Future Generation Computer Systems 113 (2020) 343–362

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Supporting food choices in the Internet of People: Automatic detection of diet-related activities and display of real-time interventions via mixed reality headsets

Klaus Fuchs [a,*], Mirella Haldimann [b], Tobias Grundmann [a], Elgar Fleisch [a]

[a] Auto-ID Labs ETH/HSG, D-MTEC, ETH Zurich, Switzerland
[b] D ONE Solutions AG, Zurich, Switzerland

ARTICLE INFO

Article history:
Received 31 January 2020
Received in revised form 28 June 2020
Accepted 6 July 2020
Available online 15 July 2020

Keywords:
Internet of People
Diet-related interventions
Mixed reality
Consumer health
Personal Internet-based healthcare
Novel ubiquitous interaction mechanisms

ABSTRACT

With the emergence of the Internet of People (IoP) and its user-centric applications, novel solutions to the many issues facing today's societies are to be expected. These problems include unhealthy diets, with obesity and diet-related diseases reaching epidemic proportions. We argue that the proliferation of mixed reality (MR) headsets as next generation primary interfaces provides promising alternatives to contemporary digital solutions in the context of diet tracking and interventions. Concretely, we propose the use of MR headset-mounted cameras for computer vision (CV) based detection of diet-related activities and the consequential display of visual real-time interventions to support healthy food choices. We provide an integrative framework and results from a technical feasibility as well as an impact study conducted in a vending machine (VM) setting. We conclude that current neural networks already enable accurate food item detection in real-world environments. Moreover, our user study suggests that real-time interventions significantly improve beverage (reduction of sugar and energy intake) as well as food choices (reduction of saturated fat). We discuss the results, learnings, and limitations and provide an overview of further technology- and intervention-related avenues of research required by developing an MR-based user support system for healthy food choices.

1. Motivation

Today's food systems focus on efficiency and the production of inexpensive and foods high in calorific value. As a result, unhealthy foods with salt, sugars, saturated fats, and trans fats have become cheaper and ubiquitously available [1]. The global supply of unhealthier foods (e.g. red meat, sweetened drinks and processed foods) has surged [2]. Consequently, excessive dietary intake has become a recognized public health priority while supporting healthier food choices is becoming paramount to improve consumer behaviors and population health [2]. These consequences emerge from unhealthy food choices that are linked to both increasingly prevalent diet-related non-communicable diseases (e.g. obesity, diabetes and cardiovascular diseases) as well as preventable premature deaths [2]. Due to their increasing incidence and continuous excessive dietary intake, related diseases have become an alarming issue globally [3,4]: Today, over 1.9 billion adults are overweight, of which 650 million adults are affected by obesity with growth rates reaching epidemic proportions [5,6]. Because diet-related non-communicable diseases involve chronic conditions, they often require long-term, cost-intensive medical treatment. As a result, increased occurrence significantly and increasingly burdens health-care systems financially [3,7–11].

Unfortunately, contemporary countermeasures (involving food labeling, nutritional education, food counseling) have been unable to shift consumer behaviors. Most consumers still seem to struggle to identify healthy foods. Despite first positive results from front-of-package labels (FoPL) in particular [12,13] only few countries have successfully implemented FoPL on a significant scale. Examples include Australia (Health Star Rating, HSR) [14], France (Nutri-Score, NS), and England (Multiple Traffic Light, MTL) [13]. Approximately three out of four shoppers do not refer to such labels at all [15]. We argue that the lack of effectivity is owed to several barriers that thwart the diffusion of such food labels. Retailers, brands, and manufacturers tend to resist voluntarily introducing food labels as they fear not only negative economic impacts due to administrative and logistical efforts but also declining revenues [16]. Hence, most countries still adhere to conventional regulation, only requiring text-based declaration at the back of food items, i.e. back-of-package labels (BoPL). However, scientific evidence strongly suggests that consumers

[*] Corresponding author.
E-mail address: fuchsk@ethz.ch (K. Fuchs).

https://doi.org/10.1016/j.future.2020.07.014
0167-739X/© 2020 Published by Elsevier B.V.

and especially low-literate citizens make significantly healthier food choices when food labels are clearly visible [13,14]. Thus, most consumers are not yet benefitting from easy-to-compare FoPL when making food choices.

In addition to food labeling, nutritional counseling and education as part of multi-purpose programs have been proposed to induce the necessary shift in consumer behavior and to counter the underlying excessive energy intake [17,18]. However, due to limited financial resources, the majority of population cannot be included in current personnel-intense counseling programs, which additionally face the challenge of low acceptance rates [19]. To this end, a series of digital solutions, mostly in form of barcode scanning mobile Health (mHealth) applications, have been developed. Unfortunately, these apps often experience low adoption, short-lived retention, and self-selection by healthy users. While they are generally seen as an inclusive and scalable support system for healthy behavior, the efficacy of such apps to support healthy food choices under realistic conditions remains contested.

First, diet-related mHealth applications require manually logging every consumed meal or food item [20–22]. Second, such mobile apps are impractical, as users must actively remember to open them during a meal or during grocery shopping in order to retrieve relevant decision support [23]. Moreover, querying such information might require consumers to use both hands when operating the phone's application, which is usually impractical when eating or shopping [20,23]. Third, studies show that current diet-related mHealth apps are primarily retained by users with above-average interest in nutrition [24], a passion that is ordinarily not prevalent in sociodemographic segments prone to diet-related diseases [25,26]. Hence, diet-related mobile applications unsurprisingly suffer from low adoption [24], short retention rates [20], self-selection bias [24], and underreporting habits [20–22]. Clearly, consumers require more user-friendly, human-centric solutions.

The emergence of the Internet of People (IoP) and the expected adoption of mixed reality (MR) headsets such as Microsoft HoloLens or Magic Leap One and their successors could soon pave the way for novel and decidedly more human-centric and user-friendly food choice support systems. The IoP holds this promise because it promotes bridging machine intelligence with human intelligence to form novel user-centric applications. We suggest that the new technical capabilities of wearable headsets in combination with a user-centric solution design can help overcome the existing drawbacks of contemporary mobile diet applications. In the dietary context, computer vision-based interpretation of video streams from headset-mounted cameras can enable the automatic detection, identification [27–29], and quantification [30–32] of diet-related activity and food items without requiring any manual user input (e.g. keeping a food diary) [20,23]. Moreover, MR headsets also enable displaying three-dimensional visualizations of interventions because spatial computing allows positioning visualizations relative to the user's periphery and detected objects, thereby achieving high presence, salience, and immersion. Based on the detected context, such visual interventions can be tailored to the user's nutritional specific needs and integrated in real time into the field of view to support headset wearers in making healthy food choices [33,34]. These aspects act as important prerequisites for effective interventions [33,34]. Both approaches – (i) automatic context detection (e.g. eating a certain food or grocery shopping) and (ii) personalized real-time interventions – offer advantages over current mobile diet applications, as they can be passively triggered and follow the just-in-time adaptive intervention paradigm [35], considered best practice in ongoing health support.

Hence, this article describes what such an IOP arena based on computer vision (CV) and visual interventions in the context of nutrition and consumer behavior might look like in the future. First, we review the current research on CV-based food detection, identification, and quantification, as well as the latest research on MR-based interventions to nudge users toward healthier food choices. Second, as one of the main contributions of this article, we develop and illustrate a holistic framework for jointly applying automatic CV-based food identification and real-time visual interventions in a vending machine (VM) setting via MR headsets. Third, to assess such a system's potential of identifying packaged food products, we present findings from a technical feasibility study. Fourth, to assess the potential of visual interventions to alter user behavior, we conducted a user study with 61 participants under realistic conditions. Finally, we identify and discuss the gaps in current research and technology and thereby open up interesting avenues of research in personalized, nutrition-targeted interventions.

2. Related work

To support users in their food choices without necessitating active user input in the real world, future IOP systems can leverage the joint application of automatic CV-based identification of food items and real-time visual interventions via MR headsets, which are expected to be mass-adopted over the next years.

2.1. Computer vision-based detection of diet-related activity

As an alternative to less intrusive methods (e.g. keeping a food diary or scanning product barcodes), computer vision (CV) based detection leads to increased ease-of-use and consistent monitoring [20,21,36,37]. Together, the introduction of AlexNet [38], the consequential development of (deep) convolutional neural networks (CNN) [39,40], and the increased affordability and availability of more performant hardware have advanced object detection (OD) and image classification across many domains, leading to novel, advanced applications. For example, CV-based detection of road signs and traffic situations has enabled the development of autonomous cars while real-time translation of image-encoded words has eased the life of travelers [41,42]. CV-based food detection [20] enables wearable MR headsets to detect diet-related user activity, identify packaged food products [34], and quantify nutritional properties of cooked meals [29] without user input. CV thereby overcomes the most relevant drawbacks of current mobile diet applications, which require logging food manually, actively taking a picture, or scanning a barcode. Specifically, wearable headsets can rely on built-in cameras that constantly produce an interpretable video feed [23]. This allows the wearer to remain hands-free, which represents a more convenient method for capturing a person's dietary contexts (at home, at restaurants, in supermarkets, or on the go). Most importantly, CV-based automatic food detection does not rely on factors otherwise determining logging: salience, memory, (bad) conscience or feelings in general, involvement or interest in nutrition.

In order to detect food items from a headset's video feed, different approaches of varying complexity and combining object detection and consequential image classification, have been suggested and validated. MobileNet [43,44] is a CNN aimed to support implementations on mobile devices with limited computational capabilities. ResNet [45,46] (presented at ILSVRC2015 by Kaiming He and his colleagues) features heavy batch normalization and 152 layers, making it computationally more intense than MobileNet. Finally, Inception v4 [47] outperformed ResNet, albeit once again featuring higher complexity.

Recent advances in deep learning and representation learning have created many such feature extractors [48]. These are generic, multipurpose, and applicable to a range of objects, including cooked meals and packaged products [49]. Neural network architectures and their hyperparameters are context-, task-, and design-dependent and hence not necessarily directly transferable to every new context or dataset without adaptation and sequential testing on realistic context-specific data. With enough training data and a suitable architecture choice, CNNs can be taught to extract features on multiple levels and may exceed human performance in some applications (e.g. face recognition) [50]. Until now, scholars have built dedicated CV pipelines, either for food items (e.g. meals) [27,29,51,52] or for packaged products in a retail environment [53–56]. Similarly, the publication of image datasets, which are central to developing CV solutions, follows this separation of meals and packaged products. For example, FoodNet101 contains 101'000 labeled images of composed dishes [27], while SKU110 contains 1.74 million images of packaged retail products [57]. To this end, a holistic solution aimed at detecting composed meals as well as grocery products needs to combine the capabilities of both streams to enable meaningfully detecting diet-related activity.

Object detection (OD) precedes image classification as it deals with identifying an area within an image as a potential candidate for classification. Recent publications on detecting retail products suggest a correct mapping of 74% mean average precision (mAP) for large supermarket datasets, over 77% mAP for smaller product datasets [55], even reaching over 93% mAP when over 60 images are available per product [58]. Similarly, meal detection in real-world images has been shown to be feasible [59,60]. In terms of meal image classification (IC), current research suggests accuracy rates of over 72% [27] for detecting category affiliation in the FoodNet101 dataset. For packaged food products, recent accuracy rates range between 48% to 69% for large supermarket-based product datasets [54], and over 95% for smaller datasets [58]. Given feature variety, as well as the vast number of meal classes and retail products, these accuracy rates are already acceptable for real-world applications. For example, CV-based meal detection and identification is already evident in mobile diet applications (e.g. Lifesum, Snaq, Bite.ai), where users can take a picture of a meal instead of searching for it via text. Similarly, retail stores are using autonomous robots that leverage CV for inventory stock keeping. Image classification accuracy and OD are expected to grow in the future with larger publicly available image datasets and hardware improvements (e.g. camera resolution).

Finally, to conclude the CV-based detection of diet-related activity, the identified food item's nutritional properties and quantity [30–32] can be retrieved. To this end, an increasing number of semantically labeled datasets are available [61], which were shown to yield reliable calorie and nutrient estimates for a meal [29,51,59,60]. Similarly, to interpret the detected activity and to nudge users toward healthy food choices during a supermarket visit, the nutritional composition data of grocery products are becoming increasingly available via open databases [62] and retrievable after prior CV identification [34].

2.2. Display of real-time interventions in mixed reality

The increasing possibility of identifying one or multiple food items within single frames in video feeds enables a user support system based on MR headsets to display corresponding visual cues to nudge users toward healthier food choices. Given sufficient accuracy (i.e. high mean average precision (mAP)), retrieving the dimensions of the detected food items becomes feasible. Further, relative user and item positioning can be approximated via spatial computing [63]. Meeting these preconditions enables displaying item-related information. For example, El Sayed and colleagues [64] and Microsoft [65] demonstrated navigating a headset wearer toward the healthiest product on a shelf while the user remained hands-free. Other potential interventions might include support in selecting healthy items from a restaurant buffet or praising a user when eating a salad and thereby form healthier habits.

Counterintuitively, given the technological feasibility of identifying food items from video feeds automatically, it seems surprising that existing research on MR-mediated food choice interventions has so far remained rather nascent. Studies on MR-mediated interventions for supporting consumers in selecting healthy food items have not advanced beyond intervention design. These studies have mainly demonstrated early-stage prototypes or are field studies involving smartphones rather than headsets. For example, smartphone-mediated MR applications have been designed to leverage CV to support consumers in identifying vegetables [66], to estimate portion sizes of composed dishes [31,67], and to help users navigate supermarkets and discover healthy food items [68]. Smartphone-based MR applications were found to be easy-to-use [68], to alter consumer behavior [69,70], and to positively improve food choices [70]. Still, outcome effects remain contested and shortcomings associated with manual logging persist.

In contrast to smartphone-mediated MR interventions, headset-mediated MR interventions allow users to remain hands-free. For example, a Google glass-based intervention study demonstrated the feasibility of automatically detecting vegetables and fruits through CV, in turn enabling food monitoring and interventions [71]. El Sayed et al. demonstrated a variance of MR visualizations aimed at improving user performance on search, selection, and ranking tasks on supermarket shelves [33]. Similarly to smartphone-based applications, wearable cameras were also shown to effectively monitor the consumption of composed dishes using CV [72]. Microsoft even patented a wearable headset able to deliver MR interventions for eating activities [65]. But although MR headset-mediated interventions on food selection have been shown to be feasible, little is known about their impact on user choice and other outcomes in the real world, nor about users' opinions on such systems and their efficacy.

