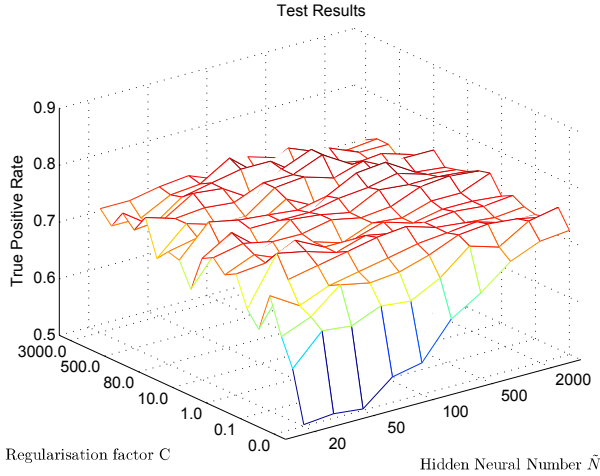


Fig. 6: The effect of hidden neural number and regularisation factor on ELM classifier. (Note: The kernel function  $g(x)$  is sigmoid function)

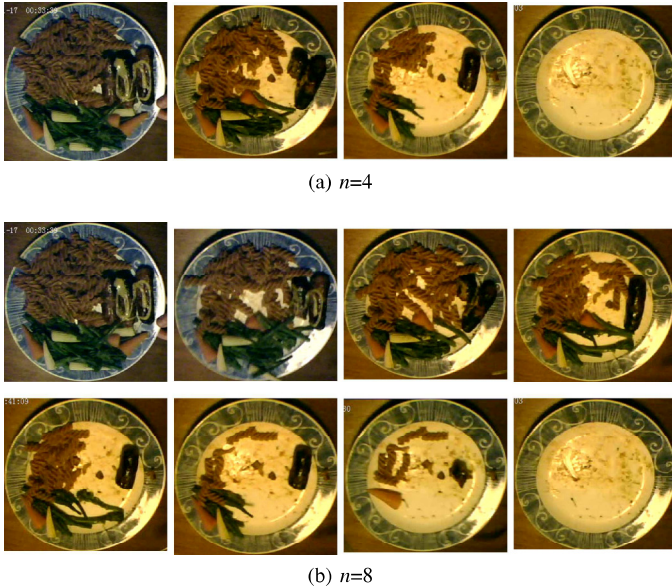


Fig. 7: Keyframe detection results for two values of  $n$ . The keyframes form a food intake log which can then be used to monitor dietary habits of a patient.

are used. As such, we evaluate our method by investigating the ability of our method to determine the proportion of food consumed across the image sequence. Keyframe detection is reliant on finding pairs of frames between which a significant amount of food has been removed. As such, a system reliably predicting the proportion of food removed between frames indicates that the system is also suitable for keyframe detection.

The estimation of the actually-consumed portion of food is based on the technique in II-B, where the histogram distance between two adjacent images is divided by the cumulative histogram distances across all pairs of adjacent images. This proportion of histogram distances at any given frame then

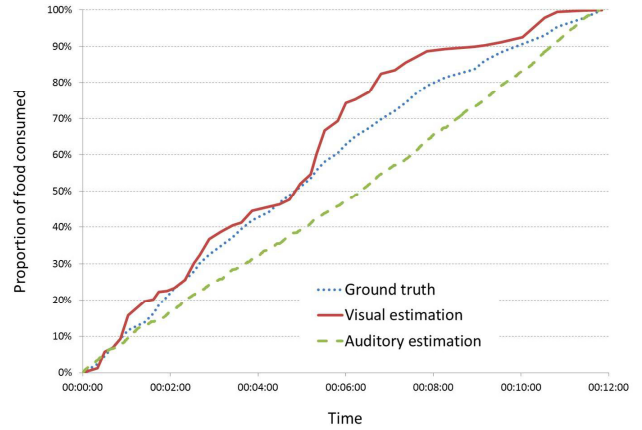


Fig. 8: Evaluating the performance of keyframe detection in estimating the proportion of food consumed.

gives an indication of the proportion of food consumed at that time. The ground truth was generated by recording the frame number at which each mouthful was consumed, and ensuring, to the best ability of the test subject, that the mouthfuls were of equal portions. At mouthful  $a$  out of a total of  $b$  mouthfuls, the ground truth proportion of food consumed is then  $\frac{a}{b}$ .

We also evaluated the ability of the sound recording to estimate the proportion of food consumed. This was done by assuming the each detected chewing activity (i.e. every bite of the jaw) was responsible for an equal volume of food. Thus, at chewing activity  $c$  out of a total of  $d$  chewing activities, the auditory estimation of food consumed is then  $\frac{c}{d}$ .

Figure 8 shows the comparison of the vision-based and auditory-based estimations to the ground truth for a sample meal. The high correlation suggests that our methods are reliable estimators of overall food consumption. Whilst the auditory data perhaps gives a better estimation of the subject’s chewing rate, the vision-based estimation is a better indicator of the actual rate of removal, and thus consumption, of food from the plate. The rate of intake of food available from Figure 8 is an important measure that can be acquired from wearable body sensors, but not reliably from manually-recorded logs.

### C. User Feedback

We further investigated the applicability of the sensor in real-world scenarios by conducting a user-feedback survey. For wearable sensors it is naturally very important to achieve high levels of comfort and discretion if they are to be welcomed by patients. Ten of the trial subjects were chosen and asked to respond to a questionnaire on matters reflecting the suitability of the sensor as a realistic alternative to manually-recorded logs. Figure V shows the questions asked and the average responses, together with the variances. For questions of a “yes or no” nature, 1 indicates a “strong no” and 5 indicates a “strong yes”.

