Wearable System for Personalized and Privacy-preserving Egocentric Visual Context Detection using On-device Deep Learning

Mina Khan
MIT Media Lab
Cambridge, MA, USA
minakhan01@gmail.com

Glenn Fernandes
MIT Media Lab
Cambridge, MA, USA
glennfer@mit.edu

Akash Vaish*
akash9712@gmail.com
MIT Media Lab
Cambridge, MA, USA

Mayank Manuja*
mayankmanuja5@gmail.com
MIT Media Lab
Cambridge, MA, USA

Pattie Maes
MIT Media Lab
Cambridge, MA, USA
pattie@media.mit.edu

ABSTRACT

Wearable egocentric visual context detection raises privacy concerns and is rarely personalized or on-device. We created a wearable system, called PAL, with on-device deep learning so that the user images do not have to be sent to the cloud for processing, and can be processed on-device in a real-time, offline, and privacy-preserving manner. PAL enables human-in-the-loop context labeling using wearable audio input/output and a mobile/web application. PAL uses on-device deep learning models for object and face detection, low-shot custom face recognition (~1 training image per person), low-shot custom context recognition (e.g., brushing teeth, ~10 training images per context), and custom context clustering for active learning. We tested PAL with 4 participants, 2 days each, and obtained ~1000 in-the-wild images. The participants found PAL easy-to-use and each model had *gt*80% accuracy. Thus, PAL supports wearable, personalized, and privacy-preserving egocentric visual context detection using human-in-the-loop, low-shot, and on-device deep learning.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Ubiquitous and mobile computing**; **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Wearable, Egocentric Camera, On-device Deep Learning, Privacy-preserving, Computer Vision, Context-aware, Active Learning, Low-shot Learning, Custom-trainable, Personalized, Human-in-the-loop

*Both authors contributed equally to this research.

UMAP '21 Adjunct, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8367-7/21/06.
https://doi.org/10.1145/3450614.3461684

ACM Reference Format:
Mina Khan, Glenn Fernandes, Akash Vaish, Mayank Manuja, and Pattie Maes. 2021. Wearable System for Personalized and Privacy-preserving Egocentric Visual Context Detection using On-device Deep Learning. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3450614.3461684

1 INTRODUCTION

Context-awareness is key for human augmentation technologies [37], and egocentric visual contexts have been useful in intelligence augmentation applications [8, 44], especially in wearable settings [28, 41]. Deep learning models can recognize a wide range of visual contexts [42] but usually need a large number of training images and massive compute power using Graphical Processing Units (GPUs). However, visual context tracking, especially sending user data to the cloud, raises privacy concerns [13, 15] and the models are also not personalized for each user, especially using privacy-preserving and human-in-the-loop deep learning in wearable settings.

We created a wearable device, called PAL, for personalized and privacy-preserving egocentric visual context recognition using on-device, human-in-the-loop, and low-shot deep learning. PAL uses on-device deep learning for privacy-preserving visual context detection so that the user data is not sent to the cloud or another device for processing. PAL also supports user input and output for human-in-the-loop training and labeling of personalized visual contexts. We used on-device models for generic object and face detection, personalized low-shot custom face and custom recognition, and semi-supervised active learning-based custom context clusters. Compared to the state-of-the-art wearable systems for personalized visual context recognition, which use at least 100 training images [24] and do not use privacy-preserving on-device deep learning, PAL uses only ~10 training images per custom context and also uses active learning-based context clustering so that the users do not have to explicitly train different contexts.

We tested PAL's device in in-the-wild wearable settings with 4 participants, 2 days each, obtaining ~1000 images. Each of the models had an accuracy of over 80%. The participants felt comfortable wearing the device, found custom training and labeling intuitive and easy, and also did not have privacy concerns with the camera.

We make three contributions in this work: i. a wearable device for privacy-preserving and personalized visual context detection using on-device and human-in-the-loop deep learning; ii. low-shot, custom-trainable, and active learning models for recognizing custom contexts, faces, and context clusters; iii. real-world evaluations of the wearable system, including in-the-wild evaluations of visual context detection models.

