Intelligent Service Robotics
https://doi.org/10.1007/s11370-024-00557-y

ORIGINAL RESEARCH PAPER

Vision-based human–machine interface for a robotic exoskeleton glove designed for patients with brachial plexus injuries

Yunfei Guo[1] · Wenda Xu[2] · Pinhas Ben-Tzvi[1,2]

Received: 24 November 2023 / Accepted: 2 August 2024

Abstract

This paper presents a novel vision-based human–machine interface (HMI) incorporated into an exoskeleton glove tailored for patients with brachial plexus injuries. Addressing the challenges posed by the loss of hand muscle control in individuals affected by these injuries, a fully automated exoskeleton glove function akin to a robotic gripper is used to prevent muscle atrophy through targeted hand muscle exercises. The proposed vision-based HMI is designed for a fully automated exoskeleton glove and incorporates computer vision techniques for the automatic identification of the target object, estimating its material and size, allowing the precise application of the required force to the target object. This novel approach enables users to efficiently grasp unknown objects with a significantly reduced failure rate. The vision-based method exhibits a grasp success rate of 87.5%, surpassing the baseline slip-grasp method's 71.9%. These results underscore the effectiveness of our vision-based HMI in enhancing the grasp functionality of the exoskeleton glove.

Keywords Human machine interface · Automated exoskeleton glove · Material classification

1 Introduction

Brachial plexus injuries (BPIs) are commonly caused by motorcycle or snowmobile accidents that damage the nerves of the arm and hand, leading to a loss of movement and sensation [1]. To restore the ability to grasp and perform everyday activities for those with BPI, automated exoskeleton gloves are used to prevent muscle atrophy due to lack of use [2–4]. As seen in Fig. 1A, BPI patients often suffer from muscle atrophy. This research is based on a previous research of an automated exoskeleton glove [2, 5] with a voice-based human–machine interface (HMI) [6, 7] that allows patients to grasp objects with the help of their healthy hand using voice-based HMIs. During the clinical experiment using the voice-based HMI (shown in Fig. 1), 60 grasp trials were conducted on two patients with brachial plexus injuries. Participants were instructed to perform 30 grasps on objects with known physical properties and 30 grasps on objects with unknown properties. The success rate of grasping known objects was 73.3%, whereas the success rate decreased to 56.7% for objects with unknown physical properties. This notable difference in success rates underscores the critical role that understanding physical properties plays in effective grasping. When the physical properties of an object are known, control programs can be tuned according to the necessary force required to secure the object. Conversely, the lack of information about an object's physical properties necessitates a trial-and-error control approach, which inherently carries a higher risk of failure. This finding aligns with previous research in robotic and human grasping, where knowledge of an object's weight, texture, and size significantly enhances the efficiency and effectiveness of grasping tasks [8, 9].

The motivation behind integrating a vision-based HMI alongside the current voice-based HMI is to devise a technique for estimating the physical attributes of target objects. This aims to improve the success rate of grasping, thereby enabling users to manipulate objects beyond those predefined. Unlike exoskeleton rehabilitation gloves that can use

✉ Pinhas Ben-Tzvi
bentzvi@vt.edu

Yunfei Guo
yunfei96@vt.edu

Wenda Xu
wenda@vt.edu

[1] Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, USA

[2] Department of Mechanical Engineering, Virginia Tech, Blacksburg, USA

Published online: 20 August 2024

electromyography (EMG) sensors to perform real-time force planning [10, 11], fully automated exoskeleton gloves are designed for patients with BPI who have little or no EMG signal from the hand and arm [1], making EMG-based approaches unsuitable. Electroencephalogram (EEG)-based HMIs can provide a force planning feature [12], but they suffer from the wearability issues of the EEG sensor [13, 14]. Voice-based HMIs are well developed and convenient to use, but not capable of grasp force estimation [6, 15, 16].

The integration of vision systems in robotic exoskeletons or grippers has been extensively explored by previous researchers, none of the existing vision-based HMI incorporates initial grasp force estimation for unknown objects using material detection. In their work, Kim et al. [17] designed an exoskeleton glove using a computer vision approach to real-time identification of the target object's location. Similarly, Pham et al. [18] proposed a vision-based method to infer grasp force on a robotic gripper, but this approach necessitates well-known physical properties of the target object. Calandra et al. [19] and Yamaguchi and Atkeson [20] introduced vision-based methods combined with tactile sensors to achieve a more stable grasp on a robotic gripper, with the vision systems primarily employed for motion and position tracking. Takamuku and Gomi [21] suggested that visual feedback of object motion could be utilized for the estimation of dynamic forces; but also assumes knowledge of the physical properties of the target object in advance.