Hence, we address the existing research gap of the joint application of CV-based detection of food items and just-in-time visual interventions to improve food choices. To this end, we present a novel conceptual, integrative framework that combines both research streams into a novel user support system for making healthy food choices. Further, we present findings from one of the first in-the-wild implementations and validations of MR-mediated purchase interventions aimed at improving food choices. Specifically, we applied a MR (MR) wearable headset-mediated intervention ($N = 61$) at vending machines (VMs) to explore the technical feasibility and potential impact of passively activated, pervasive MR food labels in affecting beverage and food purchasing choices. This article extends our previous publications on beverage choices [34] and technical feasibility [58] through multiple additional assessments. First, we expanded the assessment of beverage choices to include food choices. What follows is therefore one of the first randomized and controlled real-world intervention studies on food selection using MR interventions. We assess whether visual cues in form of front-of-package labels (i.e. Nutri-Score) influence consumers in preferring and selecting healthy or unhealthy beverages and foods. Second, we analyze consumers with low food literacy, a sociodemographic segment that is especially at risk for diet-related diseases and unlikely to enroll in traditional diet-related interventions. Third, we include an in-depth discussion on the latency of product detection via CV to assess the technical feasibility of detecting packaged products under realistic circumstances.

3. Integrative framework and implementation

To combine the advantages of automatic detection of diet-related activity and passively triggered, just-in-time interventions, we propose a novel integrative framework based on jointly applying MR headset-mediated CV and visual interventions (Fig. 2). Counterintuitively, despite its promising potential of automatic tracking and passively triggered real-time interventions, the proposed combination of food item detection and holographic interventions displayed in a wearer's MR headset represents a novelty that so far has received little attention in the relevant literature. To close this gap, we show how such an integrative framework might be created and introduce its necessary functional elements (Fig. 2). In addition, we have implemented the most important subsystems of the proposed framework in two validation studies (discussed below in the respective sections). First, by implementing a CV-based system to detect packaged food products, we corroborate the current technical capabilities of today's neural networks and present our findings, based on a conducted a technical feasibility study (Fig. 4). Second, we implemented a MR-mediated intervention system and conducted an impact study (Fig. 7) through an empirical field study. Hence, rather than considering an actual implementation of the overall framework, this article discusses the implementation and validation of its most important, yet hitherto underresearched subsystems: (i) CV-based detection of packaged products and (ii) MR mediated real-time interventions.

3.1. Mixed reality user support system for healthy food choices

In the following, we introduce the proposed integrative framework in form of a holistic user support system based on MR headsets (Fig. 2). We discuss several preconditions that are required for the proposed system to become feasible. We argue why we believe that these preconditions will highly likely be met in the coming years. Currently, our proposed system represents a vision for the IoP rather than an actual prediction with a foreseeable timeline. It might only be a matter of time until this vision becomes a reality, especially because relevant tech companies including Apple, Facebook, Microsoft, and Magic Leap have all recently announced that they will be introducing consumer-oriented MR headsets in the next few years. Whether the adoption of such headsets will occur as speedily as the adoption of smartphones or might take longer such as the adoption of VR headsets remains to be seen.

First, we assume that next-generation MR headsets (e.g. Microsoft HoloLens or Magic Leap One) and their successors will become increasingly available and will be adopted as primary interfaces to the Internet. We also assume that such next-generation MR headsets will become smaller in size, lighter in weight, more durable in terms of battery capacity, and popular to use. These improvements will increase the likelihood of users wearing these devices more or less permanently during the day. We thus expect that future generation MR headsets will be used similarly to wearable smartwatches (e.g., Apple Watch). As these devices aid users in improving their physical activities through tracking and notifications, we propose that MR headsets can support users in maintaining healthier diets.

Second, we assume that MR headset features (e.g. frontal and sideward facing cameras), will be permanently switched on and will constantly scan the environment for objects and gestures. Scanning will also enable detecting food-related activities. Through gesture detection, MR headsets allow for very natural human-to-device and human-to-object interactions, as spatial computing and holographic projections enable users to use their eyes and gestures to interact with the device as well as with applications, rather than via human-made artifacts (e.g., mouse, keyboards, or two-dimensional screens). Such input gestures are likely to include "air taps" or "blooms" (i.e. hand gestures to interact with HoloLens), which are among the principal ways of interacting with the device. On a more abstract level, drinking or eating (e.g. with knife and fork, chopsticks, hands) requires using one's hands and can therefore be considered to be a "gesture" that is detectable by the headset's cameras. Hence, these headsets have the technical capability to allow for continuous monitoring and consequential detection of diet-related activities (e.g. eating a meal or selecting groceries from a shelf aside from food items).

Third, we hypothesize that real-time CV-based detection of meals and packaged products will become available on a global scale. The previous section has shown that, although they not yet perfect, current accuracy rates for detecting and identifying food products and meals are promising. We believe that offering and maintaining CV models for interpreting meals and food products remains nontrivial. Deep learning models require large amounts of training images and have to manage a plethora of food item classes, which, moreover, are prone to high feature variety. This is the case in particular because many packaged products have a similar visual appearance and because thousands of newly packaged food items are constantly added to the market. These aspects require constantly adding newly labeled image data and consequential retraining of neural networks. But similar to the proliferation and successful adoption of other CV applications, we believe it is only a matter of time until improvements in detecting and identifying of food items from video feeds become evident. Therefore, we argue that such deep learning models will be offered by dedicated service providers, similar to today's offerings, which already enable identifying meals from photographs (e.g. Bite.ai, Snaq). We also argue that the state of the art will further improve in the near future due to the increased availability of labeled image data and improved deep learning models.

Fourth, we expect that wearable headset-based support systems offering dietary coaching will be increasingly adopted. Especially since the vast majority of consumers report being interested in nutrition, yet drop out of manual logging mHealth applications [25], such an automatic and convenient system could become popular. But even in the unexpected case, in which such wearable, headset-mediated diet coaching applications will not be mass-adopted, such devices could prove very useful in supervised counseling sessions (e.g. for diabetic or obese patients). For example, dietary monitoring is still considered a cornerstone of diabetes treatment. It is, however, widely observed that especially obese individuals underreport food intake [73]. Wearable headset-based support systems would provide less obtrusive and unbiased monitoring.

Finally, we argue that MR headsets will become capable of streaming video feeds to servers for cloud- and CV-based interpretation and corresponding detection of food items. While, in theory, devices could detect diet activity and identify meals and groceries with on-device models, we believe that cloud-based models have advantages in terms of updates, runtime, and inclusion of new products. An alternative to both approaches might be edge-computing. While potentially offering advantages in terms of privacy and latency, this still requires an online connection. Similar to gesture detection, detection of grocery shopping or eating could also be run on-device, while actual food identification could in turn rely on server-side models. We argue that especially with the upcoming roll-out of 5G telecommunication networks globally, the necessary bandwidth for uploading video streams to server-side CV interfaces will be made available at scale in the near future.

If these preconditions are met, we believe that the proposed framework in form of an MR headset mediated user support systems (Fig. 2) will become technically feasible. Further, we believe that its likely adoption as a next-generation user device will enable the MR headset to finally bring diet-related interventions to a new, unprecedented high quality and ease-of-use outperforming current mobile application-based approaches. We hold that the MR headset will achieve this goal in particular through its camera-based capturing of its environment and its ability to integrate visual cues into the field of view. As such, this system will help users to make healthier food choices through (i) the automatic detection of diet-related activity (Fig. 4) and (ii) real-time visual interventions (Fig. 7). Below, we describe the process through which the proposed system can support users in real-world scenarios.

As described in the introduction, consumers make many crucial diet-related decisions every day, for example, during consumption of meals or beverages and during grocery shopping. We suggest that our proposed MR headset-based support system can help to keep track, interpret, and improve a user's behavior. First, we assume that users are wearing their MR headset with the support system application installed and with the device activated and connected to the Web. Once users engage in a diet-related activity (e.g. selecting food products in a supermarket), the system will detect this circumstance by interpreting the video feed from the headset's cameras. In an alternative scenario, users could also wear their headset while eating in a restaurant or drinking a soda.

Once context detection confirms that a diet-related activity is taking place, the system sends video feeds to a server-side CV interface that applies object detection and image classification to assess the food items present in the current scene. To achieve this, neural networks (NNs) need to be pretrained with image databases to be able to identify meals and products. Product detection consists of two steps: (i) object detection and (ii) consequential product classification. First, the NN estimates where an object of interest might be located, and only then predicts which product is involved. This approach is well established in the related literature on packaged product identification [54,58]. In addition, detection can be improved in terms of latency and accuracy by including additional contextual information (e.g. user location) to reduce the search universe of potentially present food items. If, for example, a given restaurant is known, the potential food items are likely to be products of that restaurant, for which meal images might be available. Or, if the specific supermarket where a user is purchasing groceries, is known, the number of potentially present classes can be significantly reduced. It remains to be seen whether such augmentation with context-based metadata will generate purpose-specific knowledge graphs or simply deeper NNs (also pretrained with potentially present location data). In either case, the CV model will reveal a list of candidates and confidence values for identified food items. In case such confidence values only yield low confidence, image pooling or a recurrent NN architecture could enable repeated estimations multiple times per second in a video feed, leading to more reliable accuracy rates.

Given the successful detection, identification, and quantification of present food items, food composition databases can be queried to retrieve an item's corresponding nutritional attributes. Such properties refer to the estimated amounts of relevant nutrients per 100 grams of the food item and can include values for calories, sugar, carbohydrate, fat, saturated fat, protein, salt/sodium, dietary fiber, diverse minerals and vitamins — all potentially relevant for evaluating the quality of a food choice. In addition, food composition databases could also include further attributes such as present allergens, category affiliation, or food labels. Finally, the server-side CV would return the positions, nutritional properties, and additional attributes of one or multiple identified food items back to the system.

Next, having identified the current context, the system can compare it to a user's recent diet-related activity, preferences, and current health state, in order to interpret the impact of the currently visible food item and in turn to design the corresponding real-time intervention to be shown to the user. To this end, the nutritional properties of the identified food items are interpreted to assess their nutritional quality. Food labels such as Nutri-Score enable evaluating the amounts of present nutrients and consequentially ranking a food item's healthiness on a scale from −15 (very healthy) to 40 (very unhealthy) [74]. Likewise, composed foods could automatically be assessed by their estimated nutritional composition. Similar to dieticians, the support system must interpret a food item compared to a user's recent activity. For example, after physical exercise, an intervention for the same detected food item might look differently than after a sedentary day at the office. Similarly, recent consumption of previously detected food items could impact intervention design. For example, if a user has already eaten a chocolate bar, the impact of consuming a second one might be interpreted differently than the first one. Therefore, concepts such as the Healthy Eating Index (HEI) [75] can support users in interpreting food items based on previously consumed meals. Because users have varying preferences for certain foods, the intervention should also consider a user's dietary patterns and taste. For example, recommending a meat-based meal would entail very different user experiences for vegetarian users. Similarly, taste could also be important in designing interventions. While some users prefer spicy foods, others might have a "sweet tooth" and hence react differently to identical interventions. In this regard, current research on food recommendation systems [61] could help to research and design supportive and effective nudges.

Beyond preference, tailoring interventions is also important for health reasons, as different users (in terms of gender, age, body-mass-index, and health state) have varying recommended daily intake rates [76]. In addition, food allergies should be respected when designing interventions [68], for example, through warning signals. Icon-based notifications may prove superior over text-based notifications [77]. Finally, visual interventions need to be designed. Whether these will follow rather concrete instructions or complex three-dimensional and/or gamification-inspired animations remains to be seen.

We assume that MR support systems will keep track of users' food consumption and be able to identify their most promising improvement potentials as well as their healthiest, already established habits. Intervention design should endeavor to strengthen healthy habits (e.g. by praising a user for successful streaks) and to exploit improvement potential (e.g. by suggesting alternative food items). For example, improvement potentials might be a "reduction in sugar within yogurts" for a certain user and an "uptake in vegetables" for another user. Similar to the Apple Watch-mediated interventions on physical activity, an MR support system could also leverage behavior change theory to help users to maintain healthy habits, for example, by motivating them (e.g. "keep eating three portions of vegetables per day") or by applauding them (e.g. "you achieved your sufficient daily fruit intake level for today").

Based on the joint application of (i) the CV-based detection of diet-related activity and (ii) the display of real-time interventions in MR, the proposed next-generation MR-based user support system promises to automatically track and efficiently improve food choices over time. To validate the system's potential, we therefore conducted two validation studies for both of its subsystems: a technical feasibility study of CV-based identification of packaged products (Fig. 4) and an in-the-wild user study of real-time visual interventions in MR headsets (Fig. 7).

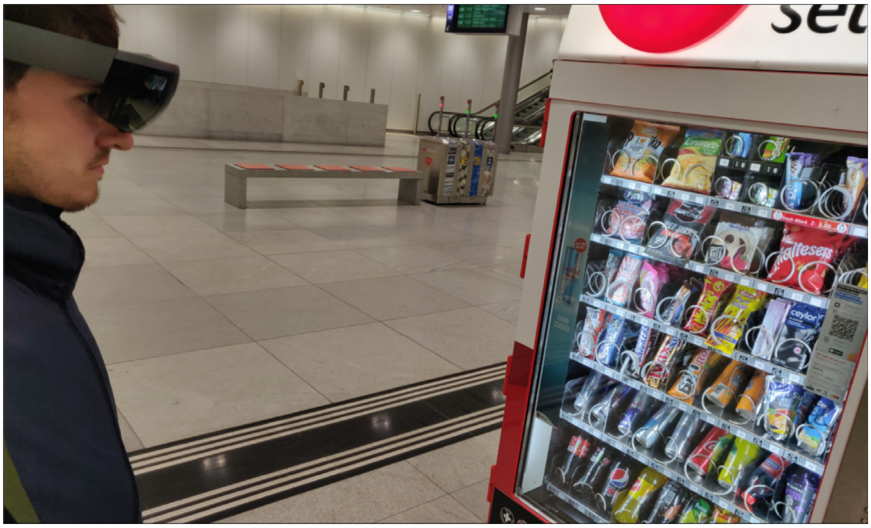
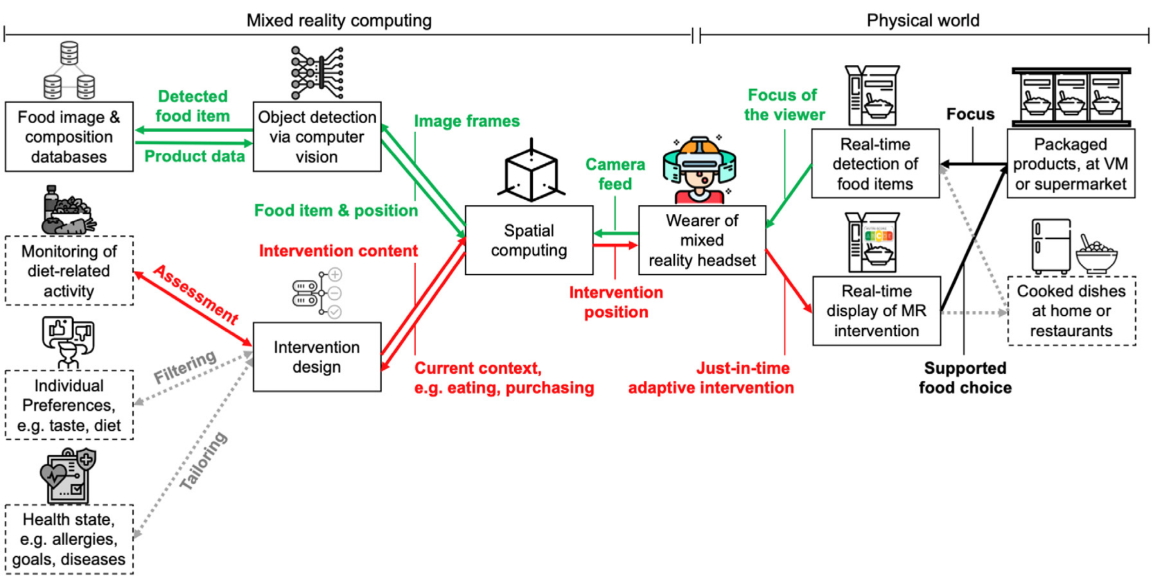


Fig. 1. Mixed reality headset wearer about to make a food choice in front of a vending machine.


