A Privacy-Preserving Wearable Camera Setup for Dietary Event Spotting in Free-Living

Giovanni Schiboni∗, Fabio Wasner†, Oliver Amft∗
∗Chair of eHealth and mHealth, FAU Erlangen-Nürnberg, Germany
Email: giovanni.schiboni@fau.de, oliver.amft@fau.de
†University of Passau, Germany

Abstract—We designed a wearable head-mounted ego-centric camera setup for dietary data collection in free-living. We addressed the problem of privacy-sensitive image content by fixing a camera on a cap’s visor pointing downwards. Salient content was maintained while drastically constraining unwanted privacy-infringing content. The privacy preservation capability of our setup was compared with literature using a modified privacy-saliency matrix. Furthermore, we implemented a dietary event spotting algorithm to reduce the amount of workload for human operator while performing analysis on a large volume of data. Transfer learning on a deep neural network was employed to perform dietary object detection and, subsequently, dietary event spotting. Average recall performance over 90% suggested the feasibility of the method.

I. INTRODUCTION

To provide high-quality dietary assessment of free-living behaviour is the first step for implementing any effective diet coaching and weight management program. Monitoring dietary activities via self-reporting techniques, e.g., questionnaires, proved to be a non-reliable source of information, providing fragmented information due to manual logging labour requested from respondents and high cognitive load. Manual logging procedures cannot cope with the complexity that characterises a dietary activity. Dietary dimensions as time, i.e., length of a meal, in-between time and counting food (fluid) intakes, as context, i.e., physical position, concurrent activity with eating, as content, i.e., food type, quantity and nutritional values, are difficult to recall from the respondents, and, in the worst case scenario, impossible. A continuous direct human observation is an expensive and non-scalable solution. Exploiting an unobtrusive wearable camera-based solution to monitor dietary activities and to provide relevant information from free-living setting may address the challenges in automatic dietary monitoring [1]. Wearable cameras may help to understand intake, compensate underestimated food recall from self-reports and improve accuracy of dietary assessment. The images of participant’s dietary activities can subsequently get manually analysed and annotated to extract knowledge.

In this paper, we present our wearable head-mounted egocentric camera setup for dietary event spotting in free-living. The following contributions are provided.

1) We fixed a camera on a cap’s visor pointing downwards, in a way that dietary activities could be monitored, at the same time excluding the surrounding environment, including people, from the field of view. Collection of privacy-infringing content was minimised while keeping a high degree of eating behaviour evidence. The amount of privacy-sensitive content was quantified by using a Privacy-Saliency Matrix (PSM) metrics.

2) We implemented a dietary event spotting algorithm that identifies dietary objects in the video sequence. Transfer learning applied to a deep neural network was employed to perform dietary object detection from a dataset collected in free-living conditions.

II. RELATED WORK

Several recent works employed wearable cameras to reduce the under-reporting effect of 24-h recall methods. Results from O’Loughlin et al. [3] demonstrated that more accurate estimate of total energy intake is provided if conventional food diary methods are integrated with a chest-mounted camera-retrieved content. A similar study, with same conclusion, was carried out by Gemming et al. [4]. Images were automatically captured by a neck-mounted camera in response to movement, heat and light (every 20-30 s). The same research group, Gemming et al. [5], provided a complementary tool to analyse and annotate environmental and social context information by human agents. They used the same SenseCam setup and three image-assisted multiple-pass 24-hour dietary recalls to assess eating behaviours. Pettitt et al. [6] implemented a wearable sensor platform composed by a micro-camera and a microphone worn on the ear. Video images of food consumed were captured from the camera and sounds, transmitted through the jaw when eating and drinking. Accuracy was investigated by comparison between assessment methods. The user had to turn on the camera before any dietary activity.