This research introduces a vision-based method capable of conducting material detection on commonly used objects through transfer learning on a dataset designed for house interior material detection. Leveraging information about the identified material, the physical properties of the target object can be estimated and subsequently utilized to calculate the appropriate grasp force. The concept of the proposed vision-based HMI is based on the natural grasping process of humans. People can pick up and lift an item without being aware of its exact weight, material, or size. Research has demonstrated that even with limited haptic feedback, humans can still achieve a secure grip based on visual information [22].

The primary contributions of this study can be summarized as follows. Initially, we applied transfer learning to the state-of-the-art house interior surface material detection techniques, adapting them for the efficient identification of materials on common objects in constrained environments. Subsequently, we carried out system integration, pairing the vision-based system with a voice-based command system [6, 7] and a slip-grasp force control system [7, 23], enabling the cohesive functioning of the exoskeleton as a whole. Lastly, preliminary grasp experiments were conducted with a healthy subject to showcase the effectiveness of the vision-based HMI in estimating the initial grasp force. The results demonstrated an increased grasping success rate, exceeding that of the baseline slip-grasp method by 15.6%.

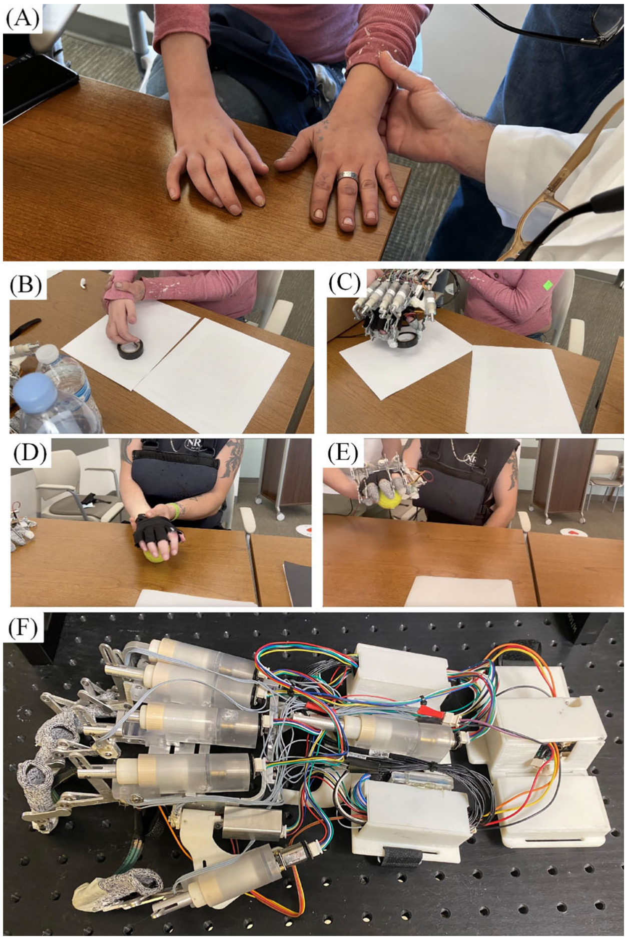

Fig. 1 Prior clinical experiments using a voice-based HMI experienced a high failure rate in grasping objects with unknown physical properties. **A** Subject with BPI on her right hand experience muscle atrophy due to lack of exercise. **B** The subject with BPI has no control over the muscles of her hand and arm. **C** The subject can grasp the target item with the aid of an exoskeleton glove. **D** The subject with BPI does not have control over his hand and arm muscles. **E** The subject can hold the target object with the assistance of an exoskeleton glove. **F** The exoskeleton glove used in the clinical experiment depicted in **B**–**E** is also employed in this research