2 RELATED WORK

PAL is a wearable system for personalized and privacy-preserving egocentric visual context detection using on-device, human-in-the-loop, and low-shot deep learning. While there are wearable systems for egocentric visual context tracking using deep learning, there are none that use on-device, human-in-the-loop, and low-shot deep learning like PAL. We divide our related work for visual context detection using deep learning into three categories.

2.1 Camera-based Deep Learning Systems

Deep learning can recognize a wide variety of visual contexts, e.g., objects and faces [42], and deep learning applications are now common in mobile environments [34]. There are deep learning-based applications using ambient cameras, e.g., for fall detection [9], activity recognition [46], and tracking museum visitors [30]. Wearable egocentric cameras have also used deep learning models, e.g., for predicting daily activities [3], visual assistance [31, 33], visual guides [40], and face recognition [5]. There are also non-egocentric wearable cameras using deep learning, e.g., for emotion recognition [43] and eating recognition [2]. However, none of these systems use on-device deep learning, especially for personalized and low-shot recognition of custom egocentric visual context.

2.2 Privacy-preserving and On-Device Learning

Research has identified several ethical concerns for wearable cameras [15], and users have also highlighted their needs for personal and bystander privacy [14]. There have been a couple of privacy-preserving approaches, e.g., privacy-preserving collaborative deep learning for human activity recognition [27] and image distort or modification [1, 7]. However, none of these systems use on-device deep learning to avoid sending data to the cloud for processing.

Deep learning commonly uses Graphical Processing Units (GPUs) but recently there have been on-device deep learning processors [45] and models [22]. On-device deep learning systems have also been used for computer vision [29], but they do not support personalized, low-shot, and human-in-the-loop visual context detection.

2.3 Personalized and Active Learning

Wearable deep learning-based egocentric visual context recognition has been used for personalized object and face recognition for memory augmentation [23, 24]. However, they do not use on-device deep learning for privacy-preserving context detection and also need ~100 images for training each personalized class, whereas our custom context and face recognition models use 1 to 10 images.

Low-shot deep learning models are common [35]. Active deep learning has also been used for activity recognition [11] and even combined with image clustering [4]. However, none of these models have been deployed in in-the-wild wearable settings, especially using on-device and human-in-the-loop deep learning.

3 DESIGN

We wanted to support wearable egocentric visual context detection. We had five key decision considerations: i. Capture the user's egocentric view but use a non-conspicuous camera position so the camera is not distracting to the wearer and people around them; ii. Enable real-time and ideally, offline detection so that the user does not have to be connected to the internet or another device for processing; iii. Enable seamless labeling of personalized contexts; iv. Enable low-shot detection of personalized contexts and efficient labeling of unrecognized contexts for active labeling; v. Enable on-device labeling so that the user data does not have to be sent to the cloud or another device for processing or to be saved there for long-term if the user does not want to. In line with our decision considerations, we made the following design decisions for camera placement, on-device deep learning, and user data labeling.

3.1 Camera Placement

Wearable cameras can be placed on the body or the face/head. On-body cameras do not always capture the same visual contexts as the user's eyes because they are distant from the user's eyes and also because the user's face may tilt or turn independent of their body. Head-mounted cameras are commonly used, e.g., GoPro and Google Glass, but they can be socially unacceptable [39] and even distracting and heavy. We decided to place a camera on the user's face to capture their visual context, but also decided to keep it small, light, and not too prominent or distracting on the user's face.

3.2 On-device Deep Learning

To preserve user privacy, we decided to use on-device deep learning for model training and inference so that user images are not sent to the cloud and can be deleted after on-device processing. The users can decide if they want their images stored on the cloud. If the user does not do real-time labeling of the custom context or if a specific context is not recognized, it is sent to the cloud so that the user can label it on the mobile/web app. The images are deleted right after the user labels them if the user has disabled long-term labeling. The users can also selectively delete specific images.

We decided to recognize a range of visual contexts, including common objects, generic faces, custom faces, and custom visual contexts, e.g., custom activities like playing pool. We decided to use low-shot human-in-the-loop learning to allow users to train custom faces and visual contexts. We also decided to use active learning and request the users to label unrecognized contexts. In order to enable efficient labeling of similar images, we decided to use context clustering so that users can label groups of images efficiently. We remove images of unknown faces to protect by-stander privacy.