Our objective was to provide a tool to mitigate issues in data collection for automatic dietary assessments in free-living. We did not perform energy intake estimation, instead we focused on the design of the data collection technical setup. We demonstrate that a specific choice of
camera position and angle of view can drastically reduce the need for censoring techniques, e.g., face detection, and image cropping, which are often expensive and errorprone. Moreover, we introduce a deep-learning based spotting approach to identify dietary objects in the footage.

III. Concept

A. Privacy Preservation

Choosing a camera position and viewing direction in ego-centric setups is challenging. On one side there is the necessity to monitor any dietary episode without losing critical information. For this reason, the field of view of the camera should be centred entirely on the eating scene. On the other hand, the camera, pointing forward the user, raises privacy issues when in public spaces. Specifically, being filmed by an unknown device, for an unknown purpose, is a particularly undesired condition for many.

Fixing the camera on the cap, pointing downwards, allowed us to reach an acceptable trade-off between privacy infringement and information retrieval. Fig. 1, shows our wearable camera setup, and Fig. 2 shows the field-of-view. In our evaluation we considered as privacy infringing content those images containing faces of secondary subjects, screens of mobile phone, and computer screen. Other variables could be considered in a larger study, e.g., bank access credentials during payments, filling documents and tattoo visible on body parts. As a consequence of the privacy-preserving configuration, the main challenge is to recognise dietary objects correctly. In fact, the images, captured from the chosen position, present unique features which are not comparable to publicly available image dataset for machine learning purposes.

B. Dietary event spotting

A main feature of free-living data is a long recording duration. A free-living monitoring setup is conceivably to work uninterruptedly for the entire day. The extended length of video recordings implies an enormous amount of data to be manually analysed and/or annotated by human expert. The process is time-consuming, error-prone and tiresome. We observed that the time period containing salient information is just a little fraction of the total monitoring duration. A method to filter out uninteresting data and speed-up the analysis process by reducing the human workload is required. Our assumption is that a dietary event is characterised by the usage of dietary objects. We implemented an event spotting algorithm to automatically identify dietary objects in the video sequence.

In order to detect dietary objects, we employed deep neural networks (DNN), specifically deep convolutional networks. We employed DNN since they represent the state-of-the-art in computer vision for object detection and recognition. The main problem that raises by using DNN is how to train the network. The DNN requires many training instances in order to work accurately. Collecting reference and training data is an expensive and laborious task. A good practice is to exploit pre-trained networks by using publicly available datasets. Unfortunately, our setup is dissimilar, regarding the angle of view, and thus the features, from other datasets in the same knowledge domain. A specific training set was therefore required in addition to existing datasets. We recorded a four days of ego-centric video material from a participant in free-living using the recording setup described below. Then, we adopted a transfer learning paradigm in order to transfer models that were pre-trained using a public dataset to our new setup. Transfer learning is a method for pre-training a model in a specific domain and transferring the model then to another domain. Transfer learning is used when dealing with time, data and computational constraints.

IV. Methods

A. Privacy-Preserving Configuration

Our egocentric camera system is composed of a Raspberry Pi Zeros combined with a camera module attached to a standard cap, all powered by a portable power bank with a capacity of 26.8 Ah (ANKER, Model A1210). The power bank was positioned in a pants back pocket or in a backpack, depending on preference. Our camera recorded in a resolution of 640x640 with around 28-30 fps. The angle of view (AOV) was 54 × 41 degrees. In order to estimate the field of view (FOV) of the camera, we applied the following equation:

$$\text{FOV} = 2\left(\tan\left(\frac{\text{AOV}}{2}\right) \times d\right)$$

were $d$ is the distance from the lens to the scene. As examples, let us consider two scenarios. If a person, 1.80 m tall, is wearing the cap while standing, the field of view is approximately 2.44 m². If the same person is sitting at a 0.72 m high table, the field of view is approximately 0.28 m². In both cases, it is really unlikely, that other people could be filmed by the camera. As quantified in next section, our configuration avoids privacy-infringing content captured in the videos.

Fig. 1: Our wearable head-mounted setup. The actual setup consists of two cameras. For our experiment, the red circled one was used and the crossed out one was switched off.
Fig. 2: Field of view of our head-mounted wearable camera setup.

