AI-Powered Smart Glasses for Sensing and Recognition of Human-Robot Walking Environments

Daniel Rossos, Alex Mihailidis, and Brokoslaw Laschowski

Abstract – **Environment sensing and recognition can allow human-robot systems to dynamically adapt to different walking terrains. However, fast and accurate visual perception is challenging, especially on embedded devices with limited computational resources. Here we developed a pair of AI-powered smart glasses for onboard real-time sensing and image classification. We used a Raspberry Pi Pico W microcontroller and an ArduCam HM0360 camera, both of which attach to the eyeglass frames using custom-designed 3D-printed mounts. We trained and optimized a lightweight and efficient convolutional neural network based on a MobileNetV1 backbone to classify the walking terrain as indoor surfaces, outdoor surfaces (grass and dirt), or outdoor surfaces (paved) using ~62,500 egocentric images from the Meta Ego4D dataset. We compiled and deployed our deep learning model using TensorFlow Lite Micro and post-training quantization to create a minimized byte array model of ~310,000 bytes while maintaining accuracy. Our system was able to predict complex walking environments with 93.6% classification accuracy with an inference time of 1.5 seconds. Our AI-powered smart glasses open new opportunities for real-world computer vision applications in human locomotion and robot control where inference on embedded devices and a low form factor (wearable integrated system) is required. Future research will focus on improving the onboard inference speed.**

I. Introduction

Visual sensing and recognition of human-robot walking environments is of growing interest. Applications of such research are widespread, ranging from the autonomous control and planning of robotic prosthetic legs and exoskeletons to interfacing with persons with visual impairments for sensory feedback. To date, AI-powered classification systems for visual perception in wearable robotics have been mainly limited to off-device inferencing using desktop computers and cloud computing. These designs lead to a reduction in mobility and a dependency on inconsistent wireless communication, which can have negative implications.

Previous research has developed head-mounted cameras with large computation systems such as Raspberry Pi 3 [1], [2], as well as chest-mounted cameras [3]-[5]. However, these designs did not integrate the vision and processing all within a single integrated system. Another key research area is deep learning, specifically convolutional neural networks (CNNs) for image recognition. The use of large models with many learnable parameters has been successful, with relatively high prediction accuracy for walking terrains, including level-ground, stairs, and other obstacles [1], [6]-[10]. Environment recognition systems have also used classical machine learning with success [11]-[13]. Both deep learning and non-deep learning models have shown high-performance but have not focused on deployment and efficiency as the main objectives. Consequently, these systems have been restricted to high-power computing and have limited deployment on mobile and embedded devices.

A fully-integrated visual perception system for environment recognition has yet to be designed, prototyped, and evaluated on edge devices with low inference speeds. This gap can be explained by limitations in mobile and embedded processing, which have only recently been alleviated by advances in hardware and deep learning model compression. Accordingly, here we developed a pair of AI-powered smart glasses with onboard sensing and image classification of real-world walking environments (Fig. 1). The mechatronics is integrated into a single device and is physically lightweight and a small form factor to not obstruct mobility or comfort. Computationally, it has sufficient memory and processing power for real-time inference with live video stream.

II. Methods

A. Mechatronics Design

The mechanical mounts for the smart glasses included two design considerations: 1) the location of the mechatronic components on the frames, and 2) the means by which these components are attached. The location of the microcontroller and

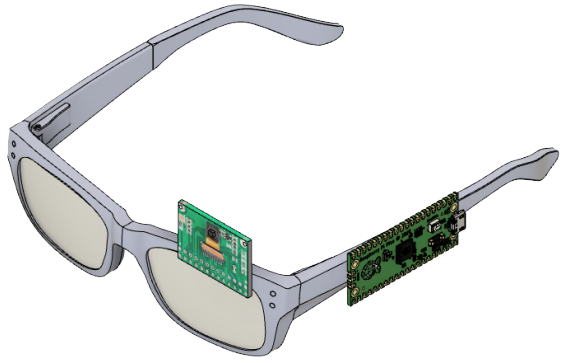

Fig. 1 Prototype design of our smart glasses, including the 3D-printed mechanical mounts used to interface the mechatronic components with the eyeglass frames.

*Research supported by the Schroeder Foundation and the AGE-WELL Networks of Centres of Excellence (NCE) program, Canada.

D. Rossos is with the Division of Engineering Science, University of Toronto, Toronto, Canada, and KITE Research Institute, Toronto Rehabilitation Institute, Toronto, Canada (e-mail: daniel.rossos@mail.utoronto.ca).