2 Related work

2.1 Slip-grasp force planning method

Slip-grasp methods are commonly used to find the appropriate grasp force for unknown objects through trial and error. Lee et al. [24] proposed a slip detection method using a customized pressure sensor to measure slippage at the fingertips of the SAFER exoskeleton glove. A hybrid slip detection method for an exoskeleton glove was proposed by Guo et al. [7] and Xu et al. [23]. This method utilizes both serial elastic actuators (SEAs) and force sensitive resistors (FSRs) to enhance its accuracy. The force controller supplements force to the fingertips when the object slips. However, this reinforcement process often leads to a laborious grasping experience where users must persistently refine the optimal grasp force through repeated failures, rendering it impractical for exoskeleton glove users. Moreover, slip detection on a robotic exoskeleton glove differs from a robotic hand or gripper due to space and size limitations. Previous researchers have designed multiple slip detection sensors for robotic hands and grippers and have achieved good results with the slip-grasp force planning method [25, 26]. However, in an exoskeleton glove application, there is insufficient space to accommodate larger and more precise slip detection sensors at the fingertips. This sensor limitation results in accuracy issues for the slip-grasp method.

In this study, the slip-grasp method proposed by Guo et al. is incorporated into the system (illustrated in step 4 of Fig. 2. Additionally, the standalone slip-grasp force control, devoid of the estimated initial grasp force, is employed as the baseline to showcase the necessity of a vision-based estimation system.

2.2 Material recognition in the wild and MINC-2500 dataset

Surface material detection using computer vision can be used to estimate the weight and the surface friction coefficient of unknown objects. The state-of-the-art material recognition dataset is material in context (MINC). Bell et al. [27] built the MINC dataset with images of human-labeled material in the real world and proposed a deep learning-based material segmentation method. Bell et al. [27] used Grad-CAM [28] method to generate a probability map from trained convolution neural network and used the conditional random field (CRF) algorithm [29] to calculate a label for each pixel. The advantage of this method is that it does not require a pixel-wise label, which is ideal for applications with limited segmented data. MINC-2500 is a subset of the MINC dataset, which contains 57,500 image patches for 23 different types of materials. However, the MINC-2500 dataset mainly contains long-shot or extra-long-shot interior design images, which are taken from a distance and contain many different objects in context. This research focuses on detecting the surface material of objects in images that are taken in a close-up or medium-close-up view. Transfer learning was performed to transfer the learned weight from MINC-2500 to a collected close-view material dataset to improve the accuracy of material classification. The setup of the neural network and the experimental results are discussed in Sects. 5 and 8.4.

3 HMI system overview

The vision-based HMI is designed to grasp an object without the need for detailed measurements in advance. The goal is to find the initial grasp force by estimating the dimension, shape, weight, and surface material of the target object. The structure of the vision-based force planning system is shown in Fig. 2. Sample images for the exoskeleton grasping environment, object category, and object material are shown in Fig. 3.

The first step of the proposed system uses voice input from a microphone to perform user verification and grasp activation [30].

After receiving a grasp command, the camera embedded in the glasses will start to take pictures and perform the following four steps on the image to estimate the target object's physical properties.

(1) The input images are sent to an object detector. Object detection will help the vision-based force planning method to understand the environment by detecting all objects in the view.

(2) The target object can be extracted based on an ARUCO marker, which is a commonly used tool in single camera pose and position estimation application [31, 32]. The placement of the ARUCO marker is shown in Figs. 2 and 3. The target object category and size are acquired, and the grasp type is determined according to the target object's category.

(3) The size of the object is calculated based on the number of pixels.

(4) The surface material of the target object is acquired by analyzing the material of the target object. Given the object's size and surface material, the object's weight can be estimated.

In the third step, the initial grasp force is calculated based on the physical properties of the target object.

Lastly, the initial grasp force is send to a slip-grasp force control system for minor adjustments [30].

4 Object detection

The state-of-the-art object detection methods are based on single shot detector (SSD) [33], Faster R-CNN [34], EfficientDet [35], and YOLOV4 [36]. Researchers have previously tested these methods on the common objects in context (COCO) dataset [37]. The inference speed and mean average precision (mAP) at 50% intersection over union (IOU) of seven distinct object detection methods are assessed on a curated dataset comprising images from SVWSUN video glass. This dataset, termed the first-person view (FPV) grasp dataset throughout the paper, serves as the basis for selecting the optimal object detection method. Figure 3 provides sample images from the FPV grasp dataset. The experimental