3.3 User Input/Output and Data Labeling

We wanted to enable custom context and face recognition. We decided to include a button so that the users can start/stop custom training sessions. We enabled two ways of labeling: Real-time labeling right after recording a custom context. and post-hoc labeling, which can be done any time after the context was recorded.

Figure 1: Wearable device with on-device deep learning, camera, open-ear audio output, microphone, and button.

In order to enable real-time labeling, we included a microphone for the users to record their labels. We decided to use audio input because it can be quickly and less visually distracting than text input. We also included open-ear audio output to give feedback to the user, e.g., confirm starting or stopping training session and replay recorded custom labels. We chose open-ear audio output as it can be delivered seamlessly to the users without the users having to explicitly check text messages or without blocking the user's real-world content. For post-hoc labeling, we decided to create a mobile/web app for labeling both user-specified sessions as well as non-session data.

Labeling a custom face or session has five steps: 1. Single press button to start custom labeling session (camera takes periodic pictures) or double press for a custom face; 2. Move around to capture the necessary context if needed; 3. Press button to stop custom labeling session; 4. Open-ear audio prompt to record custom label name; 5a. Option # 1: Keep button pressed to record label name using audio input. Listen to the audio label recording and confirm or re-record the label; 5b. Option # 2: Double press button to defer labeling to labeling on mobile/web application.

4 IMPLEMENTATION

PAL has a wearable device for personalized and privacy-preserving egocentric visual context detection, and a mobile/web app for data labeling and visualization. We describe PAL's wearable device, on-device deep learning models, and the mobile/web app below.

4.1 Wearable Device

PAL's wearable device has an ear-hook and an on-body component (Figure 1). The ear-hook has a camera and speaker, and the on-body component has a Raspberry Pi Zero and Google Coral Deep Learning Accelerator. The wire connecting the ear-hook to the on-body component has a microphone and button for user input.

Every 2 minutes, the device takes a picture and runs all the deep learning models. The user starts and stops custom training sessions (6 images per minute) by pressing the button and labels the session using audio input/output or on the mobile/web application. The device also retrieves geolocations from the mobile phone for context clustering.

The wearable device consumes maximum 0.3A at 5.25V and our 2500mAh battery lasts ~5 hours. Our camera has a 64° wide view angle and captures ~70% of the user's visual context, missing the side 200 cm and top 100 cm of a 1200 cm x 750 cm view ~1m away.

4.2 On-device Deep Learning Models

PAL has on-device visual context detection models for three kinds of tasks: i. generic object and face detection; ii. low-shot custom face and context recognition; iii. context clustering for efficient active learning of visually-similar contexts. All models are trained and inferred on the wearable device. Each model runs in ~3s and all models in ~15s.

Object and Face Detection: The *Object Detection* model is trained on 90-item Common Objects in Context (COCO) dataset [25] and *Face Detection* on Open Images v4. Both models use the MobileNet SSD v2 architecture [12, 26].

Low-shot Custom Face and Context Recognition: Both models use MobileNet v1 architecture [12]. *Custom Face Recognition* models using FaceNet [38] and needs 1-2 custom training images per face. *Custom Context Recognition* model recognizes custom activities, e.g., brushing teeth, using weight imprinting [36]. Weight imprinting adds new classes to the existing list of classes for continual learning and needs ~10 training images for each class. The custom context recognition model is an Image Classification model pre-trained on the 1000-class ImageNet dataset [6].

Context Clustering for Active Learning: We cluster visually-similar contexts, separated by geolocations, using an image embedding generator combined with a clustering model. Our image embedding generator is the second-to-last layer of an Image Classification model (MobileNet v1), pre-trained on 1000-class ImageNet dataset. We use Density-based Spatial Clustering of Applications with Noise (DBSCAN) [10] as DBSCAN does not require us to fix the number of clusters. Unlabeled clusters are displayed on the web/mobile app for labeling.

4.3 Mobile/Web App

The mobile/web app enables post-hoc labeling of custom data, including custom sessions, custom faces, and unlabeled clusters. Also, users can visualize their data using a timeline or graphical view.