A. Mihailidis is with the Institute of Biomedical Engineering and the Robotics Institute, University of Toronto, Toronto, Canada, and the KITE Research Institute, Toronto Rehabilitation Institute, Toronto, Canada (e-mail: alex.mihailidis@utoronto.ca).

B. Laschowski is with the Department of Mechanical and Industrial Engineering and the Robotics Institute, University of Toronto, Toronto, Canada, and the KITE Research Institute, Toronto Rehabilitation Institute, Toronto, Canada (e-mail: brokoslaw.laschowski@utoronto.ca).

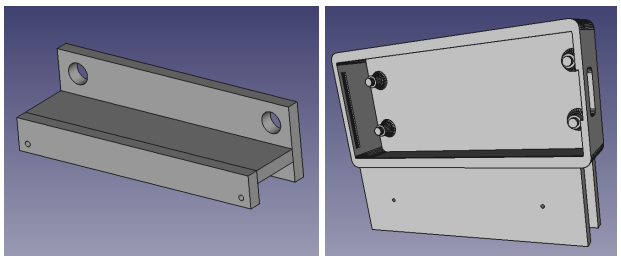

Fig. 2 The 3D-printed mounts for the camera (left) and microcontroller (right) with screw holes to secure the mechatronics to the eyeglass frames with a G-clamp-like system using rubber-tipped screws.

camera sensor was partially inspired by commercial smart glasses such as Google Glass [14] and Ray-Ban Stories [15], which have the vision sensing forward facing and the computational processing on the arms of the frames. Research has also used smart glasses [16]. This design allows for a larger processor to not obstruct the visual field-of-view while also having the camera simulate the orientation and perspective of the user - i.e., egocentric. We designed a semi-permanent mounting system that would allow our smart glasses to be applicable and transferable to a wide range of frames. This was achieved by custom-designing and 3D-printing mounting brackets for the camera and microcontroller (Fig. 2). The two main mechatronic components required to develop our system is the camera sensor to capture visual information about the walking environment, and the microcontroller for processing and computing the images (Fig. 3). With the low-power, low-latency constraints of our design, heightened scrutiny of relevant metrics and constraints are required to identify optimal components.

We used the ArduCam HM0360 VGA SPI camera due to its low power consumption, high frame rate, and high resolution [17]. The camera has a power consumption of less than 19.6mW during active VGA sampling. This low-power consumption supports the "always-on" camera design that our smart glasses aim for by ensuring that the power consumption would be minimally affected by continuous sensing. Another important feature of the camera sensor is the high frame rate. At 60 frames per second, this provides the microcontroller with a high enough sampling rate to ensure that there are no bottlenecks to the image classification resulting from the camera's framerate, while also providing updated real-world visual data, reducing lag in our smart glasses' understanding. The camera includes a relatively high resolution of 640x480 monochrome images. This resolution provides an image size large

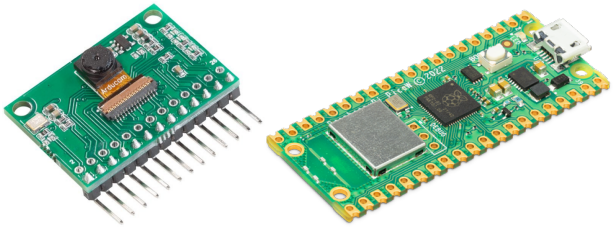

Fig. 3 The Arducam HM0360 VGA SPI camera (left) used for vision sensing and the Raspberry Pi Pico W microcontroller (right) used for the onboard image processing and computation.

Fig. 4 Example frames from the Ego4D dataset developed by Meta [18].

enough to portray environmental state information and even supports down-sampling of the resolution to allow for smaller deep learning models and faster inference predictions.

For the onboard computational processing, we used the Raspberry Pi Pico W microcontroller. This newly developed board increases memory and CPU power compared to smaller boards. The increase in processing power, larger memory, small form factor, and capability for wireless communication made this embedded microcontroller a viable solution for our smart glasses design. Compared to microcontrollers of comparable size, such as the Arduino Nano 33 BLE with a processing speed of 64 MHz, the Pico contains Dual ARM 133 MHz processors. This added processing power provides sufficient speed and parallelization to process live video streams while minimizing model inference speeds, with the goal of achieving real-time predictions using deep learning models.