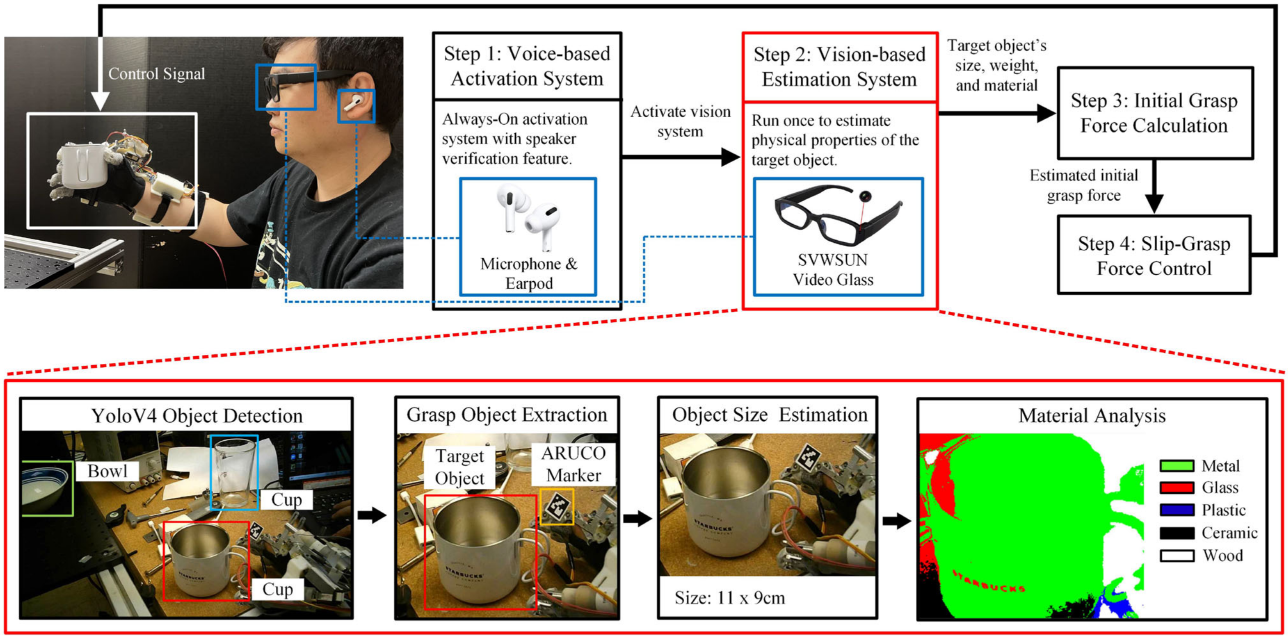
Control Signal
Step 1: Voice-based Activation System
Always-On activation system with speaker verification feature.
Microphone & Earpod
Activate vision system
Step 2: Vision-based Estimation System
Run once to estimate physical properties of the target object.
SVWSUN Video Glass
Target object's size, weight, and material
Step 3: Initial Grasp Force Calculation
Estimated initial grasp force
Step 4: Slip-Grasp Force Control
YoloV4 Object Detection
Bowl
Cup
Cup
Grasp Object Extraction
Target Object
ARUCO Marker
Object Size Estimation
Size: 11 x 9cm
Material Analysis
Metal
Glass
Plastic
Ceramic
Wood

Fig. 2 The system architecture utilized in this study comprises four main stages. Initially, there exists a voice-activated initiation system. Following this, a vision-driven estimation system is engaged, triggered by the activation. The third stage encompasses the initial calculation of the grasp force, drawing upon the inferred physical attributes from the previous step. Lastly, a slip-grasp force control mechanism is introduced to facilitate the generation of control signals and the refinement of force, all stemming from the initial grasp force calculation