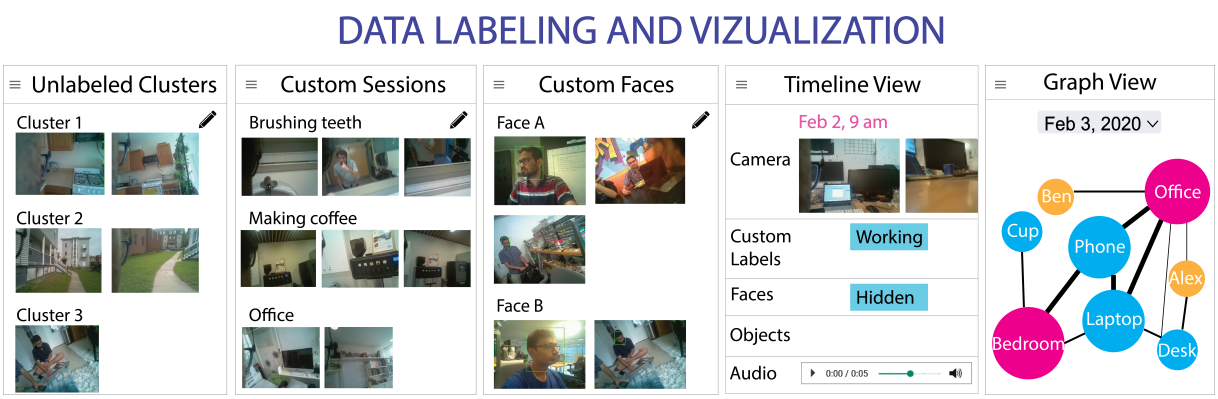

Figure 2: Interface for custom context labeling, custom face labeling, unlabeled cluster labeling, timeline data view, and graph data view.

Interface for custom context and face labeling, unlabeled cluster labeling, and timeline and graph data view is shown in Figure 2. We created our web/mobile app using Google's Polymer as it enables cross-platform deployment as a web application and a mobile application.

5 EVALUATION

We tested PAL in-the-wild with 4 participants (μ= 23.5 yrs, σ= 1.66 yrs; 3 males, 1 female; all students) wearing PAL for 2 days each (~5 hours per day based on PAL's maximum battery life). Below, we share the evaluations of our visual context detection models and also the responses from our open-ended participant interviews.

5.1 Visual Context Detection

There were a total of ~1000 images (1 image every 2 minutes) in 13 locations (9 indoors - 4 eateries, 2 shops, 1 dorm, 1 house, 1 office; 4 outdoors - 1 shopping area, 1 roadside walkway, 1 train station, 1 residential area). The details for each model are: *i. Object Detection* – 618 persons, 282 books, 48 TV screens, 45 laptops, 30 chairs, 25 bottles, 14 cars, 13 teddy bears, 8 keyboards, 7 microwaves, 7 cell phones, 6 potted plants, 5 couches, 4 bowls, 3 sandwiches, 3 trains, 2 clocks, 2 refrigerators, 2 sinks, 2 dining tables, 1 toilet, 1 umbrella, 1 bus, and 1 bicycle; *ii. Face Detection* – 180 faces; *iii. Custom Face Recognition* – 120 instances of 4 known people; *iv. Custom Context Recognition* – 7 custom activities (brushing teeth, making coffee, eating lunch, working in own office, working in open office area, playing pool, playing foosball), 10 training images each from user-initiated sessions (6 images/min), and ~350 total images; *v. Custom context clustering* – 19 indoor locations, ~300 images. We remove images of non-consenting participants to protect by-stander privacy.

Each model had over 80% accuracy (all results in Table 1). Even images partially occluded by the wearer's cheeks, eyeglasses, or hair were accurately predicted. Example images are in Figure 3.

5.2 User Experience

We conducted open-ended interviews with the participants about their experiences, which are summarized below.

Data Training, Labeling, and Visualization: The participants found the data visualizations "intuitive" and "helpful. The data training and labeling was "easy", using both audio input/output and the mobile app. Two participants mentioned that the audio output was audible in "even noisy environments", e.g., restaurants, bookstores, and train stations.

Device Comfort: The participants found the wearable device "quite comfortable" to wear and secure while walking. The participants requested a "lighter" clip-on as it "pulled on the ear-hook" and also "longer battery life".