B. Deep Neural Network: Yolo9000

We used the Darknet framework [7] which implements an easily adjustable platform for training and evaluation of network-designs on the dataset of choice. As network structure, we employed Yolo9000 [8], [9], a network that learns to predict and classify objects in images, while still maintaining high speed and performance. Girshick et al. [10] introduced a pipeline for object detection. Further developments led to networks that first joined the training process [11] and then also joined the weights for detection and classification of objects [12]. The Yolo9000 structure simplifies the detection by presenting a new way of predicting the object borders which produces less output and thus makes further classification of the object proposals faster [9].

C. Transfer Learning

In transfer learning, a base network on a base dataset is trained, and then the learned features are repurposed, or transferred, to another network, that is trained on a target dataset. We employed a pretrained network on a very large dataset, i.e., ImageNet [13] which contains 1.2 million images with 1000 categories, and then we used the same network as an initialisation for the training with our dataset. We followed the so-called fine-tuning, that implies not only replacing and retrain the classifier on top of the network, but also fine-tune the weights of the pretrained network by continuing the backpropagation. All the training sessions, for which results are reported later on, had the following fixed parameters: learning rate $= 10^{-5}$, momentum $= 0.9$, decay $= 5 \times 10^{-5}$.

D. Spotting

Individual frames were associated with a binary value:

$$ F_t = \begin{cases} 1, & \text{if } N_o \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad (2) $$

where $F_t$ is the binary value associated with frame in position $t$, $N_F$ is the total number of frames, and $N_o$ is the number of correctly identified dietary objects. A simple heuristic was chosen to characterise a true positive. If three or more consecutive frames had $F_t = 1$, a dietary event was considered to be spotted. The short duration, i.e., three seconds, was chosen in order to not miss the shortest dietary events as sip of water or single food intake. We compared class labels between ground truth and detected instances.

E. Daily Life Data Collection

A male student volunteer aged under 25 years wore our wearable setup during normal daily routine for about four days. The average recording time per day was 7.5 hours. The setup was attached in the morning and kept on until evening. The cap was taken off any time a hygiene routine was carried out. Video material of four days in free-living were collected. The user was free to carry out usual activity routines, as attending academic courses, having lunch at mensa, playing piano, commuting by public transport, and so on.

Firstly, the video was preprocessed by cropping 30 FPS output video to only one image per second. The downsampling reduced redundancy and facilitated the annotation process. Gaussian blurring was used as de-noising procedure. A total of 31.67 hours of recordings and 114021 frames were retained for analysis. Three types of annotation were employed. Objects, activity and privacy-infringing content labels. For object labels, a total number of 10427, i.e., 2.89 hours, frames were labelled as containing objects as bottle, plate, mug and glass, see Tab. I. We used the object tracker of DarkLabel [14] as tool to annotate dietary objects. For activity labels, a total number of 18607 frames, i.e., 5.15 hours, were labelled as containing activities as fluid consumption, food consumption, but also preparation of food, cleaning dishes and so on. Dietary events are not isolated events, and, commonly, other activities are carried out in the meantime, e.g., socialising or working at the desk. To assign practical meaning to activity segments, we considered gaps of less than five minutes as temporary interruption of an on going dietary event. The same heuristic was applied to interpret results of the spotting algorithm, see next section. For privacy-infringing content labels, a total number of 502 frames were labelled as containing other people’s faces or mobile and laptop screens.