Mixed reality computing
Physical world
Food image & composition databases
Detected food item
Product data
Object detection via computer vision
Image frames
Food item & position
Intervention content
Spatial computing
Camera feed
Intervention position
Wearer of mixed reality headset
Focus of the viewer
Real-time detection of food items
Focus
Packaged products, at VM or supermarket
Real-time display of MR intervention
Cooked dishes at home or restaurants
Supported food choice
Just-in-time adaptive intervention
Current context, e.g. eating, purchasing
Monitoring of diet-related activity
Assessment
Intervention design
Filtering
Individual Preferences, e.g. taste, diet
Tailoring
Health state, e.g. allergies, goals, diseases

Framework components	Reference in manuscript	Validation	Label
Computer vision-based detection of diet-related activity	Chapter 4: RQ1 – RQ3	Yes	→
Display of real-time interventions in mixed reality	Chpater 5: RQ4 – RQ7	Yes	→
Selection of packaged products in retail	Chapter 6: Dependent variable	Yes	→
Selection of cooked dishes outside of retail	Chapter 3: Integrative Framework	Out of scope	⇢
Intervention design based on diet tracking, individual preferences, health state			

Fig. 2. Integrative framework of the joint application of (i) computer vision-based detection of diet-related activity (green) and (ii) display of real-time interventions (red) in mixed reality headsets . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Implementation of validation studies

Below, we introduce the trial setup, which consists of a vending machine (VM), packaged products, wearable headset, and the study application. We decided to trial the MR user support system at a VM setup for three reasons. First, VMs contain a limited number of products that can be photographed and labeled to train a deep learning CV pipeline. Also, VMs display products without occlusion and with an active light, which aids detection and identification. Importantly, VMs offer an ideal, grid-like layout that can be used as a digital anchor to map interventions both relative to the location of the products and the machine. Second, VMs offer predominantly unhealthy food items, especially ones rich in sugar and saturated fats. As such, we expected the VM in our setup to be used by users who more frequently consume processed foods than by health-aware consumers. Third, VM product ranges remain relatively static, enabling the creation of a comprehensive, up-to-date database of ingredient data. Furthermore, it was financially feasible to purchase all the products within the machine to collect multiple pictures of the items both inside and outside the VM. We chose a Selecta VM (Fig. 1) to ensure representativeness. Selecta is the European market leader with 125’000 machines worldwide and caters to five million consumers every day. We conducted the user trial at VMs located at Zurich main railway station in Switzerland. Most Selecta machines are similarly or even equally assorted. Hence, choosing

Table 1
Products available in the VM (N = 43).

Snacks	Mean (SD)	Beverages	Mean (SD)
Weight (g)	58.2 (26.8)	Weight (ml)	400 (118.3)
Price (CHF)	2.77 (0.63)	Price (CHF)	3.18 (0.61)
Snacks by NS	Count (%)	Beverages by NS	Count (%)
A (Healthy)	0 (0%)	A (Healthy)	4 (20%)
B	3 (13%)	B	4 (20%)
C	6 (26%)	C	4 (20%)
D	7 (30.5%)	D	4 (20%)
E (Unhealthy)	7 (30.5%)	E (Unhealthy)	4 (20%)

these VMs increased the reproducibility and generalizability of this study since its impact goes beyond the few machines used here and could potentially be reproduced and applied in a similar way across VMs internationally.

For the study design, we consulted dietary experts from the Swiss Society for Nutrition (SGE-SSN). Together with these experts, we decided to focus our MR headset-mediated intervention on beverages and snacks separately. While both high-calorific snacks and sugar-sweetened beverages have been shown to play a major role in the increased prevalence of diet-related diseases [78,79], they are nonideal substitutes for one another. This relationship is intuitive, as a hungry consumer is unlikely to choose a mineral water, regardless of a well-designed intervention. To interpret the different food items, we decided to draw on the Nutri-Score (NS) framework [74] for three reasons. First, the NS food label can be converted into visual cues that can be displayed in the user's field of view. Second, growing evidence exists that this form of FoPL correlates with healthier food choices in purchasing environments [80]. Third, the NS framework includes both a food- and a beverage-specific rating for nutritional quality. In order to realize the Nutri-Score in its original, intended form, not only a product's nutrients, but also its relative share of fruit, vegetable, and nuts requires accounting for, as the Nutri-Score credits such ingredients with a bonus on the score. Regarding content (see product characteristics and NS ratings in Table 1), the Selecta machine offers different assortments for both categories. Distribution of beverage products is uniform, ranging from healthy (e.g. mineral water) to unhealthy items, and thereby offering a variety of healthy substitutes for consumers to choose from. For snacks, the range is skewed toward unhealthy items with fewer healthy alternatives to choose from. Available types of beverages are equally distributed regarding nutritional quality (from A to E according to Nutri-Score; Table 1). Beverages in the Selecta VM include mineral water (still and sparkling), juices, energy drinks, energy-reduced and sugared soft drinks. On average, bottles contain 400 ml and cost CHF 3.18 (USD 3.30). Standard deviation is rather small, indicating a more or less identical price across different types of beverages. Available snacks include chewing gums, cakes, donuts, chips, chocolate bars, waffles, and beef jerky. On average, they cost 2.77 CHF (USD 2.87) and weigh 58.2 grams on average.

To conduct our studies (technical feasibility and user study), we purchased all products available in the VM and manually entered their properties including nutritional composition into a database hosted on a dedicated server. In total, the number of products (snacks and beverages) available over the course of the study in a VM was 109. However, Selecta's cyclical changes to the product assortment reduced the final product universe in the user study to 43 products from which users could choose (Table 1).

4. Computer vision-based detection of diet-related activity

Based on one of our previous studies, we now provide insights into the technical feasibility of current CV models supporting the correct detection and identification of food items using MR headsets. We compared different NN architectures and their corresponding accuracy rates for image classification and OD, respectively [58]. We decided to create one large labeled image dataset and to evaluate classification and detection tasks separately. To assure realistic conditions in terms of resolution and quality (Fig. 2), the pictures used for training the CV models were recorded using Microsoft HoloLens or comparable mobile devices. To train the NNs and consequential inference, we used Google Cloud with P100 GPU instances and TPU v2 instances.

We decided to evaluate how many labeled image instances of a product are required to achieve suitable performance to support the detection and identification of food items. We chose this approach due to the still relatively limited availability of publicly accessible labeled training data for packaged products, and also because labeled product images are potentially expensive to acquire or generate for the millions of existing products. In this context, we also released this study's labeled dataset containing 295 images of VMs assortments with 10'035 labeled instances (5646 beverages, 4389 snacks) to stimulate research on packaged retail products [81]. To the best of our knowledge, our dataset (N = 10'035 labeled product image instances) represents the largest publicly available dataset in this domain that contains product identifiers (GTINs) and therefore allows integration of nutrient data. As collecting such labeled image data requires significant time and effort, we assume that this situation will improve over time and that this paper represents an important and first milestone, by moving from synthesized lab data [54] to real-world product detection [58].

We compared three current models for (i) image classification and (ii) OD. Using (i) Inception ResNet V2 [47], (ii) ResNet50 V2 [45], and (iii) MobileNet V2 [44] as classification networks, we used a subset of popular CNNs architectures available for image classification of differing complexity. The corresponding, implemented networks (ODNs) were: (i) Inception ResNet V2 [47] for classification, with Faster RCNN [82] for OD; (ii) ResNet50 V2 [45] for classification, with SSD and Focal Pyramid Networks (RetinaNet) [83] for OD; and (iii) MobileNet V2 [44] for classification, with SSD [84] for OD. We trained all NNs by fine-tuning from existing checkpoints. For image classification, those checkpoints stemmed from Tensorflow hub and were pretrained with the ImageNet 2012 dataset [85], while those for OD originated from the Tensorflow API [86], pretrained with the COCO 2014 dataset [87].

Image classification operations are the backbone of OD operations as they classify an object that was previously found to be potentially interesting within an image. Thus, successful OD requires reliable image classification performance. Hence, we raised two research questions (RQs). **RQ1** Can current NNs yield sufficiently high accuracy for image-based product classification in a realistic retail environment? Based on the results of the image classification tasks, we addressed the same objective for OD, which includes image classification as a subsequent task. This leads to **RQ2:** Can current NNs in combination with object detection networks (ODNs) yield sufficiently high enough mean average precision (mAP) for image-based product detection?

Next, we manually labeled products via defining image patches from pictures taken from the VM (Fig. 3). Second, to ensure consistency across the multiple subsets of varying $k \in [0,100]$ images, we chose a subset of N = 39 products from the VM, for which the total labeled dataset included at least 100 (training) + 20 (test) instances per product class. This step allowed us to evaluate the NN architectures for any k smaller or equal to 100 for the N = 39 products. For product classification, we excluded the 20 instances per class as a holdout dataset. Similar to product detection, the holdout dataset including randomly sampled 20% of the images guaranteed that there were at least 100 instances of every product in the training set. Thus, the test set for OD was imbalanced in a number of classes, yet included at least 20 instances

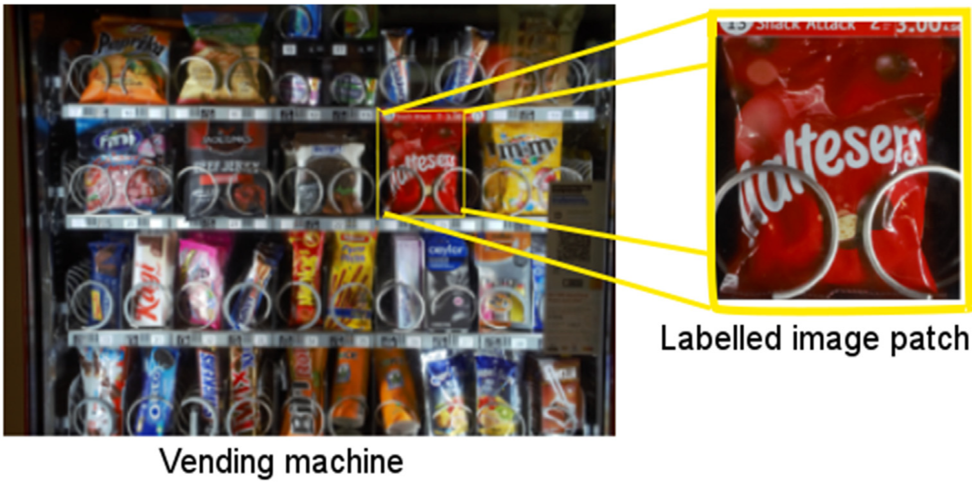
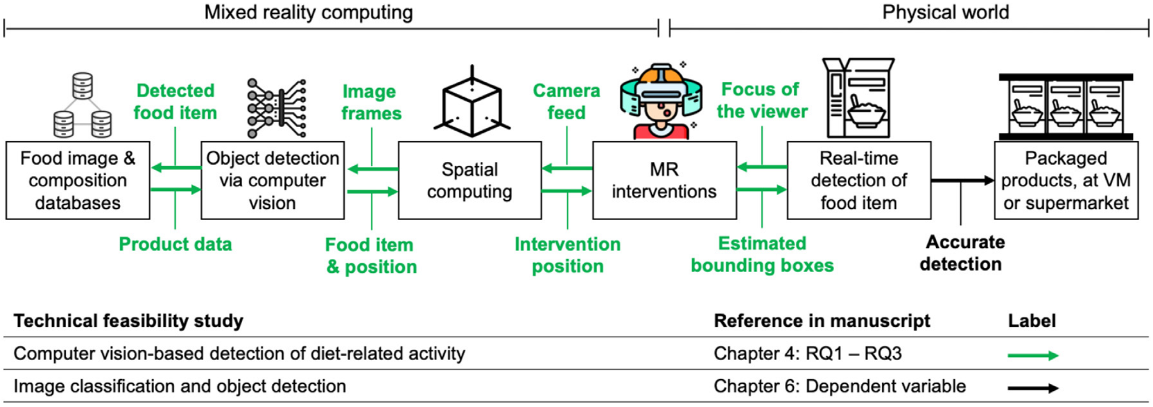


Fig. 3. Vending machine with a labeled, rectangular image patch (e.g., Maltesers snack 100 g). (Product classes: 109 in total, of which N = 43 were available during the user study and N = 39 were captured with over 100 labeled image patches).



Technical feasibility study	Reference in manuscript	Label
Computer vision-based detection of diet-related activity	Chapter 4: RQ1 – RQ3	→
Image classification and object detection	Chapter 6: Dependent variable	→

Fig. 4. Computer vision-based detection of diet-related activity using MR headsets.

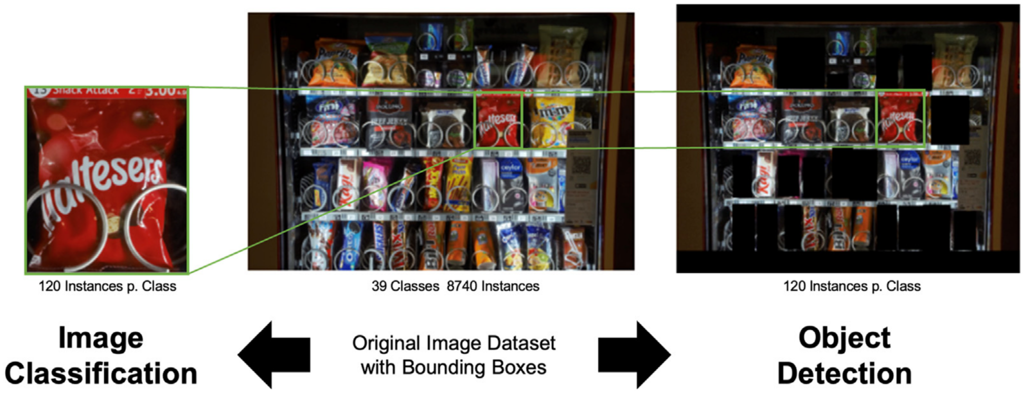


Fig. 5. Generation of training data with *k* images per product class for training the neural networks in the image classification and object detection tasks.

of every product. For each *k* smaller than 100, subsamples of the entire training dataset were chosen. For image classification, images not needed were simply excluded from the dataset. For the OD task, the parallel occurrences of certain product instances within VM made this approach impossible (e.g. RedBull is present four times in every Selecta VM). We therefore randomly chose instances in the training set to algorithmically cover surplus image patches with black overlays until every labeled product existed exactly *k* times across all training images (Fig. 5). Training images without labeled instances (i.e. fully blacked out) were excluded. For every *k*, the training dataset looked slightly different, since different images were blacked out. However, the holdout set remained constant for all datasets.