Camera: None of the participants mentioned having major issues wearing the camera in public or privately. The participants took off the device whenever needed. Two participants mentioned liking the "small" and "invisible" camera.

5.3 Discussion

We created a wearable device, called PAL, for privacy-preserving and personalized egocentric visual context detection in wearable settings. We tested PAL in in-the-wild real-world settings with 4 participants, 2 days each. Our results show an accuracy of over 80% for for generic object and face detection, custom face and activity recognition, and custom indoor location clustering (total ~1000 images). The participants found the device comfortable to wear, the wearable audio input, audio output, and camera usable in public and private settings, and the wearable human-in-the-loop custom context training and labeling easy and intuitive. Thus, our initial tests show that PAL can support personalized and privacy-preserving egocentric visual context detection in wearable settings. We aim to further deploy PAL with more deep learning models for computer vision and test the models for more real-world human augmentation applications.

6 APPLICATIONS

We have used PAL to provide real-world habit habit-formation support in egocentric visual contexts [17, 20] as well as for language learning and memory support [19]. We also added other sensors to PAL's platform and open-sourced PAL as a modular platform for behavior change applications [16]. Users want behavior

Table 1: PAL's on-device models and their in-the-wild evaluations with 4 participants, 2 days each, and total ~1000 images

Models	Training	Testing
Object Detection[1]	90-item COCO [25]	98.8% accuracy, F1-score = 0.79, ~1000 instances
Face Detection[1]	Open Images v4 [21]	88.8% accuracy, F1-score = 0.9, ~180 instances
Custom Face Recognition [2,3]	1-2 custom images/person	86.9% accuracy, 4 known people, 120 instances
Custom Context Recognition[2,4,5]	10 custom images/session	87.2% accuracy, 7 custom activities, ~350 images
Custom Context Clustering[2,5,6]	N/A	82% accuracy, 19 indoor locations, ~300 images

[1]MobileNet SSD v2, [2]MobileNet v1 [12], [3]FaceNet [38], [4]Weight Imprinting [36], [5] Pre-trained on ImageNet

Figure 3: Example images from in-the-wild evaluations of PAL's on-device deep learning models (total ~1000 images).

change support and tracking in visual contexts [18], and we envision that PAL can be used for self-tracking and context-aware behavior change applications, e.g., Just-in-time Adaptive Interventions (JITAIs) [32]. PAL's egocentric visual context detection can also be further used for other applications, e.g., memory support for people with Alzheimer's or their caretakers.

7 CONCLUSION

Egocentric visual contexts can provide detailed contextual information, but wearable cameras are a privacy concern. Also, deep learning-based visual context detection models usually require large datasets and compute power. We created a wearable device, called PAL, for privacy-preserving and personalized egocentric visual context detection using on-device, low-shot, and human-in-the-loop deep learning. PAL's on-device deep learning enables user privacy as user data does not have to be sent to the cloud or another device for processing. We deployed models to not only detect generic faces and objects but also to recognize custom faces and contexts via low-shot learning and even create custom clusters for efficient active learning. We tested PAL in in-the-wild wearable settings, and not only did the visual context detection models perform well, but also the participants found it easy to use PAL for egocentric visual context detection. Thus, PAL uses on-device, low-shot, and human-in-the-loop deep learning for personalized and privacy-preserving egocentric visual context detection, and paves the way for several wearable context-aware applications using personalized egocentric visual contexts.

REFERENCES

[1] Rawan Alharbi, Mariam Tolba, Lucia C Petito, Josiah Hester, and Nabil Alshurafa. 2019. To Mask or Not to Mask? Balancing Privacy with Visual Confirmation Utility in Activity-Oriented Wearable Cameras. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 3 (2019), 1–29.