<table>
<thead>
<tr>
<th></th>
<th>Bottle</th>
<th>Glass</th>
<th>Mug</th>
<th>Plate</th>
<th>Annotated Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1</td>
<td>550</td>
<td>497</td>
<td>631</td>
<td>2973</td>
<td>2793</td>
</tr>
<tr>
<td>Day 2</td>
<td>222</td>
<td>1</td>
<td>2</td>
<td>1290</td>
<td>1462</td>
</tr>
<tr>
<td>Day 3</td>
<td>949</td>
<td>1836</td>
<td>3</td>
<td>3393</td>
<td>2938</td>
</tr>
<tr>
<td>Day 4</td>
<td>81</td>
<td>1104</td>
<td>1069</td>
<td>2137</td>
<td>3234</td>
</tr>
<tr>
<td>Total</td>
<td>1802</td>
<td>3438</td>
<td>1723</td>
<td>9793</td>
<td>10427</td>
</tr>
</tbody>
</table>

TABLE I: Number of dietary object instances in dataset B.
technologic application and the negative impact of privacy concerns. The authors introduced the PSM as 2-by-2 matrix, see Fig. 3, to quantify the privacy-saliency trade-off. To compare our setup with the ones evaluated in [2], we modified the PSM. A quadrant of the matrix, labelled as Qx in Fig. 3 where x is the number of the quadrant, indicates the percentage of the total number of frames with specific characteristics, instead of absolute values. Quadrant 1 (Q1) indicates the percentage of the total number of frames that contains evidences of eating behaviour and exhibit no privacy concerns. Quadrant 2 (Q2) indicates the percentage of the total number of frames that contains evidences of eating behaviour and exhibit privacy concerns. Quadrant 3 (Q3) indicates the percentage of the total number of frames with no relevant dietary and no privacy-infringing information. Quadrant 4 (Q4) indicates the percentage of the total number of frames with no relevant eating behaviour information and privacy-infringing information. Value Q2.1 indicates the percentage of the total number of frames with relevant eating behaviour information that exhibit privacy concerns. Value Q4.3 indicates the percentage of the total number of frames with no relevant eating behaviour information that exhibit privacy concerns.

Fig. 3: Modified PSM providing a framework for studying the balance between privacy concerns and evidence of eating in video. Firstly introduced in [2], we modified the PSM to better compare different frameworks.

In order to evaluate the dietary event spotting, we performed a leave-one-out cross validation, considering each day as a fold. Performances were measured by exploiting the following evaluation metrics.

### B. Overlapping Score

An identified candidate match was compared with the ground-truth object bounding boxes on the same frame by using the Overlapping Score (OS), also called intersection over union. The OS is analytically described as:

$$ OS = \frac{|GT \cap \omega|}{|GT \cup \omega|} $$  

where $\omega$ is the window region detected by the object detector and GT is the window region define by the ground-truth labels.

### C. Precision and Recall

A True Positive (TP) was identified any time a GT-related OS was larger than a threshold $T_{OS}$. In our framework $T_{OS}$ is fixed at 0.5. We employed Precision (P) and Recall (R) as evaluation metrics. These metrics were derived as follows:

$$ P = \frac{\text{Recognised Objects}}{\text{Relevant Objects}}, \quad R = \frac{\text{Recognised Objects}}{\text{Retrieved Objects}}. $$  

Relevant objects was the number of objects that appears in the video sequence, retrieved objects represented the number of bounding boxes with OS > $T_{OS}$, while recognised objects was the number of bounding boxes with a GT-related OS > $T_{OS}$.

### D. Mean Average Precision

For the object detection evaluation, we relied on mean Average Precision (mAP). Firstly, Average Precision (AP) was computed to summarise the shape of the precision/recall curve. AP is defined as the mean precision at a set of eleven equally spaced recall levels $[0, 0.1, \ldots, 1]$:

$$ AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} p_{\text{interp}}(r) $$  

the interpolated precision $p_{\text{interp}}$ at a certain level $r$ is defined as the highest precision found for any recall level $r' \geq r$:

$$ p_{\text{interp}}(r) = \max_{r' \geq r} p(r') $$  

Ultimately, mAP was computed as weighted average of the APs from all classes. Analytically, mAP is expressed as:

$$ \text{mAP} = \frac{\sum_{i=1}^{C} N_i \cdot \text{AP}}{N_c} $$  

where $C$ is the number of classes, $N_i$ was the number of instances of class $i$ and $N_c$ was the overall number of instances for all classes. The use of a weighted average was justified by the highly unbalanced classification problem we had to cope with.