We considered it important to highlight the aspect of "fast-paced" in our research questions, especially because our technical setup aimed to address the realistic conditions experienced by consumers in a retail environment. These include that consumers can expect instant results (i.e. within one second), as they potentially encounter thousands of products on their journey through a retail store or a train station. The duration of one second proved well-suited to multiple user tests, where we equipped users with a HoloLens device in front of the VM. As different NN architectures feature varying complexity levels and accuracy rates, the trade-off between latency, accuracy, the number of available training images, and the computational environment (i.e. mobile device or cloud backend) became another focus of our study and led to **RQ3**: Can object detection run at a real-time latency with a frequency rate of at least multiple frames per second on user devices?

If detection over multiple frames is an option, this may theoretically increase accuracy by image pooling the confidence and object predictions over multiple frames. Frame pooling is possible since the headset continuously uses simultaneous localization and mapping (SLAM) algorithms to map headset surroundings. This map enables inferring the position of the headset and the camera parameters from the 3D position of a pixel on the image via ray tracing. This, in turn, allows for capturing multiple instances of a food item, yet to be detected. Next, the average confidence for all frames per class and the average position of all overlapping boxes can be taken and the item class with the highest average confidence can be selected as the prediction. This provides the NN with multiple possibilities for finding the object, assuming that NN accuracy is so high that it will on average predict the correct object more often than the incorrect one. Such image pooling allows for recurrent NNs in which the previous image instances of an item are refed to the next prediction. Therefore, fast inference speed may boost accuracy through image pooling. If accurate image classification (RQ1) and object detection (RQ2) are possible at fast inference speed (RQ3), then the technical feasibility of the proposed detection and interpretation of diet-related activity through MR headsets can be confirmed.

5. Display of real-time interventions in mixed reality

Next, we assessed the potential impact of the proposed MR headset-based support system on actual user behavior, i.e. beverage and snack choices from the vending machine. As such, this section extends our preliminary results [34]. Also, it includes (1) the addition of food items to the analysis, and thus extends previous analysis of beverages [34], (2) an additional analysis of food choice behavior in users with low food literacy, and (3) in-depth discussion of the intention-to-use of MR headset wearers. Given technical feasibility, the efficacy of the visual intervention is of course a further requirement that a support system must meet to change behavior.

To conduct our in-the-wild field user study (at Zurich main railway station), we needed to ensure fast product detection relative to the headset wearer's position, in order to prevent any selection bias among participants arising from intervention appearance.

When evaluating whether to use a hard-coded or CV-based layout in our impact study, we compared both methods and ultimately opted to use a hard-coded layout to validate the potential impact of the MR intervention. We implemented and tried the OD and image classification using NNs and image pooling, and indeed observed successful and correct detection for most products within a few seconds under realistic conditions. Unfortunately, reproducible, quantitative assessments of product recognition rates under real conditions are impossible due to the many moderating factors impacting inference frequency. More specifically, several factors, among others, varying product orientation (i.e. non front-facing products), lighting conditions, angle and distance to the VM, configured minimum confidence thresholds to display an intervention, and Internet connection speed, all heavily impact the time needed to detect products via the headset. In our development phase, we used Microsoft HoloLens to stream image frames to a Google Cloud Platform GPU to assess the feasibility of using the OD pipeline presented in this paper to identify products in the VM. We had to use a server-side NN because HoloLens (version 1) was unable to run TensorFlow on-device. Fifteen colleagues at our lab tried the headset and self-reported (through qualitative feedback) that they observed detection rates of approximately 70% of all products within approximately three seconds in the VM, when in WiFi, under non-ideal lighting conditions and with high minimum confidence thresholds to minimize the number of false positive detections. But since HoloLens did not support fast websocket transfers, we also compared it to a OnePlus 6T Android device, where detection was much faster with approximately 85% of products detected within two seconds. Especially the discussed edge cases in [58] hindered a perfect detection of all products (e.g. backside-rotated products, mineral waters look very similar). Due to the many moderating factors involved, an exact quantification of such CV-based detection rates in the wild is impractical and was not conducted by us.

To ensure rigorous analysis of the potential impact of the visual intervention on product selection behavior and to avoid selection bias due to varying inference speed (early, delayed, or potentially false display), we hard-coded the relative position and mapping of products to the fixed VM layout. Given that study participants should not be impacted by technical issues, we proceeded with the hard-coded layout for the purpose of intervention assessment. This allowed instant visualization of all Nutri-Scores once the VM was detected (i.e. 100% of products detected within 0.1s). The intervention within the MR app was then mapped on the VM with its 49 boxes surrounding each product respectively. In addition, we included an explanation menu, start button, introductory text, and a submenu to display nutrients about a particular product (Fig. 6). We argue that this does not seriously limit the generalizability of the results, as technical feasibility is ensured, and as detection accuracy and latency will only improve with future generation headsets and further improved models.

As described in the system introduction, our intervention design included displaying product-specific Nutri-Score labels (Fig. 7). Therefore, we color-coded the products' surrounding boxes with frames color-coded to the corresponding Nutri-Scores (Fig. 6). The interventions were designed using FreeCAD and Unity UI elements. We also included an optional detailed nutrient display, which could be opened via the Microsoft Hololens Clicker devices and displayed the detailed nutritional details (similar to what a color-coded declaration of nutrients would look like, i.e., indicating why a Nutri-Score is high or low; see Fig. 8, right). The product database was specified as an object in the app's C# code. Product nutrients were cached locally on HoloLens to minimize interaction time. At setup time, every VM box was assigned a product key, thus granting access to all the nutrients displayed at runtime. Even if product allocation to boxes changed (e.g. by introducing a new product), changes could be made relatively fast. The updated app could be redeployed on the study device within minutes, after retrieving the new product data including nutrients from the study's server. Nutrients included energy, sugar, saturated fat, sodium, protein, fiber, the share of fruit/vegetable/nuts (to calculate the Nutri-Score), as well as the respective Nutri-Score value, ranging from A (healthy) to E (unhealthy). In addition, every product was labeled either as a "Snack" or as a "Beverage" based on the product identifier allocated to each box.

To assess intervention efficacy, we conducted a randomized controlled trial (RCT). Hence, recruited study participants were allocated to either a treatment (TG) (Figs. 8 and 9, right) or a control group (CG) (see Figs. 8 and 9, left). Before new users

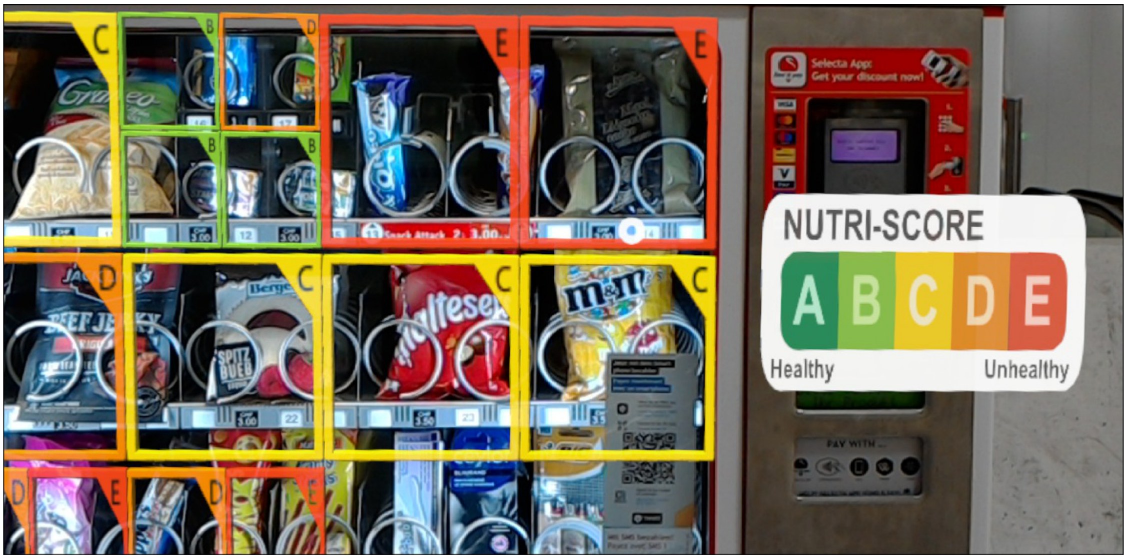
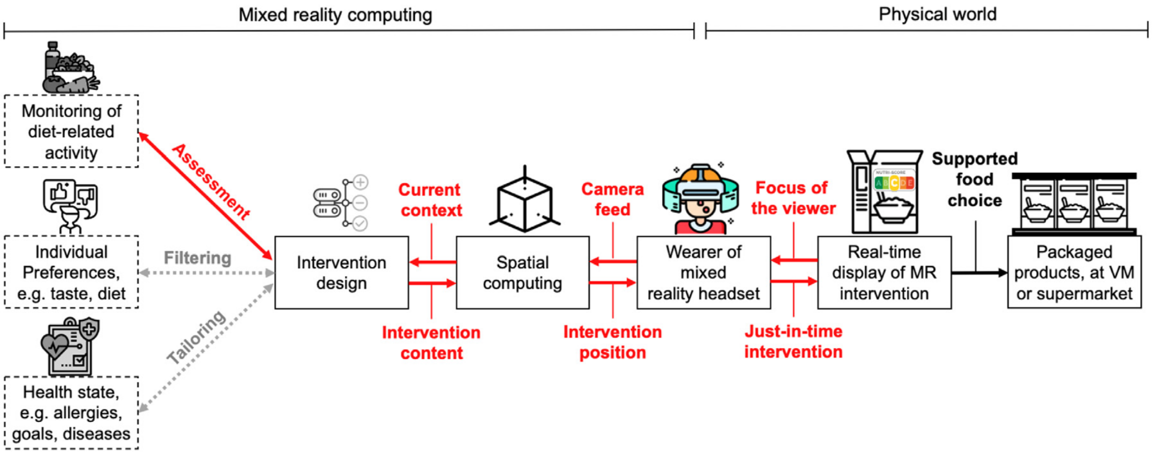


Fig. 6. Computer-vision on packaged products allows novel human–object interactions, e.g. passively triggered diet interventions in the absence of markers or barcodes (Screenshot of Microsoft HoloLens viewing the study's vending machine).


Mixed reality computing
Physical world
Monitoring of diet-related activity
Individual Preferences, e.g. taste, diet
Health state, e.g. allergies, goals, diseases
Assessment
Filtering
Tailoring
Intervention design
Current context
Intervention content
Spatial computing
Camera feed
Intervention position
Wearer of mixed reality headset
Focus of the viewer
Just-in-time intervention
Real-time display of MR intervention
Supported food choice
Packaged products, at VM or supermarket

Impact study	Reference in manuscript	Label
Display of real-time interventions in mixed reality	Chpater 5: RQ4 – RQ7	→
Selection of packaged products in retail	Chapter 6: Dependent variable	→
Intervention design based on diet monitoring, individual preferences, health state	Out of scope	⇢

Fig. 7. Display of real-time interventions in mixed reality.

could use the app, the study supervisor prepared the headset and HoloSelecta app, and familiarized users with the headset. As the experiment (i.e. TG and CG) was implemented within one app, the supervisor manually entered the user identifier and user allocation on the headset prior to new users receiving the headset. In addition, the supervisor selected the menu language (English or German) for every user and recalibrated the VM layout to fit perfectly onto the box grid. To support mapping from virtual to reality, the machine-sized quad was displayed with 50% transparency. Around 20 times per second, the app logs all user interactions, including the focal point, the submenu status, and product selection. This tracking is sent with the survey to a server for continuous evaluation and stored on-device for persistence.

We aimed to examine how an MR headset-mediated nutrition label, in this case Nutri-Score (NS), impacts actual selection of snacks and beverages. To this end, we conducted a non-blinded, supervised randomized controlled trial RCT with a follow-up survey involving 61 users at a VM at Zurich main railway station in Switzerland. When users were allocated to the TG, they received the Nutri-Score in color when considering and purchasing snacks (Fig. 8, right) or beverages (Fig. 9, right). In the control group (CG), users saw white frames during the purchase process, i.e. when selecting snacks (Fig. 8, left) or beverages (Fig. 9, left). Two research questions interested us in this setting. **RQ4:** Does the MR mediated purchase intervention impact the nutritional quality of snacks chosen at the VM? And, correspondingly, **RQ5:** Does the MR mediated purchase intervention impact the nutritional quality of beverages chosen at the VM?

We used convenience sampling for this study, as users were proactively asked to participate when approaching the VM area. Initially, the supervisor set a new three-digit user identifier on his laptop. Next, he asked the prospective user to take a brief introductory survey. After initializing the survey with the respective user identifier and user language, users provided information about their gender, age, and education. In addition, the supervisor guessed users' height and weight for sampling. Users were not

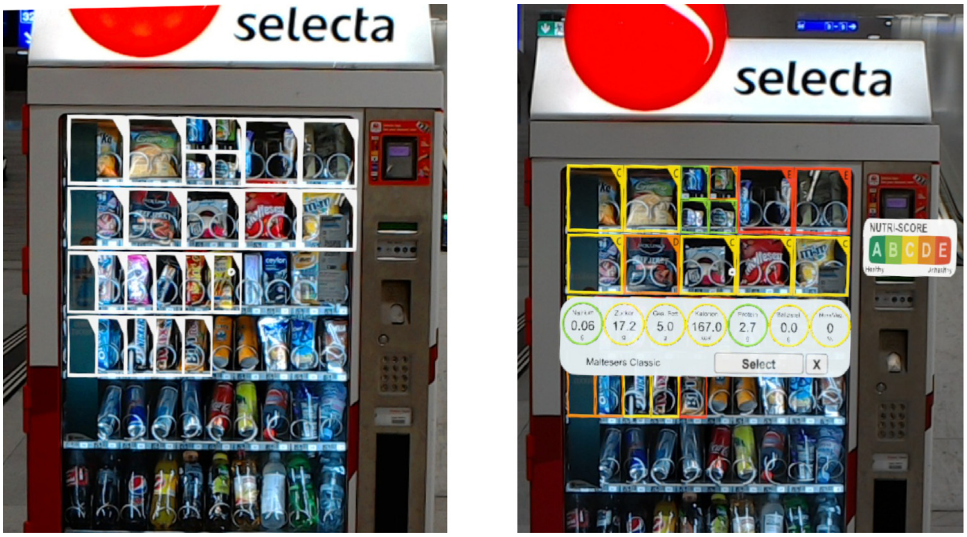


Fig. 8. MR Intervention: (left) CG selecting snacks; (right) TG selecting snacks (with details).