[2] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. 2020. FitByte: Automatic Diet Monitoring in Unconstrained Situations Using Multimodal Sensing on Eyeglasses. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[3] Daniel Castro, Steven Hickson, Vinay Bettadapura, Edison Thomaz, Gregory Abowd, Henrik Christensen, and Irfan Essa. 2015. Predicting daily activities from egocentric images using deep learning. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers (ISWC '15)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/2802083.2808398

[4] Luiz F. S. Coletta, Moacir Ponti, Eduardo R. Hruschka, Ayan Acharya, and Joydeep Ghosh. 2019. Combining clustering and active learning for the detection and learning of new image classes. *Neurocomputing* 358 (Sept. 2019), 150–165. https://doi.org/10.1016/j.neucom.2019.04.070

[5] Ovidiu Daescu, Hongyao Huang, and Maxwell Weinzierl. 2019. Deep learning based face recognition system with smart glasses. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '19)*. Association for Computing Machinery, New York, NY, USA, 218–226. https://doi.org/10.1145/3316782.3316795

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[7] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. 2018. Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–18.

[8] Aiden R. Doherty, Steve E. Hodges, Abby C. King, Alan F. Smeaton, Emma Berry, Chris JA Moulin, Siân Lindley, Paul Kelly, and Charlie Foster. 2013. Wearable cameras in health: the state of the art and future possibilities. *American journal of preventive medicine* 44, 3 (2013), 320–323. Publisher: Elsevier.

[9] Anastasios Doulamis and Nikolaos Doulamis. 2018. Adaptive Deep Learning for a Vision-based Fall Detection. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference (PETRA '18)*. Association for Computing Machinery, New York, NY, USA, 558–565. https://doi.org/10.1145/3197768.3201543

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

[11] H M Sajjad Hossain and Nirmalya Roy. 2019. Active Deep Learning for Activity Recognition with Context Aware Annotator Selection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. Association for Computing Machinery, New York, NY, USA, 1862–1870. https://doi.org/10.1145/3292500.3330688

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (April 2017). arXiv:1704.04861 [cs.CV]

[13] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. 2014. Privacy behaviors of lifeloggers using wearable cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 571–582.

[14] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. 2014. Privacy behaviors of lifeloggers using wearable cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 571–582.

[15] Paul Kelly, Simon J. Marshall, Hannah Badland, Jacqueline Kerr, Melody Oliver, Aiden R. Doherty, and Charlie Foster. 2013. An ethical framework for automated, wearable cameras in health behavior research. *American journal of preventive medicine* 44, 3 (2013), 314–319. Publisher: Elsevier.

[16] Mina Khan, Glenn Fernandes, and Pattie Maes. 2021. PAL: Privacy-preserving Audio, Visual, and Physiological Contexts for Wearable Context-aware Behavior Change Support. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

[17] Mina Khan, Glenn Fernandes, and Pattie Maes. 2021. PAL: Wearable and Personalized Habit-support Interventions in Egocentric Visual and Physiological Contexts. In *Proceedings of the Augmented Humans International Conference*.

[18] Mina Khan, Glenn Fernandes, and Pattie Maes. 2021. Users want Diverse, Multiple, and Personalized Behavior Change Support: Need-finding Survey. In *International Conference on Persuasive Technology*. Springer.

[19] Mina Khan, Glenn Fernandes, Utkarsh Sarawgi, Prudhvi Rampey, and Pattie Maes. 2019. PAL: A Wearable Platform for Real-time, Personalized and Context-Aware Health and Cognition Support. *arXiv preprint arXiv:1905.01352* (2019).

[20] Mina Khan, Glenn Fernandes, Akash Vaish, Mayank Manuja, Agnis Stibe, and Pattie Maes. 2021. Improving Context-aware Habit-support Interventions using Egocentric Visual Contexts. In *International Conference on Persuasive Technology*. Springer.

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982* (2018).

[22] Nicholas D Lane, Sourav Bhattacharya, Akhil Mathur, Petko Georgiev, Claudio Forlivesi, and Fahim Kawsar. 2017. Squeezing deep learning into mobile and embedded devices. *IEEE Pervasive Computing* 16, 3 (2017), 82–88.

[23] Hosub Lee, Cameron Upright, Steven Eliuk, and Alfred Kobsa. 2016. Personalized object recognition for augmenting human memory. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 1054–1061.