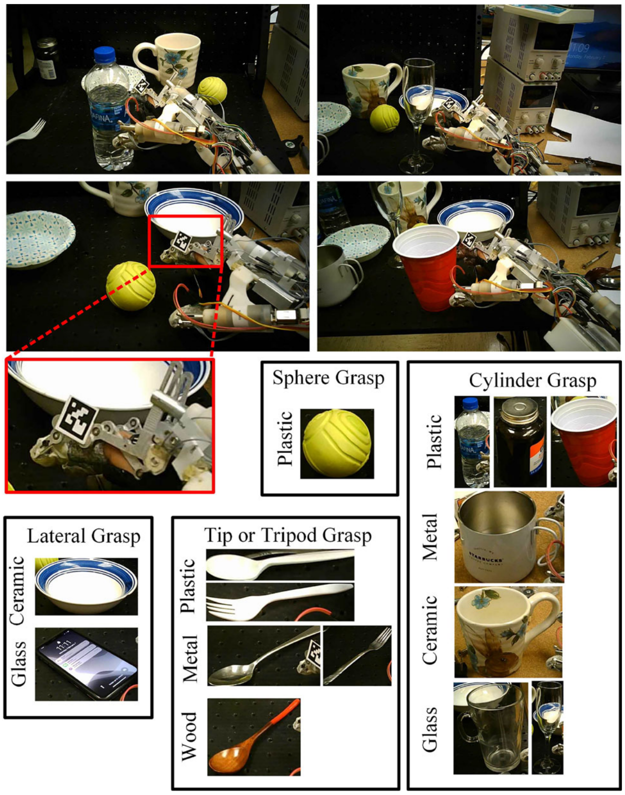

Fig. 3 Sample images for the exoskeleton grasping environment, object category, and object material

results are shown in Fig. 8. According to the experiments, YOLOV4 was selected as the object detection method used in this research; it better balanced speed and mAP than other methods.

4.1 Target object detection

The target object is determined by the outcome of the VOLOV4 algorithm and its spatial relationship to the ARUCO marker. The output from the YOLOV4 object detection algorithm is an object category vector **c**, an object bounding box vector **B**, and an object center vector **S**. The *n*th object detected in an image belongs to category ${}^{n}c$.

$$\mathbf{c} = [{}^{1}c, {}^{2}c, \ldots, {}^{n}c] \tag{1}$$

For the *n*th object detected in an image, the object's bounding box $\mathbf{{}^{n}b}$ is the combination of the upper left corner $\mathbf{{}^{n}p_{ul}} = ({}^{n}x_{\mathrm{ul}}, {}^{n}y_{\mathrm{ul}})$ and the lower right corner $\mathbf{{}^{n}p_{lr}} = ({}^{n}x_{\mathrm{lr}}, {}^{n}y_{\mathrm{lr}})$.

$$\begin{aligned}\mathbf{B} &= [\mathbf{{}^{1}b}, \mathbf{{}^{2}b}, \ldots, \mathbf{{}^{n}b}] \\ &= [({}^{1}x_{\mathrm{ul}}, {}^{1}y_{\mathrm{ul}}, {}^{1}x_{\mathrm{lr}}, {}^{1}y_{\mathrm{lr}}), \ldots, ({}^{n}x_{\mathrm{ul}}, {}^{n}y_{\mathrm{ul}}, {}^{n}x_{\mathrm{lr}}, {}^{n}y_{\mathrm{lr}})]\end{aligned} \tag{2}$$

For the *n*th object detected in an image, the center of the pixel of the detected object is located at $\mathbf{{}^{n}s}$ calculated from the bounding box $\mathbf{^{n}b}$.

$$\begin{aligned}\mathbf{S} &= [\mathbf{^{1}s}, \mathbf{^{2}s}, \dots, \mathbf{^{n}s}] \\ &= [(^{1}x_s, ^{1}y_s), \dots, (^{n}x_s, ^{n}y_s)]\end{aligned} \tag{3}$$

The target object is selected based on the distance to the ARUCO marker located on the exoskeleton glove. The output of the ARUCO application programming interface (API) contains the center coordinate of the marker: $\mathbf{s_m} = (x_m, y_m)$.

The exoskeleton glove used in this research is right-handed with the ARUCO marker placed on the index finger linkage (see Fig. 3). A weighted distance function was customized to find the distance between the ARUCO marker center coordinate $\mathbf{s_m}$ and the detected nth object center $\mathbf{^{n}s}$:

$$\begin{aligned}^{n}d &= w_0(x_m -^{n} x_s) + w_1(^{n}y_s - y_m) \\ &\quad + \sqrt{(x_m -^{n} x_s)^2 + (y_m -^{n} y_s)^2}\end{aligned} \tag{4}$$

where ^{n}d is the nth object distance between the object center and the ARUCO marker center. w_0 is the weight that serves as the penalty for the object located on the right of the marker, and w_1 is the weight that serves as the penalty for the object located above the marker. $(^{n}x_s, ^{n}y_s)$ is the coordinate of the center of the object from the vector of the center of the object $\mathbf{^{n}s}$. The grasped object's index i can be found by minimizing the customized distance function ^{n}d:

$$^{i}d = \min(^{1}d, ^{2}d, \dots, ^{n}d) \tag{5}$$

The category of the target object is ^{i}c, the bounding box is $\mathbf{^{i}b}$, and the center coordinate is $\mathbf{^{i}s}$.