The Pico contains 64 kB SRAM and 2 MB QSPI flash memory, which is greater than other microcontrollers such as the Arduino Nano 33 BLE. This increased memory is important for running learning algorithms directly on the embedded device and providing flexibility in the type of models that can be loaded, including more memory-intensive models such as deep convolutional neural networks. The Pico also has a small form factor of 21mm x 51.3 mm, which is essential for our design to be easily integrated into eyeglasses and provide minimal obstruction to mobility or comfort. The Pico can wirelessly communicate and interface with external robotic devices and computers via a CYW43439 chip, which supports single-band 2.4 GHz Wi-Fi connection and Bluetooth 5.2.

B. Computer Vision and Deep Learning

We created a new image dataset based on the Ego4D dataset developed by Meta [18]. The full Ego4D dataset includes more than 3,670 hours of egocentric (first-person) video collected by 923 subjects from 74 locations worldwide (Fig. 4). Images were collected using head-mounted wearable cameras, which made the dataset highly applicable to our computer vi-

Table 1. Breakdown of the class distributions in our new image dataset adapted from the Meta Ego4D dataset [18].

Environment class	Total frames	Number of videos	Frames per video
Indoor surfaces	24,189	51	474
Outdoor surfaces (paved)	21,362	51	418
Outdoor surfaces (grass and dirt)	16,958	31	547

sion application. In addition to an appropriate camera angle, the video clips were pre-labelled to identify the scene in which the videos were recorded.

Building on the annotations by Meta, a subsection of the Ego4D dataset applied to our task. We created new labels for 1) indoor surfaces, 2) outdoor surfaces (grass and dirt), and 3) outdoor surfaces (paved). See Table 1 for the class distributions. Using the videos, we sampled at one frame per second to collect images. To reduce the required memory storage to run our deep learning model, images in our dataset were downsampled to 96x96 pixels before being used for training, therein minimizing the staging area requirements for the microcontroller. This down sampling is a constraint imposed by the computing hardware. To help reduce overfitting during training, we added random horizontal reflections, image zooms, slight rotations, and contrast changes. These augmentations were deemed appropriate since such effects would likely occur in real-world walking while wearing the glasses. Finally, all image frames were converted to grayscale to mimic the conditions of the camera sensor (Fig. 5).

Fig. 5 Example frames of each environment class, including indoor surfaces (top row), outdoor surfaces - grass and dirt (middle row), and outdoor surfaces - paved (bottom row).

For classification and automatic feature engineering, we used the baseline model of MobileNetV1 in TensorFlow with an alpha value of 0.25, thereby reducing the network's width and learnable parameters to lower the computational demand on the embedded hardware. An additional 2D convolutional layer was added before the MobileNet base model to expand the input dimensions of the grayscale images to a 3-channel image as required by the MobileNet layer. The MobileNet layer is followed by a 2D global average pooling layer to reduce the dimensionality of the 2D output, followed by a fully connected layer with a softmax activation to predict the three environment classes (Table 2). The MobileNet architecture was selected as the underlying model similar to [8] because the depth-wise separable convolutional layers aid in efficient and accurate image classification. The model contained ~219,300 parameters and was trained using TensorFlow on Google Colab. The dataset was split into training (70%), validation (15%), and test (15%) sets. To avoid data leakage between test and validation sets, the source videos for frames within the training and validation sets were different.

Finally, we converted the deep learning model to a TensorFlow Lite model using a quantization method converting the floating-point numbers to 8-bit integers and resolving incompatible tensor operations. The TensorFlow Lite model was then converted using the TensorFlow Lite Micro tooling to produce a byte array usable by the microcontroller for onboard inference. Our final model was a size of 0.31MB. To quantify inference speed, the inference loop takes the most recent image from the camera and loads it to the microcontroller's memory. The image is then loaded into memory as input to the model, and the resulting label for that frame is derived.

III. Results

Our deep learning model achieved a training accuracy of 97.7%, training loss of 0.07, validation accuracy of 93.2%, and validation loss of 0.41 (Fig. 6). During inference on the test set, the embedded model achieved an overall prediction accuracy of 93.6%, f1-score of 93.6%, precision of 93.7%, and recall of 93.6%. The multiclass confusion matrix in Table 3 shows the distribution of the prediction accuracies for each walking environment. The neural network most accurately predicted outdoor surfaces - grass and dirt (96.8%), followed by outdoor surfaces - paved (94.7%) and indoor surfaces (90%). The total onboard inference loop time is 1.5 seconds from reading the image to outputting the label.

IV. Discussion

Here we present a preliminary development of AI-powered smart glasses with onboard visual sensing and recognition of

Table 2. The lightweight and efficient deep learning model used for image classification of walking environments.