The results show encouraging levels of satisfaction both in ease of fixation and lightness of the device, indicating the

TABLE V: User Feedback Score

Feedback questions	Mean/variance (1~5)
Does the device fit easily on the ear?	4.1/0.54
Is the device comfortable to wear?	2.9/0.99
Is the device light enough to wear?	4.5/0.28
Does the device effect your eating?	2.4/1.38
How many hours a day would you be prepared to wear the device?	<1h (3), 1~3h(2) 3~5h (2), >5h(3)

suitability of physically combining a microphone and camera into wearable sensor for this task. Whilst the comfort level had only an average response, the device itself was modified from an off-the-shelf purchase which did not focus design around comfort. The effect of the device on eating also receive an average response, but due to the high variance in these responses and the fact that each subject only wore the device for one meal, this perhaps would become less of a problem as accustomisation prevailed. The final row of the table suggests that many patients would be willing to wear the device for a significant amount of time if analysis of the data could help with dietary treatment, an encouraging response on the device's overall suitability.

#### IV. CONCLUSIONS

In this paper, we have demonstrated a novel wearable body sensor to monitor a subject's food intake behaviour. We have combined an in-ear microphone with a miniature camera to form an integrated multi-sensor wearable device. Chewing activities are detected from the microphone which consequentially triggers the detection of keyframes by using the integrated miniature camera. This compiles a visual log of the food intake to provide a doctor with a summary of eating behaviour characteristics, such as consumption speed, preference of certain food types and overall consumption levels.

We target our system to work in real life scenarios where dynamic background noise and simultaneous sounds exist. Several efforts have been taken to improve the computational efficiency and noise robustness of the system. For instance, we reduced the sound features from 76 to 27 by feature selection process to shorten the computation time. Extreme Learning Machines was used due to the efficiency of training and robust to the noise. The experimental results show that three activities can be successfully detected from sound with over 80% detection rate, but the drinking activity failed to be detected due to its low sound level and suppression by the background noise. Keyframes were detected by searching for ellipses representing plates, and comparing colour histograms between adjacent images. The visual logs demonstrate a clear overview of the consumption of food, and we further showed that the proportion of food consumed can be reliably estimated using inter-image colour histogram distances. The results derived demonstrate the practical value of the proposed system.

#### REFERENCES

[1] O. Amft. A wearable carpad sensor for chewing monitoring. In *Sensors, 2010 IEEE*, pages 222–227, nov. 2010.

[2] Oliver Amft, Mathias Stger, and Gerhard Trster. Analysis of chewing sounds for dietary monitoring. In *In UbiComp 2005*, pages 56–72. Springer, 2005.

[3] Oliver Amft and Gerhard Troster. Methods for detection and classification of normal swallowing from muscle activation and sound. In *Pervasive Health Conference and Workshops, 2006*, pages 1–10, 29 2006-dec. 1 2006.

[4] Louis Atallah, Julian J.H. Leong, Benny Lo, and Guang-Zhong Yang. Energy expenditure prediction using a miniaturized ear-worn sensor. *Medicine & Science in Sports & Exercise*, 43(7):2688–2710, Jul 2010.

[5] M Blighe, A. Doherty, A. F. Smeaton, and N. E. O'Connor. Keyframe detection in visual lifelogs. In *Proc. PETRA*, 2008.

[6] Gertjan J. Burghouts and Jan-Mark Geusebroek. Performance evaluation of local colour invariants. *CVIU*, 13(1):48–62, 2009.

[7] Fitbit Co. <http://www.fitbit.com>, 2011.

[8] EUROSTAT. Overweight and obesity - BMI statistics. Technical report, 2011.

[9] S. Giannarou and G-Z. Yang. Content-based surgical workflow representation using probabilistic motion modeling. In *Proc. MIAR*, pages 314–323, 2010.

[10] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Margin based feature selection - theory and algorithms. In *Proceedings of the twenty-first international conference on Machine learning, ICML '04*, pages 43–, New York, NY, USA, 2004. ACM.

[11] Guang-Bin Huang, Dian Wang, and Yuan Lan. Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, 2:107–122, 2011. 10.1007/s13042-011-0019-y.

[12] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[13] T. M. Nguyen, S. Ahuja, and Q. M. J. Qu. A real-time ellipse detection based on edge grouping. In *Proc. Systems, Man and Cybernetics*, 2009.

[14] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, aug. 2005.

[15] A. M. Prentice, A. E. Black, W. A. Coward, and T. J. Cole. Energy expenditure in overweight and obese adults in affluent societies: an analysis of 319 doubly-labelled water measurements. *Eur Journal Clin Nutr*, 50(2):93–97, 1996.

[16] Jenna L. Scisco, Eric R. Muth, Yujie Dong, and Adam W. Hoover. Slowing bite-rate reduces energy intake: An application of the bite counter device. *Journal of the American Dietetic Association*, 111(8):1231–1235, 2011.

[17] Denis Serre. *Matrices: Theory and Applications*. Springer-Verlag New York, 2002.

[18] U.S. Department of Health and Human Services. Summary health statistics for u.s. adults: National health interview survey 2009. *Vital and Health Statistics*, 10(249):1–217, December 2010.

[19] J. C. Witschi. Short-term dietary recall and recording methods. *Nutritional Epidemiology*, 4:52–68, 1990.

[20] H. Zhang, B. Li, and D Yang. Keyframe detection for appearance-based visual slam. In *Proc. IROS*, pages 2071–2076, 2010.