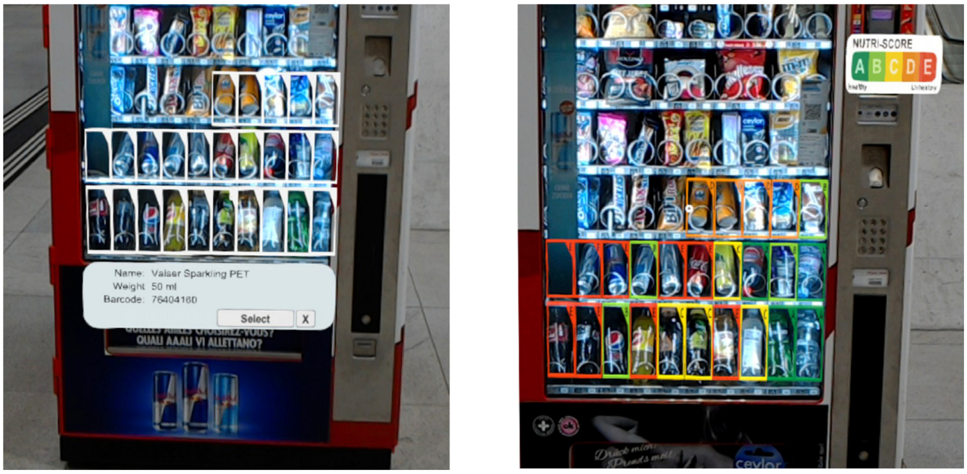


Fig. 9. MR intervention: (left) CG selecting beverages (with details); (right) TG selecting beverages.

asked directly to state their age and gender to prevent priming. Correct weights and heights were collected in a post-hoc survey.

After users completed the introductory survey, the supervisor retrieved the pseudonymized data on his machine. Next, an algorithm placed users into either the treatment group or the control group based on balanced sampling. This was done to ensure that the sum of differences between the groups for every basic item (age, gender, (estimated) height, (estimated) weight, and education was minimal, in order to create balanced samples between TG and CG. When users received the HoloLens device, they were shown a welcome screen and a tutorial. This provided details of how to control the app with head movements and the clicker. Next, users were shown an overview of the available snacks and beverages (as well as the corresponding Nutri-Scores for the TG, i.e. white frames for the CG). The HoloLens Clicker was used for product selections in the app instead of gestures, since gestures by novice users might not have been detected immediately.

The experiment required users to conduct four choice tasks. Every task was first described on an explanatory screen, i.e. before users could begin the task. For the first task, users were asked to purchase a snack of their own choice. Second, users were asked to select a beverage. The "Select" button enabled users to purchase their selected product and to finish the task. Selections were logged under every user identifier. After purchasing both products, users were asked to identify the healthiest snack and the healthiest beverage. This enabled us to assess whether the intervention led users to selection healthy products, or 'only' increased their ability to identify healthy products. During every task, users were only allowed to select one box related to the task. Finally, users took a final, post-hoc survey including items on usage antecedents and randomization checks (Table 4). All items were scored using a seven-point Likert scale (1: strong disagreement to 7: strong agreement). The final survey marked the end of the study, with users offered the selected items for free.

With the post-hoc survey data available, we further investigated the effectiveness of real-time interventions in high-risk user groups, as well as the self-reported intention to use such a system. We thus aimed to answer two research questions. **RQ6**: Is the real-time intervention also effective for risk users, i.e. ones who are prone to make unhealthy food choices? And **RQ7**: The MR support system being novel, do users intend to use such a system again in the future? The analysis of the sample description was performed via descriptive statistics to obtain the means, frequencies (n), and percentages (%). Chi-square and independent sample t-tests were used to test for differences across TG and CG users. All statistical analyses were performed using Python.

Table 2
Accuracy of image classification per *k* labeled images.

Model	Number of images used for training image classification						
	3	6	10	20	40	60	100+
Inception	**0.74**	**0.90**	**0.94**	**0.94**	**0.97**	**0.97**	**0.98**
ResNet	0.65	0.85	0.90	0.92	0.96	0.96	0.97
MobileNet	0.62	0.70	0.86	0.91	0.94	0.95	0.95

Table 3
Mean average precision (mAP) for object detection per *k* labeled images.

Model	Number of images used for training object detection				
	20	30	40	60	100+
Inception	0.42	0.58	0.65	0.75	0.82
ResNet	**0.78**	**0.83**	**0.89**	**0.94**	**0.93**
MobileNet	0.40	0.45	0.53	0.66	0.72

Table 4
Inference speed for the different neural networks on-device and on cloud.

Model	Latency in ms (fps)		
	On-device[a]	Cloud[b] (total)	Cloud[b] (prediction)
Inception	./.[c]	480 (2.08)	241 (4.15)
ResNet	2200 (0.45)	160 (6.25)	76 (13.16)
MobileNet	80 (12.50)	120 (8.33)	31 (32.26)

[a] OnePlus 6T.

[b] Google Cloud P100 GPU.

[c] Did not run on mobile device.

Significance between the treatment and the control group was measured using Mann–Whitney-U tests, since our sample did not follow normal distribution.

6. Results

We then set forth to assess (i) the technical feasibility of CV-based detection of food items in a purchase process (RQ1–RQ3) and (ii) the potential impact of visual real-time on choice behavior (RQ4–RQ7).

6.1. Summarized results of the technical feasibility study

Based on our positive results regarding RQ1, RQ2, and RQ3, we argue that the automatic OD of food items in next-generation MR headsets is technically feasible. Especially noteworthy, headset-mounted cameras allow reliable identification of food items under realistic conditions in near real-time while only requiring limited amounts of labeled training images.

To address the technical feasibility of image classification under realistic conditions in-the-wild, we compared the performance of the three neural networks, i.e. Inception ResNet V2 (''Inception''), ResNet50 V2 (''ResNet''), and MobileNet V2 (''MobileNet'') in our first research question (RQ1). Concretely, we tested the potential of the three NNs to correctly classify labeled image patches, i.e. to identify which one of the total N = 39 products (with over 120 image patches) in the VM sample are displayed in an image snippet. We used classes with a sufficiently large number of *k* labeled image patches and selected 100 (training) + 20 (test) instances for each of the N = 39 applicable product classes as the training dataset. As depicted in Table 2, the more complex the Inception network performs, the less sophisticated (but faster) the model becomes in terms of accuracy of the image classification task. Accuracy may range between 0 (always false predictions) and 1 (always correct predictions). We assume that accuracy rates of 95% (90%) may be considered sufficient for a user support system to meaningfully interpret a user's general dietary behavior. Our study suggests that Inception requires six images for a 90%, and 26 images for a 95% accuracy. Further, the other networks converge to nearly perfect accuracy with a growing number of available training images and are also able to reach 90% and 95% accuracy in image classification. These accuracies would however require labeled training data. For further details on technical feasibility assessment including more in-depth discussion of challenging edge cases, please refer to our previous publication [58].

Next, we assessed the potential of ODNs to support product detection within images of the retail environment (RQ2). The OD task included the identification of image patches that each contain a product and the subsequent correct classification of the detected image patches. Concretely, the ODNs must (i) detect the position of objects in the VM assortment and (ii) correctly classify the detected objects against the labeled ground truth. Finally, the ODN's respective performance can then be calculated by the mean average precision (mAP) [87]. The mAP requires an intersect over union (IoU) between the estimated and the true label [87] of 0.5, as recommended by similar studies in other fields. Mean average precision can range between 0 (none of the predictions achieved an overlay with the actual product of over 0.5) and 1 (all of the predictions achieved an overlay with the actual product of over 0.5). The corresponding, implemented ODNs are (i) Inception ResNet V2 [47] for classification, with Faster RCNN [82] for OD (Inception), (ii) ResNet50 V2 [45] for classification, with SSD and Focal Pyramid Networks (RetinaNet) [83] for OD, iii) MobileNet V2 [44] for classification, with SSD [84] for OD.

Table 3 shows that only the ResNet/RetinaNet architecture achieves an mAP of over 0.9 and that none of the networks achieved an mAP of over 0.95. Concretely, 42 images were required for an mAP of 0.9. For a subset of products with more than 150 labeled image patches available, an mAP of over 0.95 was reached. The ResNet/RetinaNet even reaches an accuracy of 98.6%, when used with all available data (i.e. classes have varying numbers of training images, from 100 to 1000 images, with a mean around 250). Inception was able to achieve a 94.5% mAP when the entire dataset was used. The MobileNet architecture was unable to achieve an mAP of over 90% in our study. This architecture choice seems to play a much larger role as the focal pyramid network architecture of the RetinaNet was far superior when used to detect objects. However, one explanation might be the perfect amount of regularization achieved through downsized images of 320 × 320 pixels, which the other pipelines were not optimized for. Hence, we found that OD yields acceptable results given the right architecture design and a sufficient number of at least 42 labeled training images. For further details on detection assessment, please refer to our previous publication [58].

As described in the introduction to our proposed user support system, CV-based detection of diet-related activities will likely be conducted by interpreting video streams sent from MR headsets. Theoretically, image pooling enables further reducing the error rates of incorrect image classification by using multiple video frames per second to detect objects and by using a mean confidence pooling approach. For a network already achieving high accuracy (e.g. 94% for RetinaNet), predicting the same food item across multiple (i.e. slightly different) frames in a video stream, and given that the user's movement is not rapid, is far more likely than making the same false prediction for a food object multiple times per second.

Thus, can OD run at a real-time latency with a frequency rate of at least multiple frames per second on user devices (RQ3)? Unfortunately, HoloLens 1 fitted with a Unity UWP was neither able to run Tensorflow OD models on the device, nor support fast websocket-based transfers to the server. Therefore, we used a comparable mobile device that supports Android and has native support for Tensorflow Lite for latency validation (i.e. OnePlus 6T) to assess the models' inference speed on user devices. It can be argued that at least future-generation MR reality headsets will have the computational capabilities of such a mobile phone. Therefore, these inference speed estimates could be interpreted as lower bounds to future detection rates. For cloud-based validation, we used a Linux computer with support for gRPC (faster transfer of image data than REST).

Table 4 shows that only MobileNet is fast enough to run in memory on the mobile device, achieving predictions at 12 frames per second (fps). On cloud, MobileNet performs slower, as image upload slows down prediction speed. Nevertheless, it can still achieve 8 fps, while the inference rate per single image is 80 ms for inference with MobileNet on the phone and 2200 ms with Resnet. The more complex models perform better in the cloud, with MobileNet taking around 120 ms, ResNet 160 ms, and Inception taking 480 ms for each prediction. Increasing resolution from 320 × 320 to 640 × 640 prolongs latency equally for all networks by 10 ms. These results are in line with similar studies [86]. In theory, through image pooling, the achievable accuracy could approximate 100%. The chance of not finding the right object within each second is reduced by an increasing number of predictable frames per second: $error = (1 - mAP)^{fps}$. Because this requires the user to record "different" images (i.e. the user slowly moves), and probably would still not help identify edge cases (i.e. distorted shapes, reflections, occlusions, etc.), this is a rather theoretical consideration, but should encourage further research. The trade-off between fast neural networks and their respective accuracy remains a very important topic for academia as well as for real-world use-cases. Given the possible frame rate of 6.25 predictions per second for RetinaNet-based OD (including image classification with ResNet) through gRPC on the cloud, we argue that the inference speed of sufficiently accurate models is sufficient for image pooling.

6.2. Summarized results of our user study

Next, we assessed the impact of visual real-time interventions on actual food choices, as postulated by RQ4 and RQ5. Participants in our in-the-wild study had a mean age of 29.83 (SD = 13.38) years, 33% were female, and 53% had tertiary education. While 56% of participants stated that they rarely use VMs, 21% reported monthly usage, and 16% reported weekly or daily usage. We did not find any significant differences between the Treatment Group (TG) (N = 31) and the Control Group (CG) (N = 30) across any of the sample dimensions, which indicates successful randomization of the randomized controlled trial (RCT). The study protocol envisaged that participants undergo four tasks in a consequential sequence: (1) choose a snack of their own choice and (2) select a beverage of their own choice. After selecting their true preferences (they were given the products for free after successful participation in the study), they were asked to identify (3) the healthiest snack and (4) the healthiest beverage (see Table 5).

Regarding impact on snack choices (RQ4), we observed significant differences. Users in the TG with the Nutri-Score intervention selected products with 48% less saturated fat per 100 g on average. This result can be interpreted as a large reduction. Other differences include TG users choosing healthier snacks (−5.8 Nutri-Score points), as well as a 14% reduction in sugar and a 9% reduction in calories. The Nutri-Score improved from 11.37 (CG) down to 10.00 (TG) (Scale −15 = healthy to 40 = unhealthy). Saturated fat content of the selected snack reduced from 9.95 g/100 g (CG) down to 5.16 g/100 g (TG).

When identifying the healthiest snack, highly significant differences between TG and CG were observed. The average Nutri-Score points of the chosen snacks improved from 10.57 (CG) down to 4.84 (TG). The real-time visual intervention seemed to greatly support identifying healthy snacks. Among other qualitative feedback, the TG and CG both stated the VM offers little to no substitutes for certain snack products. The only rather healthy snacks were chewing gums, a product that can hardly be considered a perfect substitute for consumers who planned to buy a salty sack. In fact, no snack received the healthiest Nutri-Score "A" (less than -1 Nutri-Score point). Nevertheless, (i) the significant improvement in saturated fat (and the improvements in other nutrients) within the snack assortment with limited healthy substitutes available, as well as (ii) the significantly improved ability to identify healthy snacks via the real-time intervention led us to accept RQ4 (see Table 6).

Regarding users purchasing a beverage of their own choice (RQ5), significant differences were observed, as shown in our previous publication [34]. Users with the Nutri-Score intervention selected products with 5.8 Nutri-Score points less on average

Table 5
Sample description of study participants (N = 61).

Age (years)	Mean (SD)	Gender	Count (%)
	29.83 (13.38)	Female	20 (32.8%)
		Male	38 (62.8%)
		Other	3 (4.9%)
MR experience	Count (%)	Education	Count (%)
none at all	26 (42.6%)	Primary	12 (19.7%)
some (tried)	29 (44.3%)	Secondary	14 (23.0%)
a lot (frequent)	4 (6.6%)	Tertiary	32 (52.5%)
N.A.	4 (6.6%)	NA	3 (4.9%)
Weight (BMI)	Count (%)	VM frequency	Count (%)
Underweight	2 (3.3%)	Infrequently	34 (55.7%)
Normal	46 (75.4%)	Monthly	13 (21.3%)
Overweight	7 (11.5%)	Weekly	5 (8.2%)
Obese	6 (9.8%)	Almost daily	5 (8.2%)
		N.A.	4 (6.6%)

VM: Vending machine.

Table 6
Snack choices across treatment group (N = 31) and control group (N = 30).