Network layer	Output shape	Number of parameters
Convolutional 2D	1x96x96x3	30
MobileNet_V1_0.25	1x3x3x256	218,544
Global average pooling	1x256	0
Fully connected	1x3	771

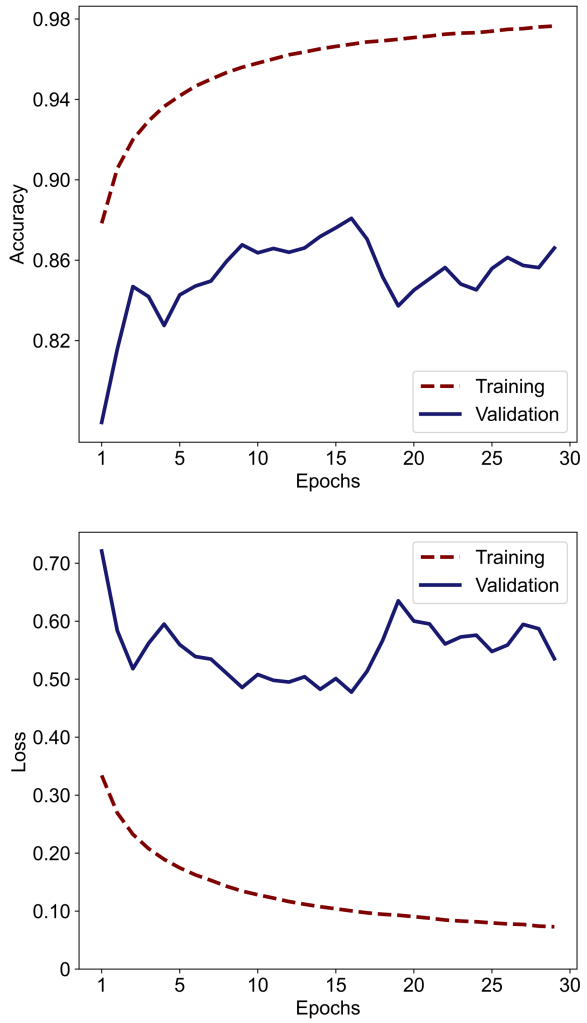

Fig. 6 The loss and accuracy curves on the training (maroon) and validation (blue) sets.

walking environments, which can wirelessly interface with humans and/or robots for locomotion. We designed the system to attach directly to eyeglass frames with a Raspberry Pi Pico W microcontroller and an ArduCam HM0360 camera sensor. We then developed a preliminary environmental recognition model to evaluate the computational feasibility of our fully-integrated system. Powered by deep learning, our model was trained and evaluated using a custom-built dataset with ~62,500 egocentric images of real-world walking environments adapted from the open-source Ego4D dataset by Meta [18]. Our system was able to predict different complex walking terrains in real-time (1.5 seconds) with an overall accuracy of 93.6%. These results demonstrate the potential to develop fast yet accurate visual perception systems deployed on embedded devices.

Compared to previous studies that have been limited by computational power and historically unable to support on-device inference due to the high memory demands of machine learning, our system leverages advances in microcontroller hardware, efficient neural network architectures, and compression algorithms to develop a fully-integrated system. As a result, our smart glasses can process and classify images of the walking environment in real-time without a dependency on inconsistent wireless communication to desktop computers or cloud computing.

Furthermore, compared to other leg and chest-mounted systems [8]-[10], our smart glasses offer several benefits due to its human-centered design. Our system was trained and evaluated using egocentric images, also known as first-person vision. This point-of-view closely mimics the biological vision system and takes into consideration the orientation of the user's head, which have practical implications for inferring locomotor intent. Additionally, our smart glasses do not have explicit requirements for the pose or viewing angles compared to other systems [19] that have relied on manual heuristics and meticulous rule-based thresholds for both the users and environments.

Nevertheless, our design has some limitations. The inference speed of our perception system is still somewhat high, which could impede real-time control. This would be especially important for rehabilitation robotics to dynamically adapt to different walking environments. Another means of future development would be to increase the number of environmental states in our classification model. Our model was developed to assess the feasibility of a fully-integrated smart glasses system using a high-performance deep learning model. However, increasing the number of environmental states would allow our system to be more applicable to a wider range of applications such as autonomous driving with powered wheelchairs or providing sensory feedback to persons with visual impairments. Overall, our AI-powered smart glasses open new opportunities for computer vision applications in human locomotion and robot control where inference on embedded devices and a low form factor (wearable integrated system) is required.

Table 3. The multiclass confusion matrix showing image classification accuracies (%) during inference on the test set. The columns and rows are the predicted and labelled classes, respectively. The classes include indoor surfaces (IS), outdoor surfaces - grass and dirt (OS-GD), and outdoor surfaces - paved (OS-P).