In order to penalise false positives and false negatives we computed the average F-measure as:

$$ \text{F-Measure} = 2 \times \frac{P \times R}{P + R}. $$

### VI. RESULTS

We made a indicative comparison between the privacy-preservation capability of our setup with the ones in [2], see Fig. 6, i.e., (B) head-mounted, (C) neck-mounted, (D) chest-mounted. The two dataset were different in number of participants, sampling frequency and total number of frames. Our dataset was collected by one participant, at 1 FPS, and it had a total number of 114021 frames. Thomaz dataset was collected by five participants, at 1/30 FPS, and it had a total number of 14422 frames. Our setup, in terms of privacy preserving, performed comparable to the best of their setups, i.e., the neck-mounted. In four days
of recordings, approximately 8 hours per day, less than 100 frames were recorded containing other people’s faces, and just 2 frames during dietary activity. During no eating activity, the main source of privacy infringing content was the screen of the mobile phone. Less than 400 frames were collected containing images of mobile or laptop screen. Concerning images from no eating context were less than 0.55% of the total amount of frames.

Precision, recall and F-measure of dietary event spotting are shown in Fig. 5. When the threshold was varied from 0.1 to 0.8, average precision increased from a minimum of 78% to 98%, average recall decreased from of 100% to 55%, and average f-measure decreases from of 90% to 64%.

Precision, recall and F-measure of dietary object detection and mAP are shown in Fig. 5. An overall mAP value of 51% reached. The individual object detection performances, by tuning threshold values, have a similar trend. One exception is the bottle class that reached the highest precision and exhibited the lowest recall.

VII. DISCUSSIONS

The impact of camera position and angle on the amount of privacy infringing content is evident from the modified privacy-saliency matrix evaluation. Nevertheless, our experiment can be considered just a feasibility study due to its small scale, i.e., one participant study. In order to validate our methodology, a larger number of participants, with different background and demographics, should be employed. Different participants’ habits could critically affect the privacy-saliency matrix evaluation. For instance, our participant, worked for most of the time at their laptop, implying a scarce usage of mobile phone. The lack of phone use needs could explain the small ration of frames in the fourth quadrant of matrix (A) in Fig. 4.

From the machine learning point of view, transfer learning proved to be crucial in order to correctly handle the scarcity of training instances due to the peculiarity of the image features captured from the chosen position. Further complicating the recognition process, high intra-class variance and a high inter-class similarity was observed. Dietary objects of the same class exist in different shapes, colours, dimensions, e.g, a 0.33l green pet bottle and a 1.5l transparent glass bottle. Moreover, dietary objects of different classes were similar in shapes, colours and dimensions, e.g, a cylindrical mug and a can. A high number of training instance should be provided in order to cover as many cases as possible or a taxonomy of labels with a higher resolution should be designed in order to cope with the high variability of object in a free-living setting. Furthermore, the correct detection of objects, when the user is performing manipulative tasks or in specific posture, e.g., laying down on the bed, can be prevented by occlusions.

VIII. CONCLUSIONS

The contribution of this paper is two-fold. We firstly demonstrated that the position and orientation of the camera constrain the amount of privacy-infringing content in egocentric videos. Specific position and angle of view, in fact, reduce the necessity to employ error-prone anonymisation techniques to limit privacy-concern, e.g., face detection, motion filtering, and image cropping, which are typically expensive and cumbersome to carry out. We also implemented a dietary event spotting algorithm, based on fine-
tuning of a DNN, in order to reduce the amount of workload for human operators, while performing analysis on large volume of data. Retrieval performance suggested that our approach is feasible.

ACKNOWLEDGEMENT
This work has been partially funded by the European Union H2020 MSCA ITN ACROSSING project (GA no. 616757).

REFERENCES