[24] Hosub Lee, Cameron Upright, Steven Eliuk, and Alfred Kobsa. 2018. Personalized Visual Recognition via Wearables: A First Step Toward Personal Perception Enhancement. In *Personal Assistants: Emerging Computational Technologies*. Springer, 95–112.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[27] Lingjuan Lyu, Xuanli He, Yee Wei Law, and Marimuthu Palaniswami. 2017. Privacy-Preserving Collaborative Deep Learning with Application to Human Activity Recognition. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1219–1228. https://doi.org/10.1145/3132847.3132990

[28] Steve Mann. 1997. Wearable computing: A first step toward personal imaging. *Computer* 30, 2 (1997), 25–32. Publisher: IEEE.

[29] Akhil Mathur, Nicholas D Lane, Sourav Bhattacharya, Aidan Boran, Claudio Forlivesi, and Fahim Kawsar. 2017. Deepeye: Resource efficient local execution of multiple deep vision models using wearable commodity hardware. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 68–81.

[30] Mauro Mezzini, Carla Limongelli, Giuseppe Sansonetti, and Carlo De Medio. 2020. Tracking Museum Visitors through Convolutional Object Detectors. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 352–355. https://doi.org/10.1145/3386392.3399282

[31] Davide Mulfari. 2018. A TensorFlow-based Assistive Technology System for Users with Visual Impairments. In *Proceedings of the Internet of Accessible Things (W4A '18)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3192714.3196314

[32] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Ann. Behav. Med.* 52, 6 (May 2018), 446–462.

[33] A Nishajith, J Nivedha, Shilpa S Nair, and J Mohammed Shaffi. 2018. Smart cap-wearable visual guidance system for blind. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 275–278.

[34] Kaoru Ota, Minh Son Dao, Vasileios Mezaris, and Francesco G. B. De Natale. 2017. Deep Learning for Mobile Multimedia: A Survey. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3s (June 2017), 34:1–34:22. https://doi.org/10.1145/3092831

[35] N O'Mahony, Sean Campbell, Anderson Carvalho, L Krpalkova, Gustavo Velasco Hernandez, Suman Harapanahalli, D Riordan, and J Walsh. 2019. One-Shot Learning for Custom Identification Tasks; A Review. *Procedia Manufacturing* 38 (2019), 186–193.

[36] Hang Qi, Matthew Brown, and David G Lowe. 2018. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5822–5830.

[37] Roope Raisamo, Ismo Rakkolainen, Päivi Majaranta, Katri Salminen, Jussi Rantala, and Ahmed Farooq. 2019. Human augmentation: Past, present and future. *International Journal of Human-Computer Studies* 131 (2019), 131–143. Publisher: Elsevier.

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[39] Dana Schuster. 2014. The revolt against Google 'Glassholes'. *New York Post* 14 (2014).

[40] Lorenzo Seidenari, Claudio Baecchi, Tiberio Uricchio, Andrea Ferracani, Marco Bertini, and Alberto Del Bimbo. 2017. Deep Artwork Detection and Retrieval for Automatic Context-Aware Audio Guides. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3s (June 2017), 35:1–35:21. https://doi.org/10.1145/3092832

[41] Thad Starner, Bernt Schiele, and Alex Pentland. 1998. Visual contextual awareness in wearable computing. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215)*. IEEE, 50–57.

[42] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* 2018 (2018).

[43] Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. 2020. EMO: real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services (MobiSys '20)*. Association for Computing Machinery, New York, NY, USA, 448–461. https://doi.org/10.1145/3386901.3388917

[44] Cassandra Xia and Pattie Maes. 2013. The design of artifacts for augmenting intellect. In *Proceedings of the 4th Augmented Human International Conference (AH '13)*. Association for Computing Machinery, New York, NY, USA, 154–161. https://doi.org/10.1145/2459236.2459263

[45] Hoi-Jun Yoo. 2020. Deep Learning Processors for On-Device Intelligence. In *Proceedings of the 2020 on Great Lakes Symposium on VLSI (GLSVLSI '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3386263.3409103

[46] Yanyi Zhang, Xinyu Li, Jianyu Zhang, Shuhong Chen, Moliang Zhou, Richard A. Farneth, Ivan Marsic, and Randall S. Burd. 2017. CAR - a deep learning structure for concurrent activity recognition: poster abstract. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN '17)*. Association for Computing Machinery, New York, NY, USA, 299–300. https://doi.org/10.1145/3055031.3055058