	IS	OS-GD	OS-P
IS	90.0	4.7	5.4
OS-GD	0.7	96.8	2.6
OS-P	2.8	2.5	94.7

ACKNOWLEDGMENT

We want to thank members of our Bionics Lab, part of the Artificial Intelligence and Robotics in Rehabilitation Team at the KITE Research Institute, Toronto Rehabilitation Institute, for their support, especially A. Garrett Kurbis for his support with the TensorFlow Lite and Hannah Smegal for her support with the 3D-printing. This research is dedicated to the people of Ukraine in response to the 2022 Russian invasion.

REFERENCES

[1] L. Novo-Torres, J.-P. Ramirez-Paredes and D. J. Villarreal, "Obstacle recognition using computer vision and convolutional neural networks for powered prosthetic leg applications", *in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3360–3363, Jul. 2019.

[2] R. L. da Silva, N. Starliper, B. Zhong, H. H. Huang, and E. Lobaton, "Evaluation of embedded platforms for lower limb prosthesis with visual sensing capabilities." *arXiv*, Jun. 26, 2020.

[3] A. H. A. Al-Dabbagh and R. Ronsse, "Depth vision-based terrain detection algorithm during human locomotion," *IEEE Trans. Med. Robot. Bionics*, vol. 4, no. 4, pp. 1010–1021, Nov. 2022.

[4] K. Karacan, J. T. Meyer, H. I. Bozma, R. Gassert, and E. Samur, "An environment recognition and parameterization system for shared-control of a powered lower-limb exoskeleton," *in 2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, pp. 623–628, Nov. 2020.

[5] G. Khademi and D. Simon, "Convolutional neural networks for environmentally aware locomotion mode recognition of lower-limb amputees," *in 2019 ASME Dynamic Systems and Control Conference*, Nov. 2019.

[6] N. E. Krausz and L. J. Hargrove, "Recognition of ascending stairs from 2D images for control of powered lower limb prostheses," *in 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 615–618, Apr. 2015.

[7] V. Rai, D. Boe, and E. Rombokas, "Vision for prosthesis control using unsupervised labeling of training data," *in 2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pp. 326–333, Jul. 2021.

[8] A. G. Kurbis, B. Laschowski, and A. Mihailidis, "Stair recognition for robotic exoskeleton control using computer vision and deep learning," *in 2022 International Conference on Rehabilitation Robotics (ICORR)*, pp. 1–6, Jul. 2022.

[9] A. G. Kurbis, A. Mihailidis, and B. Laschowski, "Development and mobile deployment of a stair recognition system for human-robot locomotion." *bioRxiv*, Apr. 28, 2023.

[10] D. Kuzmenko, O. Tsepa, A. G. Kurbis, A. Mihailidis, and B. Laschowski, "Efficient visual perception of human-robot walking environments using semi-supervised learning." *bioRxiv*, Jun. 29, 2023.

[11] B. Pan et al., "COPILOT: Human-environment collision prediction and localization from egocentric videos." *arXiv*, Mar. 26, 2023.

[12] A. Sharma and E. Rombokas, "Improving IMU-based prediction of lower limb kinematics in natural environments using egocentric optical flow," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 699–708, 2022.

[13] E. Tricomi et al., "Environment-based assistance modulation for a hip exosuit via computer vision," *IEEE Robot. Autom. Lett.*, vol. 8, no. 5, pp. 2550–2557, May 2023.

[14] "Google Glass Teardown." http://www.catwig.com/google-glass-teardown/.

[15] "Discover Ray-Ban Stories Features." https://www.ray-ban.com/canada/en/discover-rayban-stories/clp.

[16] O. Tsepa, R. Burakov, B. Laschowski, and A. Mihailidis, "Continuous prediction of leg kinematics during walking using inertial sensors, smart glasses, and embedded computing." *bioRxiv*, Feb. 13, 2023.

[17] "Arducam HM0360 VGA SPI Camera Module for Raspberry Pi Pico," https://www.arducam.com/product/arducam-hm0360-vga-spi-camera-module-for-raspberry-pi-pico-2/.

[18] K. Grauman et al., "Ego4D: Around the world in 3,000 hours of egocentric video," *in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990, Jun. 2022.

[19] N. E. Krausz, T. Lenzi, and L. J. Hargrove, "Depth sensing for improved control of lower limb prostheses," *IEEE Trans. Biomed. Eng.*, vol. 62, no.11, pp. 2576–2587, Nov. 2015.