Task	Item			
	Snacks			
	TG[e]	CG[e]	ΔTG-CG (%)[f]	P
1. Select a snack of your own choice				
NS[a]	10.00	11.37	−1.37	.35
Energy[b]	324.09	354.90	−30.81 (−9%)	.18
Saturated fat[c]	5.16	9.95	−4.79 (−48%)	0.003*
Sugar[c]	23.51	27.26	−3.75 (−14%)	.13
Sodium[c]	0.43	0.27	+0.16	.23
Protein[c]	9.30	8.32	+0.98	.35
Fiber[c]	0.98	1.49	+0.51	.29
Time[d]	52.43	46.14	+6.29	.25
3. Identify the healthiest snack				
NS[a]	4.84	10.57	−5.73	0.000*
Energy[b]	238.83	394.70	−155.87 (−39%)	0.000*
Saturated fat[c]	2.79	5.36	−2.57 (−48%)	0.000*
Sugar[c]	7.03	13.18	−6.15 (−46%)	0.001*
Sodium[c]	0.28	0.80	−0.52	0.000*
Protein[c]	5.69	15.09	−9.40 (−62%)	0.000*
Fiber[c]	0.55	2.52	−1.97 (−77%)	0.002*
Time[d]	36.23	31.27	+4.96	.16

*Significant at 5% level.
[a] Nutri-Score has a point scale from −15 (very healthy) to 40 (very unhealthy).
[b] in KJ/100 g.
[c] in g/100 g.
[d] in seconds.
[e] mean values.
[f] Percentage change for differences over 1 g/100 g or 10 KJ/100 g.

Table 7
Beverage choice across treatment group (N = 31) and control group (N = 30).

Task	Item			
	Beverages			
	TG[e]	CG[e]	ΔTG-CG (%)[f]	*P*
2. Select a beverage of your own choice				
NS[a]	−0.97	4.80	−5.77	0.009*
Energy[b]	22.03	33.47	−11.44 (−34%)	.06
Saturated fat[c]	0.05	0.03	+0.02	.37
Sugar[c]	4.88	6.79	−1.91 (−28%)	0.049*
Sodium[c]	0.02	0.02	0	.43
Protein[c]	0.34	0.23	+0.11	.38
Fiber[c]	0.00	0.00	0	.50
Time[d]	35.52	28.19	+7.33	.15
4. Identify the healthiest beverage				
NS[a]	−14.37	−13.48	−0.89	.27
Energy[b]	0.63	2.07	−1.44	.27
Saturated fat[c]	0	0	0	.16
Sugar[c]	0.15	0.49	−0.34	.28
Sodium[c]	0	0	0	.35
Protein[c]	0	0.02	−0.02	.16
Fiber[c]	0	0	0	.50
Time[d]	23.97	17.88	+6.09	.38

*Significant at 5% level.
[a] Nutri-Score has a point scale from −15 (very healthy) to 40 (very unhealthy).
[b] in KJ/100 ml.
[c] in g/100 ml.
[d] in seconds.
[e] mean values.
[f] Percentage change for differences over 1 g/100 ml or 10 KJ/100 ml.

(and therefore healthier products), with significantly less sugar (28%), and significantly less calories (34%, albeit only on a 90% confidence interval) per 100 ml. The Nutri-Score declined from 4.8 (CG) down to −0.97 (TG) (Scale −15 = healthy to 40 = unhealthy). The sugar content of the selected beverage declined from 6.79 g/100 ml (CG) down to 4.88 g/100 ml (TG). The energy content of the selected drink reduced from 33.47 KJ/100 ml (CG) down to 22.03 KJ/100 ml (TG). Regarding users selecting the healthiest beverage, no significant differences between TG and CG were observed. The average Nutri-Score declined from −13.48 (CG) down to −14.37 (TG). These scores are both very close to the perfect score (−15), as most users correctly selected mineral water to be the healthiest drink, a widely known fact that even most users in the control group were well aware of. Qualitative feedback from both TG and CG included spontaneously remembering mineral water being healthiest, when asked to identify the healthiest beverage. Most users seemed able to correctly identify the two different mineral waters as the healthiest option. Only a few users selected orange juice as the healthiest option. The significant improvements in Nutri-Score, energy (−34%), and sugar (−28%) within the beverage assortment led us to accept RQ5.

As discussed more in-depth in our previous publication [34], sociodemographic segments prone to diet-related diseases can benefit from passively triggered, visual real-time interventions in MR headsets. Was the real-time intervention also effective among risk users, who are prone to unhealthy food choices? (RQ6). When comparing overweight (BMI over 25 kg/m^2) to non-overweight users, the intervention seems more effective among overweight users. Both overweight and non-overweight users exhibited improved values for sugar, energy, and Nutri-Score for the selected beverages when receiving the intervention (TG). When comparing higher educated (e.g. tertiary education) to less educated users, the intervention seems to be supportive in both segments: Less educated users experience a significant improvement of the Nutri-Score of the selected product, while higher educated ones choose healthier products on average. Similarly, preexisting food literacy (measured by a food literacy questionnaire) correlates with significant improvements in the Nutri-Score of selected products. Nevertheless, also less food-literate users experience improved Nutri-Score, sugars, and energy of the selected products. Further statistical tests comparing overweight versus non-overweight, highly educated versus less educated, food literate versus illiterate receivers of the intervention do not suggest that the intervention makes either group perform the tasks significantly better. Given that all risk groups (i.e. less food-literate, less educated, and overweight users) benefit from the real-time intervention, we accepted RQ6.

After the experiment, all participants took a final survey. This included questions on usage antecedents and additional randomization checks, to better understand whether study participants would reuse a similar system in the future (RQ7). All items were encoded as 7-stage Likert scales, ranging from 1 (strong disagreement) to 7 (strong agreement). We found that the MR-based support system is quite popular with study participants. Qualitative feedback included the desire to take the HoloLens device home or requesting that such a visual intervention app be released for their smartphones. Negative feedback included the weight of the HoloLens headset and lack of familiarity with or understanding of the Nutri-Score concept. One participant even suggested developing a Nutri-Score intervention for color-blind people, as they might be unable to interpret the red-to-green color-coding of product frames.

Users in the TG stated a significantly higher intention to use (3 items) and had high performance expectations after trying out the intervention themselves (5 items). They mentioned that the headset system is helpful, educative, supports faster and healthier product decision, and thus encourages starting or maintaining a healthy diet. In addition, they stated that their social environment (2 items) supported and favored them using such a system. Observations showed that all three constructs (Intention to use, performance expectation, and social influence) ranked significantly higher in the TG than in the CG. Interestingly, users of both groups felt rather unobserved and claimed to have selected their “true, unbiased” behavior. They also users indicated that they had enjoyed the HoloSelecta experiment and considered the experience “fun” and “exciting”. Therefore, we accepted RQ7, as users states that they were highly motivated to use such a headset-based support system for healthy food choices if available.

7. Discussion

Although the majority of people claim to be interested in nutrition and maintaining a healthy lifestyle [25], counterintuitively perhaps diet-related diseases such as obesity and cardiovascular diseases are steadily increasing [88]. Current smartphone-mediated dietary mHealth applications fail to be mass-adopted [25] and are often ineffective [20] due to effort-intensive and error-prone manual logging, aside from the salience required and delayed interventions. Thus, developing such new and more user-centric approaches certainly seems to be called for. To this end, we introduce a novel integrative framework in line with the principles of the Internet of People (IoP) leveraging the joint application of CV-based detection of diet-related activities and just-in-time visual interventions via next-generation MR headsets.

Our proposed framework describes how such a vision might be realized in the future. Our technical feasibility study demonstrated the current capabilities and limitations of NNs in supporting such a framework while our impact study examined its potential to support healthier food choices. Specifically, we were interested in whether wearable smartglasses, which leverage CV and MR real-time interventions, support healthier food selections at VMs, i.e. locations where consumers may in fact (intend to) purchase unhealthier foods and beverages. While most research in this domain so far has focused on classifying food products via CV through pre-fabricated image datasets [53–55], or has only assessed visual interventions via augmented reality on handheld devices [89], our study contributes to the literature by leveraging MR headsets and interventions, thereby overcoming the drawbacks of current smartphone-mediated interventions.

Our proposed integrative framework currently represents a vision for the IoP rather than a prediction with a foreseeable timeline. However, since relevant tech companies (including Apple, Facebook, Microsoft, and Magic Leap) have announced that they intend to release their consumer-oriented MR headsets as next-generation computer systems in the coming years, it might only be a matter of time for such a vision to become reality. Whether such headsets will be adopted as quickly as smartphones or will take longer (such as the adoption of VR headsets) remains to be seen. Even if mass adoption of wearable headsets might still be years away, our proposed system might already be enabled in diet-counseling programs to educate patients in making healthier food choices. Also, the discussed subsystems (i.e. MR interventions and CV-based detection of diet-related activity) might also be adopted independently by home assistant systems such as Google Home or Amazon Alexa. Future versions of these systems could leverage cameras to detect eating activity, for example, or automatically display visual feedback after ordering food online. Still, the insights from the validation study lead us to conclude that implementing such as framework is likely to be technically feasible in terms of improving food choices.

Compared to existing studies on detecting packaged products [53–55], our technical feasibility study is the first to collect, label, and apply real-world images for product detection. To the best of our knowledge, the collected dataset of 10’035 labeled product instances represents the largest labeled image dataset to include product identifiers (GTINs). It therefore allows integration of product metadata. As labeled image data remains one of the largest limitations to identifying packaged products, our study demonstrated higher accuracy rates for product classification and OD due to the increased number of labeled instances per product. Compared to accuracy rates of up to 85.3% in [54], we observed product classification accuracy rates of 95%–97.7% (see RQ1) with 100 images per class via transfer learning [58]. Furthermore, we concluded that at least six (for 90% accuracy) to 26 images (for 95% accuracy) are required to train relevant models (see RQ2). Our study therefore provides insight into the feasibility of image classification on packaged products. It also demonstrates that real-world implementation of image detection is feasible. Thus, a limited amount of investment and effort will be required in future studies.

Moreover, we observed that OD under real-world conditions is possible, yet heavily depends on architecture choice (see RQ3). The ResNet/RetinaNet architecture implemented in our study achieved acceptable results, i.e. mean average precision (mAP) of over 90%, after being trained on 42 images. An mAP of over 95% was observed when 100 images per product were used. Again, this is much higher than in related studies, which leveraged less images per product and observed an mAP of 73.5%–76.93% [54]. In further studies, we believe it would be interesting to assess the accuracy of already performant networks by using video stream data and by factoring latency in detail. Especially the trade-off between the complexity of the neural network models and their respective inference speed (i.e. a device’s upload speed to a cloud-based server and the headset’s own computational capability) is worth assessing. Our study suggests that server-side GPU-supported predictions are much faster compared to local predictions as they are limited by the device’s hardware. Differences in latency allow for improved accuracy when predicting products on the server, especially via image pooling. Given ResNet’s mAP of 0.94, the device can only predict once every 2.2 s. Server-side predictions can make 13.75 predictions in the same amount of time, even factoring in the image transport via gRPC, and thus yielding a theoretical error of close to zero: $(1 - 0.94)^{13.75}$.

Of course, these results do not hold in the real world, as many factors impact accuracy (e.g. lighting conditions, connectivity, orientation, product assortment, etc.). A particular feature of our research context was that HoloLens did not support TensorFlow at the time of implementation, meaning we could not compare on-device prediction in the real world. Once HoloLens 2 is available, it is will possible to realistically assess device-based capability of predicting packaged products in the real world. To conclude, the current advantages of cloud-based inference and the advantages of image pooling over multiple frames in a video sequence may prove a promising approach to achieving higher degrees of accuracy and mAP. In other words, the vision of a global detection of packaged products is likely to occur. Yet whether this will occur via one or many NNs, which, for example, might be connected via a knowledge graph, remains to be researched. Given a certain location, an MR headset application might then be able to retrieve currently valid pretrained NNs to make a local prediction on the respective shelf or VM, and thereby benefit from specially trained networks. An alternative global NN for predicting products would in turn have to be pretrained on labeled training data from a sufficient number of shelves to predict products on all types of shelves (e.g. from VMs to large supermarkets). Hence, our study contributes to the nascent body of research on the CV-based identification of food items by demonstrating the feasibility of detecting packaged products [90].

Regarding our impact study, this paper presents the first real-world validation of a MR headset-mediated intervention on packaged products. Previous studies have leveraged handheld tablets [89], monochromatic glasses with text-based interventions [71], or wearable clothes-integrated cameras [72], all with varying disadvantages. For example, handheld devices may prove impractical during shopping, text-based interventions non-trivial to interpret, and textile-integrated cameras need a second screen to display interventions. The only document describing the displays of diet-related interventions inside a MR headset is a patent by Microsoft [65]. However, nothing is yet known about such a system’s efficacy under realistic conditions. We therefore conducted an in-the-wild user study using Microsoft HoloLens (1st generation) to validate the potential impact of visual interventions during the purchase process at a VM. We observed significant differences regarding food and beverage choices (RQ4, RQ5): significant reductions in sugar and calories among beverages and saturated fat in food items.

This study hence contributes to existing work on purchase-related mHealth [91] and food labels [74] that both examine two-dimensional static cues and their impact on food decisions. Our findings contribute to previous research by suggesting that dynamic, three-dimensional environmental cues can achieve a similar effect, and could even be shown automatically in the absence of a printed food label. In fact, we argue that in some cases our proposed solution is superior because it does not require salience or active user input. As such, it enables hands-free shopping, and thereby overcomes one of the strongest barriers of mHealth-based solutions [20,21,23]. We also find that MR headset wearers more easily identify healthy snacks when supported by the intervention (see Table 7). Additionally, our impact study concluded that irrespective of their health state or nutritional literacy users were able to profit from the intervention and indicated high motivation to use a wearable device for food choice support (RQ6, RQ7). This finding is encouraging, as food-illiterate users are traditionally underserved by contemporary mHealth interventions. Finally, we found that the MR-based support system was popular among study participants, whose qualitative feedback included the desire to take the MR headset device home.

Our findings have certain limitations, of course. We believe that four limitations warrant special attention. First, we applied a convenience sampling and single-blinded observation to our research design. We mainly sampled male commuters and travelers, who might not be entirely representative of a population. Further studies should apply a stratified sampling approach to achieve a more representative sample. Moreover, the technology used and the "single-blindness" of our study may have affected the outcome, especially for participants who required more technical support than others. Fortunately, users who were less familiar with the intervention technology did not significantly differ regarding the nutritional quality of their selected items. Nonetheless, with further investment in the technological set-up of the study, future research designs should take note of users' technological familiarity and degree of blindedness in the intervention setup. We recommend designing research settings in which users systematically feel less observed.

Second, the VM research setting represents another potential limitation regarding the transferability of the results. We considered this point in multiple ways. Since there is a general lack of training data for packaged products, we selected a substitute that could also be "observed in the wild" [92] and be generalized to a supermarket setting. A VM, if understood as a shelf with products, is comparable to a collection of shelving found in a retail store. Furthermore, we selected a representative VM, operated by the European market leader (Selecta: 0.125M machines worldwide and five million consumers daily). From such a perspective, it would be interesting to compare these findings with other studies on the same or other VM contexts, in order to increase the generalizability of our findings. It would also be worth conducting validation studies – moving from VMs to supermarkets – when training data becomes available on a large scale. We believe crowdsourcing product image data might be a viable option for populating labeled image databases from shelves around the world.