4.2 Target object size estimation

The size of the target object is determined by pixel count relative to the ARUCO marker. To ensure accurate estimation, the user must align the ARUCO marker and the target object manually, maintaining a close to equal distance from the camera. This allows for estimating the target object's size from a 2D image perspective.

The detected object's bounding box $\mathbf{^{i}b}$ can be transferred from pixel coordinates to camera coordinates, and then to marker coordinates. The coordinates are explained in Fig. 4. The Euclidean distance between the points **e** and **f** in the marker coordinates is the length of the object (w) in centimeters (points are shown in Fig. 4). The Euclidean distance between points **f** and **g** in the marker coordinates is the height of the object (h) in centimeters. The size of the ARUCO marker is 2 cm in width and 2 cm in height. The size of the target object can be determined in the same coordinate system by analyzing the number of pixels relative to the ARUCO marker.

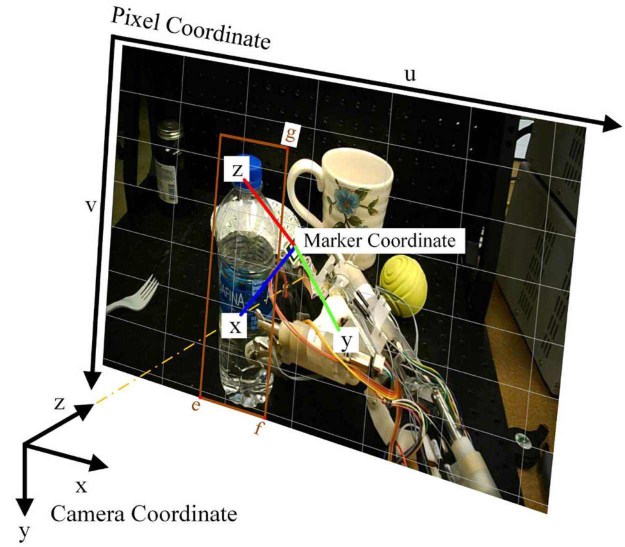

Fig. 4 Illustration of the camera, marker, and pixel coordinates

The following method can be used to convert points from pixel coordinates to marker coordinates. The ARUCO API outputs the rotation vector (**r**) in the axis-angle representation and the center coordinate (**t**) of the marker in the camera coordinates. To transfer a point $\mathbf{p_p} = (u, v)$ from the pixel coordinates to the camera coordinates $\mathbf{p_c} = (x_c, y_c, z_c)$, the following equations are used:

$$x_c = \frac{u - s_x}{f_x} d_z \tag{6}$$

$$y_c = \frac{v - s_y}{f_y} d_z \tag{7}$$

where d_z is the distance from the marker to the camera in the camera coordinates. s_x and s_y are the coordinates of the principle point in the camera coordinates. The s_x and s_y values utilized in this study are 640 and 360, measured from SVWSUN video glass. f_x and f_y are focal lengths of x and y axis in pixels. The f_x and f_y values utilized in this study are 1184 and 1249, measured from SVWSUN video glass.

To transfer a point $\mathbf{p_c} = (x_c, y_c, z_c)$ from the camera coordinate to the marker coordinate $\mathbf{p_m} = (x_m, y_m, z_m)$, the following equations are used:

$$\mathbf{R} = \text{Rodrigues}(\mathbf{r}) \tag{8}$$

$$\mathbf{p_m} = \mathbf{R}^T(\mathbf{p_c} - \mathbf{t}) \tag{9}$$

where Rodrigues formula was used to build a transformation matrix **R** from the axis-angle representation rotation vector **r**. **t** is the marker coordinate center represented in the camera coordinates.

5 Material classification

There are two common approaches to detect the surface material of an object, including image classification based on center pixels and semantic segmentation on the entire image [27, 38, 39]. The most widely used material classification datasets are the Flicker Material dataset (FMD), MINC, and Open Surface dataset. There are only limited pixel-wise annotated images provided, and most of these annotated images are furniture from the interior of a house, which is very different from this application. Due to the limited availability of annotated data, a pixel-wise supervised classification method such as UNet [39, 40] cannot be used. For this application, the center pixel classification method was used to classify the material of a given object image, and the conditional random field (CRF) [29] method was used for segmentation. Material segmentation is used to visualize the classification result.