The research setting may have provided a third potential source of limitation. We argue that the layout of the VM does not limit the external validity of CV-based detection, as OD and image classification are conducted in two separate steps. First, the NN estimates where an object of interest could be, and only subsequently predict which product might be involved. Therefore, detecting "areas of interest" needs to proceed independently of hardcoding layout and is reliable for detecting most product positions. Therefore, a new product or changed position does not necessarily entail a false prediction, but rather nonprediction if the product is unknown in the training data (i.e. confidence thresholds will likely be too low to make a meaningful prediction). This approach is similar to the architectures of Tonioni's work on detecting products on supermarket shelves (e.g. Grozi datasets, i.e. GP-180 and Grozi 3.2k) [54,55]. Our CV pipeline therefore also yields similar results to [54,55] when trained and applied to those datasets, and as such does not inherently possess a performance bias toward VMs.

Relatedly, a fourth potential limitation may arise from combining the research environment and surrounding research design choices. This limitation primarily involves hard-coding our headset setup. As discussed, given the high dependence on external factors (e.g. lighting conditions, distance and angle, connectivity, configured confidence thresholds), we opted for a hard-coded layout to control for such potentially varying factors. Integrating these factors into our context might have delayed inferring certain products and displaying related interventions, and therefore might have impacted users' choices. Hence, we opted to hard-code the positions within the impact study to avoid biased selections based on delays. However, we argue that hard-coding does not necessarily limit outcome generalizability in terms of the impact study. Moreover, further technological improvements (e.g. availability of labeled image data, improved connectivity via 5G, edge computing, improved headset hardware), which support the ubiquitous CV-based recognition of food products, will reduce differences in inference frequency between products over time.

In the future, research is needed to leverage CV in order to scale such MR interventions to supermarket shelves with thousands of products. Further, validation studies are needed to compare interventions to printed labels (e.g. inside the VM) or to mHealth interventions aimed at improving food selection. If wearable smartglasses become widely adopted consumer devices, integration of MR interventions into personnel-based counseling programs can be expected. In turn, these approaches will be able to complement traditional nutrition interventions in situations where consumers intend to make food purchase selections. Future research could shed light on more comprehensive system designs and different interactions between associations, concepts, goals and awareness, as well as outcomes potentially optimizing MR-based nutrition interventions. Finally, MR headsets could support developing future FoPL labels, as such headsets could be tested with MR glasses and potentially integrate eye and gaze tracking. This aspect might be especially useful for developing new or tailored FoPL labels, in turn enabling comparing different labels without physical changes.

8. Open challenges and future avenues of research

This section adds to our aforementioned research ideas and discusses the most pressing challenges and the resulting research potentials identified above. We selected the following topics as these promise to be important avenues of research, with a view to successfully implementing IOP-based user support systems in the context of healthier food choices.

Currently, reliable global CV-based food detection models are still missing. More specifically, disjunctive, pretrained special-purpose models exist for packaged products as well as composed meals, yet need to be combined to support diet-related interventions via MR headsets. Merging these domains, i.e. developing CV models capable of detecting meals and packaged products, also seems worth investigating. To this end, the potential of knowledge graphs, which enable searching only within a relevant, small subset of all globally available meals or products, is fascinating. For example, if the support system, based on a user's location, knows that he or she is currently at a specific store or restaurant, the CV model could focus on assessing the possible food item candidates available at that particular venue rather than searching among all food items.

Furthermore, limitations of CV-based detection of diet-related activities include edge cases, for which visual appearance is a non-ideal proxy for identifying nutritional composition. Edge cases also include look-a-like food and beverage products (e.g. Coca Cola and Coca Cola Zero) or some types of composed meals (e.g. nutritional quality of similarly looking soups may very strongly). It will hardly be possible through CV alone to identify differences that are not subject to human vision but instead are inferred through taste (e.g. salt content of a soup). Here, alternative, novel methods including near-infrared spectroscopy (NIRS) could enable differentiating sugar-free and sugar-rich items, for example. Still, we consider the autonomy of user input to be more importance than attaining perfect accuracy. In contrast to CV applications such as self-checkouts (e.g. Amazon Go) or inventory stock keeping via robots, diet tracking does not require perfect, yet rather acceptable accuracy [20,36]. False positives are less critical: In the dietary domain, tracking general behavior is often more important than accuracy. In fact, if accurate dietary assessments are needed, state-of-the-art methods include blood sampling or spot-urine assessments via medical laboratories. Nevertheless, a validation of MR-mediated tracking with food diaries or bio samples remains unprecedented.

Detection of eating behaviors and activities is likely to be an important avenue of research to extend the IoP based framework presented. Aside from detecting food items, MR headsets can also assess a user's eating speed and eating times via CV [93]. For example, growing evidence shows that slower eating and certain eating time windows correlate with improved health [94]. The scarcity of such studies likely stems from sampling, educating, and monitoring participants requiring great effort. While self-reporting may ameliorate this situation, researchers must then account for various biases related to food habits. MR headsets offer the chance to improve sampling size and tracking eating activities (e.g. speed and frequencies). Such setups could also plausibly be extended with other tested systems for identifying food activities (e.g. smart watches, electromyography, or automated wrist motion tracking). Such approaches could help overcome some of the major limitations in nutritional science, and thus enable better understanding the relationship between a user's intentions, intake, behavior and health outcomes.

Food label design might represent another interesting avenue of research. Intervention efficacy as observed in our impact study might also partly arise from food label design. Although Nutri-Score represented the state-of-the-art in terms of effectivity and adoption in previous studies, other frameworks may prove more effective over time. Such potentially more personalized interventions could include interpreting recent activities (i.e. previously consumed food or physical workouts), personal food preferences (i.e. taste, preferring a vegetarian diet), information on an individual's health state (e.g. diseases or allergies). Study questions in this area could include comparing different intervention designs, potentially ranging from monetary reward systems to sophisticated three-dimensional gamification elements (e.g. avatars). Furthermore, interventions could integrate other non-food-related information such as feedback on product sustainability (e.g. palm oil content). Food waste prevention could be assessed through MR-mediated interventions. Headset cameras could track leftovers on a plate or hand gestures indicating food waste. Finally, MR headsets could also support developing future front-of-package labels, as they could be tested with MR glasses capable of measuring eye-tracking and gaze. This aspect might be useful for developing new or tailored food labels, enabling comparison of different labels without physical changes.

Another very important study aspect of CV-mediated MR headsets is user-intervention interaction. CV and MR allow for just-in-time-interventions (JITAI), considered best practice in digital health interventions [95]. Users can thus be prompted, nudged, or reminded anytime or just in time prior to, during, or directly after eating. To this end, we conducted an explorative analysis that compared users who spent varying lengths of time in the real-time intervention. We measured head poses through constant logging of the focal point. As expected, users who spent less time in the intervention seem to be "pre-determined" and only look at a few products in the assortment (presumably due to a preformed preference). In contrast, users who spent longer in the intervention "browsed around" and checked multiple alternative food items (Fig. 10). Hence, different exposure times to visual intervention impact food choice. Therefore, future intervention design could leverage the ability to measure eye tracking to adapt interventions to users' current focus. Future studies could thus research when such prompts ideally should occur, whether prompt archetypes exist that better cater to certain user groups, and of course what such prompts should look like (e.g. messages, avatars, scores). A large body of research has studied these topics in the area of mHealth, which makes it interesting to see how previous findings might translate to the MR context. We suspect that next-generation MR headsets will become less obtrusive for users, who could therefore potentially adhere to diet interventions more long-term than happens with short-lived mHealth retention. Hence, research on MR-mediated JITAI will be able to better study feedback and motivation mechanisms, which would be central to developing a sustainable, personal food choice support system.

Another important research topic that remained underdiscussed in this paper, yet is central to the IoP paradigm are users. Our user study found that participants indicated high intention-to-use and performance expectations, as well as hedonic motivation. This suggests that with the uptake of MR headsets in the future, such diet intervention formats could become popular. However, we believe that IOP approaches as demonstrated in this paper allow for a deeper understanding of users' eating habits. Studies on eating behaviors are often conducted in the laboratory or under controlled conditions [96]. Observation may lead users to reduce their calorific intake, even if observation is missing [97]. Future studies could assess whether introducing MR headsets might have such effects, whether users experience patronization, or even whether users grow accustomed to their wearable headset over time and behave naturally. Such study findings would be interesting to discuss in the context of diet pattern detection and in the context of improving consumer behavior in the long-term, as users might revert to their former (unhealthy) behavior once an intervention or the feeling of monitoring wanes.

Finally, but not trivially, validation studies are required to compare existing dietary interventions with novel MR-mediated interventions. For example, printed labels (e.g. inside the VM) or mobile health interventions aimed at improving food selection need to be compared to MR-mediated interventions, especially in future long-term scenarios. We have provided various pragmatic and theoretical arguments about the drawbacks of existing approaches (see Section 2). Such claims, however, require further validation. Validation studies should not be seen as a trivial exercise. They instead might highlight the necessity of using existing approaches for particular tasks in the diet-related behavior change process. For example, personal contact (e.g. in form of counseling or food education) might remain necessary and cannot simply be replaced but rather needs to be extended with diet-related IOP approaches. As such, integrating MR-based diet interventions in personnel-based counseling programs represent an interesting possible foray for nutrition studies. In turn, these approaches could complement traditional nutrition interventions in situations where consumers intend to make food purchase selections. Future research could shed light on more comprehensive system designs and different interactions between associations, concepts, goals, and awareness, as well as on outcomes potentially leading to optimized MR and IOP based nutrition interventions.

9. Conclusion

We have considered what an IOP-based user support system for improved diet behaviors might look like in the future and have focused on real-time detection and intervention of food items. We have discussed the necessary elements of such a system. Based on our previous studies, we have highlighted various aspects of real-time detection and intervention. CV- and MR-mediated identification and intervention of packaged products is still a

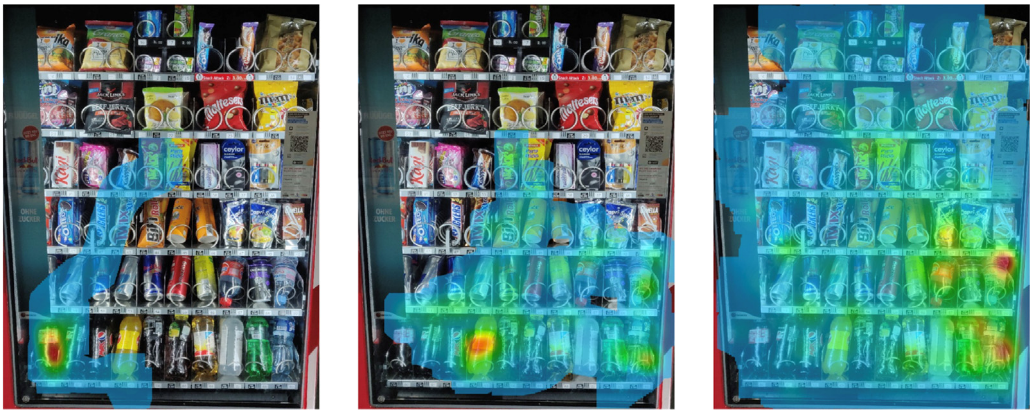


Fig. 10. Head pose tracking of different users who spent varying amounts of time in the intervention (left: short to right: long).

nascent field and lacks publicly available datasets and published research. Nevertheless, our studies provide insight into how image classification and OD represent promising approaches to an IOP-based user support system. Finally, we provide interesting avenues for future research, which could enable IOP-based user support systems that help users to make healthier food choices.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This manuscript is an extended paper based on prior research conducted and published in two previous conferences papers [34, 58]. We would like to thank the reviewers of the 9th International Conference on the Internet of Things for the best paper award and the encouraging feedback that paved the way toward this manuscript. Second, we published a subset of the results from the user study on the impact of MR-mediated purchase interventions aimed at supporting healthy beverage choices [34] and wish to thank the reviewers and audience of the 27th ACM International Conference on Multimedia for their fruitful feedback. Submitting our previous work to both conferences helped to shape the conceptual integration of both concepts, i.e., automatic sensing of diet-related activities and corresponding real-time interventions via MR headsets, into a more holistic vision for supporting users in making healthier food choices in the emerging Internet of People. In addition, we thank the Selecta Group for providing us with the opportunity to conduct the user study at their VMs at Zurich main railway station in Zurich, Switzerland. Finally, we thank everyone who volunteered to participate in the user study presented here.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.future.2020.07.014.

References

[1] F. Branca, et al., Transforming the food system to fight non-communicable diseases, BMJ (2019).

[2] J.D. Stanaway, et al., Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: A systematic analysis for the global burden of disease stu, Lancet (2018).

[3] B.M. Popkin, The nutrition transition and obesity in the developing world, J. Nutr. 131 (3) (2001) 866S–870S.

[4] F. Sassi, M. Devaux, M. Cecchini, E. Rusticelli, The obesity epidemic: analysis of past and projected future trends in selected OECD countries, OECD Health Work. Pap. 45 (45) (2009).

[5] R.J.R. Levesque, Obesity and overweight encyclopedia of adolescence, 2018, [Online]. Available: http://www.who.int/mediacentre/factsheets/fs311/en/ [Accessed: 23-Jan-2016].

[6] N. Alexandratos, J. Bruinsma, World Agriculture Towards 2030/2050: The 2012 Revision, Vol. 20, ESA Work. Pap. No. 12-03, Food Agric. Organ. UN, 2012, p. 375, no. 4.

[7] A. Omran, The epidemiological transition, Int. Encycl. Popul. (1982) 172–175.

[8] B.M. Popkin, An overview on the nutrition transition and its health implications: the bellagio meeting, Public Health Nutr. 5 (1a) (2002) 93–103.

[9] K. Srinath Reddy, M.B. Katan, Diet, nutrition and the prevention of hypertension and cardiovascular diseases, Public Health Nutr. 7 (1A) (2004) 167–186.

[10] H.H. Vorster, The emergence of cardiovascular disease during urbanisation of africans, Public Health Nutr. 5 (1A) (2002) 239–243.

[11] P. Puska, Nutrition and global prevention on non-communicable diseases, Asia Pac. J. Clin. Nutr. 11 (SUPPL. 8) (2002) 755–758.

[12] V. Kalnikaite, J. Bird, Y. Rogers, Decision-making in the aisles: Informing, overwhelming or nudging supermarket shoppers?, Pers. Ubiquitous Comput. 17 (6) (2013) 1247–1259.

[13] C. Julia, et al., Perception of different formats of front-of-pack nutrition labels according to sociodemographic, lifestyle and dietary factors in a french population: Cross-sectional study among the NutriNet-Santé cohort participants, BMJ Open (2017).

[14] Z. Talati, S. Pettigrew, H. Dixon, B. Neal, K. Ball, C. Hughes, Do health claims and front-of-pack labels lead to a positivity bias in unhealthy foods?, Nutrients (2016).

[15] K.G. Grunert, J.M. Wills, L. Fernández-Celemín, Nutrition. knowledge, Nutrition knowledge and use and understanding of nutrition information on food labels among consumers in the UK, Appetite 55 (2) (2010) 177–189.

[16] E. Volkova, et al., 'Smart' RCTs: Development of a smartphone app for fully automated nutrition-labeling intervention trials, JMIR mHealth uHealth 4 (1) (2016) e23.

[17] M. Dehghan, N. Akhtar-Danesh, A.T. Merchant, Childhood obesity, prevalence and prevention, Nutr. J. 4 (Table 1) (2005) 24.

[18] B. Swinburn, G. Sacks, E. Ravussin, Increased food energy supply is more than sufficient to explain the US epidemic of obesity, Am. J. Clin. Nutr. (90) (2009) 1453–1456.

[19] R.F. Kushner, Barriers to providing nutrition counseling by physicians: a survey of primary care practitioners, Prev. Med. (Baltim). 24 (6) (1995) 546–552.

[20] R. Steele, An overview of the state of the art of automated capture of dietary intake information, Crit. Rev. Food Sci. Nutr. 55 (13) (2015) 1929–1938.

[21] T. Vu, F. Lin, N. Alshurafa, W. Xu, Wearable food intake monitoring technologies: A comprehensive review, Computers 6 (1) (2017) 4.

[22] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, S. Cagnoni, Automatic diet monitoring: a review of computer vision and wearable sensor-based methods, Int. J. Food Sci. Nutr. 68 (6) (2017) 656–670.

[23] K.L. Fuchs, M. Haldimann, D. Vuckovac, A. Ilic, Automation of data collection techniques for recording food intake: A review of publicly available and well-adopted diet apps, in: 9th International Conference on Information and Communication Technology Convergence: ICT Convergence Powered by Smart Intelligence, ICTC 2018, 2018.

[24] L.M. König, G. Sproesser, H.T. Schupp, B. Renner, Describing the process of adopting nutrition and fitness apps: Behavior stage model approach, J. Med. Internet Res. 20 (3) (2018).

[25] L.M. König, G. Sproesser, H.T. Schupp, B. Renner, Describing the process of adopting nutrition and fitness apps: Behavior stage model approach, J. Med. Internet Res. 20 (3) (2018).

[26] K. Fuchs, V. Huonder, D. Vuckovac, A. Ilic, Swiss FoodQuiz: Inducing nutritional knowledge via a visual learning based serious game, in: Proceedings of the 30th European Conference on Information Systems (ECIS), 2016, no. June, pp. 1–15..

[27] P. Pandey, A. Deepthi, B. Mandal, N.B. Puhan, Foodnet: Recognizing foods using ensemble of deep networks, IEEE Signal Process. Lett. (2017).

[28] O. Beijbom, N. Joshi, D. Morris, S. Saponas, S. Khullar, Menu-match: Restaurant-specific food logging from images, in: Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, 2015.

[29] A. Myers, et al., Im2Calories: Towards an automated mobile vision food diary, in: Proceedings of the IEEE International Conference on Computer Vision, 2015.

[30] Y. Yue, W. Jia, M. Sun, Measurement of food volume based on single 2-d image without conventional camera calibration, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2012.

[31] T. Stütz, R. Dinic, M. Domhardt, S. Ginzinger, Can mobile augmented reality systems assist in portion estimation? a user study, in: ISMAR 2014 - IEEE International Symposium on Mixed and Augmented Reality - Media, Arts, Social Science, Humanities and Design 2014, Proceedings, 2014.

[32] Y. Ando, T. Ege, J. Cho, K. Yanai, Depthcaloriecam: A mobile application for volume-based food calorie estimation using depth cameras, in: MADiMa 2019 - Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, Co-Located with MM 2019, 2019.

[33] N.A.M. Elsayed, B.H. Thomas, K. Marriott, J. Piantadosi, R.T. Smith, Situated analytics: Demonstrating immersive analytical tools with augmented reality, J. Vis. Lang. Comput. 36 (2016) 13–23.

[34] K. Fuchs, T. Grundmann, M. Haldimann, E. Fleisch, Impact of mixed reality food labels on product selection: Insights from a user study using headset-mediated food labels at a vending machine, in: MADiMa 2019 - Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, Co-Located with MM 2019, 2019, pp. 7–15.

[35] I. Nahum-shani, S.N. Smith, K. Witkiewitz, L.M. Collins, B. Spring, S.A. Murphy, Just-in-time adaptive interventions (JITAIs): An organizing framework for ongoing health behavior support, 2014.

[36] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, S. Cagnoni, Automatic diet monitoring: a review of computer vision and wearable sensor-based methods, Int. J. Food Sci. Nutr. (2017) 1–15.

[37] K. Fuchs, M. Haldimann, D. Vuckovac, A. Ilic, Automation of data collection techniques for recording food intake: a review of publicly available and well-adopted diet apps, in: 2018 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, South Korea, 2018, pp. 58–65.

[38] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1–9.

[39] Z.-Q. Zhao, P. Zheng, S. Xu, X. Wu, Object detection with deep learning: A review, 14 (8) 2018.

[40] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, Vol. 2017-Janua, 2017, pp. 2261–2269.

[41] S.A. Radzi, M. Khalil-Hani, Character recognition of license plate number using convolutional neural network, in: Lecture Notes in Computer Science, in: (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011.

[42] Y. Liu, H. Huang, Car plate character recognition using a convolutional neural network with shared hidden layers, in: Proceedings - 2015 Chinese Automation Congress, CAC 2015, 2016.

[43] A.G. Howard, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, ArXiv.

[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C.C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, Arxiv.Org.

[46] J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object detection via region-based fully convolutional networks, 2016.

[47] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

[48] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[49] A. Karpathy, T. Leung, Large-scale video classification with convolutional neural networks, in: CVPR 2014, 2014.

[50] P.J. Phillips, et al., Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms, Proc. Natl. Acad. Sci. (2018).

[51] J. Li, R. Guerrero, V. Pavlovic, Deep cooking: Predicting relative food ingredient amounts from images, in: MADiMa 2019 - Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, Co-Located with MM 2019, 2019.

[52] F. Kong, Automatic Food Intake Assessment using Camera Phones ProQuest Diss. Theses, 2012.

[53] W. Geng, et al., Fine-grained grocery product recognition by one-shot learning, 2 (2018) 1706–1714.

[54] A. Tonioni, L. Di Stefano, Domain invariant hierarchical embedding for grocery products recognition, Comput. Vis. Image Underst. (January) (2019).

[55] A. Tonioni, L. Di Stefano, A deep learning pipeline for product recognition on store shelves, 2019, arXiv.

[56] L. Karlinsky, J. Shtok, Y. Tzur, A. Tzadok, Fine-grained recognition of thousands of object categories with single-example training, in: CVPR, 2017.

[57] E. Goldman, et al., Precise detection in densely packed scenes, in: CVPR2019, 2019.

[58] K. Fuchs, T. Grundmann, E. Fleisch, Towards identification of packaged products via computer vision, in: The 9th International Conference on the Internet of Things (IoT 2019), 2019.

[59] T. Ege, K. Yanai, Multi-task learning of dish detection and calorie estimation, in: ACM International Conference Proceeding Series, 2018.

[60] T. Ege, K. Yanai, Simultaneous estimation of dish locations and calories with multi-task learning, IEICE Trans. Inf. Syst. (2019).

[61] S. Haussmann, et al., Foodkg: A semantics-driven knowledge graph for food recommendation, 2019.

[62] S. Gigandet, Openfoodfacts mobile applications, 2019, Openfoodfacts.com.

[63] A.L. Kor, Qualitative spatial reasoning for orientation relations in a 3-d context, in: Advances in Intelligent Systems and Computing, 2019.

[64] N.A.M. ElSayed, B.H. Thomas, K. Marriott, J. Piantadosi, R.T. Smith, Situated analytics: Demonstrating immersive analytical tools with augmented reality, J. Vis. Lang. Comput. 36 (2016) 13–23.

[65] R. Jerauld, Warable food nutrition feedback system, 2017, US009646511.

[66] G. Waltner, et al., MANGO - Mobile augmented reality with functional eating guidance and food awareness, in: ICIAP 2015 Work., Vol. 1, 2015, pp. 425–432.

[67] M. Domhardt, et al., Training of carbohydrate estimation for people with diabetes using mobile augmented reality, J. Diabetes Sci. Technol. (2015).

[68] J. Ahn, J. Williamson, M. Gartrell, R. Han, Q. Lv, S. Mishra, Supporting healthy grocery shopping via mobile augmented reality, ACM Trans. Multimedia Comput. Commun. Appl. (2015).

[69] S.G. Dacko, Enabling smart retail settings via mobile augmented reality shopping apps, Technol. Forecast. Soc. Change (2017).

[70] S.C. Isley, R. Ketcham, D.J. Arent, Using augmented reality to inform consumer choice and lower carbon footprints using augmented reality to inform consumer choice and lower carbon footprints, 2017.

[71] H. Jiang, J. Starkman, M. Liu, M.C. Huang, Food nutrition visualization on google glass: Design tradeoff and field evaluation, IEEE Consum. Electron. Mag. (2018).

[72] H. Chen, W. Jia, X. Sun, Z. Li, Y. Li, Saliency-aware food image segmentation for personal dietary assessment using a wearable computer, Meas. Sci. Technol. 025702 (2015) 25702.

[73] A. Sallè, M. Ryan, P. Ritz, Underreporting of food intake in obese diabetic and nondiabetic patients, Diabetes Care (2006).

[74] C. Julia, S. Hercberg, Nutri-score: Evidence of the effectiveness of the french front-of-pack nutrition label, Ernährungsumschau Int. 64 (12) (2017) 181–187.

[75] A.M. Buffington, Are vending machine selections healthier? trends in dietary quality of vending machine food and beverage selections among NHANES participants age 6-19 years between 2003-2012, Diss. Abstr. Int. Sect. B Sci. Eng. (2018).

[76] K. Fuchs, T. Barattin, M. Haldimann, A. Ilic, Towards tailoring digital food labels: Insights of a smart-RCT on user-specific interpretation of food composition data, in: MADiMa 2019 - Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, Co-Located with MM 2019, 2019.

[77] K. Fuchs, M. Zeltner, M. Haldimann, A. Ilic, Icon-based digital food allergen labels for complementation of text-based declaration, in: 28th European Conference on Information Systems (ECIS 2020), 2020.

[78] G.M. Singh, et al., Global, regional, and national consumption of sugar-sweetened beverages, fruit juices, and milk: A systematic assessment of beverage intake in 187 countries, PLoS One (2015).

[79] M. Luger, M. Lafontan, M. Bes-Rastrollo, E. Winzer, V. Yumuk, N. Farpour-Lambert, Sugar-sweetened beverages and weight gain in children and adults: A systematic review from 2013 to 2015 and a comparison with previous studies, Obes. Facts (2018).

[80] C. Julia, S. Hercberg, Development of a new front-of-pack nutrition label in France: the five-colour nutri-score, Public Health Panor. 3 (4) (2017) 712–725.

[81] T. Grundmann, K. Fuchs, Holoselecta dataset mendeley data, 2020, [Online]. Available: https://data.mendeley.com/datasets/wm83krbf32/1 [Accessed: 21-Jan-2020].

[82] S. Ren, K. He, R. Girshick, J. Sun, Faster r-CNN: Towards real-time object detection with region proposal networks IEEE trans, Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149.

[83] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: Proc. IEEE Int. Conf. Comput. Vis., Vol. 2017-Octob, 2017, pp. 2999–3007.

[84] S. Zhang, L. Wen, X. Bian, Z. Lei, S.Z. Li, Single-shot refinement neural network for object detection, 2017, pp. 4203–4212.

[85] O. Russakovsky, et al., Imagenet large scale visual recognition challenge, 2014, arXiv Artif. Intell..

[86] J. Huang, et al., Speed/accuracy trade-offs for modern convolutional object detectors, 2017, arXiv.

[87] T.-Y. Lin, et al., Microsoft COCO: Common objects in context, 2015, Arxiv. Org.

[88] World Health Organization, World health organization fact sheet: Cardiovascular diseases (CVDs), World Health Organ. (2017).

[89] N.A.M.M. Elsayed, B.H. Thomas, K. Marriott, J. Piantadosi, R.T. Smith, Situated analytics: Demonstrating immersive analytical tools with augmented reality, J. Vis. Lang. Comput. 36 (2016) 13–23.

[90] S. Knez, L. Šajn, Food object recognition using a mobile device: Evaluation of currently implemented systems, Trends Food Sci. Technol. (2020).

[91] E. Dunford, et al., Foodswitch: A mobile phone app to enable consumers to make healthier food choices and crowdsourcing of national food composition data, JMIR mHealth uHealth 2 (3) (2014) e37.

[92] A. Chamberlain, A. Crabtree, T. Rodden, M. Jones, Y. Rogers, Research in the wild: Understanding 'in the wild' approaches to design and development, in: Proceedings of the Designing Interactive Systems Conference, DIS '12, 2012.

[93] C. Pettitt, J. Liu, R.M. Kwasnicki, G.Z. Yang, T. Preston, G. Frost, A pilot study to determine whether using a lightweight, wearable micro-camera improves dietary assessment accuracy and offers information on macronutrients and eating rate, Br. J. Nutr. (2016).

[94] J.S. Lee, G. Mishra, K. Hayashi, E. Watanabe, K. Mori, K. Kawakubo, Combined eating behaviors and overweight: Eating quickly, late evening meals, and skipping breakfast, Eat. Behav. (2016).

[95] L. Yardley, T. Choudhury, K. Patrick, S. Michie, Current issues and future directions for research into digital behavior change interventions, Am. J. Prev. Med. 51 (5) (2016) 814–815.

[96] E. Robinson, C.A. Hardman, J.C.G. Halford, A. Jones, Eating under observation: A systematic review and meta-analysis of the effect that heightened awareness of observation has on laboratory measured energy intake, Am. J. Clin. Nutr. (2015).

[97] S.S. Holden, N. Zlatevska, C. Dubelaar, Whether smaller plates reduce consumption depends on who's serving and who's looking: A meta-analysis, J. Assoc. Consum. Res. (2016).

Klaus Fuchs is the Associate Research Director of the Auto-ID Labs ETH/HSG. He is earning his Ph.D. in Information Management from ETH Zurich in 2020. He holds a M.Sc. in Management, Technology and Economics from ETH Zurich.His research interests includedigital health, digital receipts and computer vision.

Mirella Haldimann Ph.D., is a Data Science and Analytics Consultant at D ONE Solutions. She earned her Ph.D. in Industrial Engineering from Linköping University in 2019. She also holds a M.Sc. in Technology Management from the Rotterdam School of Management (RSM). After she served as a postdoctoral researcher at the Management, Economics and Technology (D-MTEC) department of the Eidgenössische Technische Hochschule (ETH) Zurich. Her areas of research interests include business model innovation, technology management, and digital health.

Tobias Grundmann M.Sc., completed his master studies in Management, Technology and Economics at ETH Zurich. After his studies, he joined Holo One, where he implemented computer vision-based applications in the industrial context. His areas of research interests include computer vision and digital health.

Elgar Fleisch has a double appointment at ETH Zürich and University St. Gallen (HSG). At ETH, he is a full professor of information management, at HSG of technology management. The research activities of Elgar Fleisch and his team focus on applications, economic impacts, and infrastructures of mobile and ubiquitous computing.