Since this application focuses on grasping daily used objects as shown in Fig. 3, the number of classes in MINC-2500 was reduced from 23 to 5, which include ceramic, metal, glass, plastic, and wood.

5.1 Material classification challenges

Initially, the deep learning material classification method was trained and tested on MINC-2500 and achieved good accuracy. The original MINC dataset material patch classification was trained on VGG-16, AlexNet, and InceptionV1 in 2014. The VGG-16 architecture was used as a performance baseline to test the new networks, which achieved high classification accuracy in the ImageNet challenge: InceptionResNetV2 and ResNet152V2. Moreover, networks that achieve similar classification accuracy were tested, but have faster inference speeds: InceptionV3, ResNet50V2, and MobileNetV2. In addition to different network architectures, the NetVLAD pooling method was tested, which is a clustering-based pooling method commonly used in speaker verification, face detection, and place recognition [41].

The model's pre-trained weights are sourced from ImageNet, and training halts if the validation loss fails to decrease for ten consecutive epochs. To evaluate the training outcome, a custom dataset (referred to as the close-view material dataset in the rest of the paper) was employed, mirroring the application's use case. This dataset comprises images sourced from the FMD dataset and those gathered online. Figure 7 illustrates sample images from this dataset. The close-view material dataset encompasses 169 images for each of the five categories.

The training results and model performance comparison are shown in Table 1. According to training results, ResNet50V2, MobileNetV2, and InceptionV3 are the top 3 networks that achieve a good time and performance balance in the MINC-2500 validation set. However, the MINC-2500 does not have a perfect generalization to material classification. The context in the MINC dataset is very different from that of this application, which prevents the network from finding a correct label during testing on the close-view material dataset. NetVALD clustering pooling layer also does not improve accuracy. To solve the generalization issue, transfer learning was performed to retrain the model in the Close-view Material dataset. Transfers from ImageNet and MINC-2500 weight were experimented. The results are shown in Table 2. The results show that the transfer from MINC-2500 using ResNet50V2 has the best accuracy when testing on the Close-view Material dataset.

5.2 Proposed approach: transfer leaning using ResNet50V2

Based on the experimental results from the previous section, ResNet50V2 was used to transfer the weight from ImageNet to the MINC-2500 dataset. The number of material classes in MINC-2500 is reduced to metal, ceramic, plastic, glass, and wood. The input layer is modified to match the MINC-2500 size, the convolution blocks from ResNet50V2 have not been modified, and the weight is trained using the initial value from ImageNet. The output of the convolution layer consists of 2048 feature maps $\mathbf{M}_{[\mathbf{12}\times\mathbf{12}\times\mathbf{2048}]}$. The pooling layer uses global average pooling to group the feature maps $\mathbf{M}_{[\mathbf{12}\times\mathbf{12}\times\mathbf{2048}]}$ to $\mathbf{M}_{[\mathbf{1}\times\mathbf{1}\times\mathbf{2048}]}$ and classified into five classes multiplied by weight $\mathbf{W}_{[\mathbf{5}\times\mathbf{2048}]}$. Due to the low generalization accuracy of the MINC-2500 dataset, the MINC-2500 weight was transferred to the close-view material dataset using the same architecture. The training and inference procedure is shown in Fig. 5.

When inferring on a sample image, the ResNet50V2 network was modified to output a class probability map ${}_{\mathbf{c}}\mathbf{P}_{[\mathbf{1}\times\mathbf{5}]}$ and a feature-map-sized class probability map ${}_{\mathbf{f}}\mathbf{P}_{[\mathbf{12}\times\mathbf{12}\times\mathbf{5}]}$ using Grad-CAM [28]. The Grad-CAM is generated using the following equation:

$$ {}_{\mathbf{f}}\mathbf{P} = \sum_{n=1}^{2048} {}^{\mathbf{n}}\mathbf{W}\,{}^{\mathbf{n}}\mathbf{M} \tag{10} $$

Where, ${}^{\mathbf{n}}\mathbf{M}$ is the nth feature map and ${}^{\mathbf{n}}\mathbf{W}$ is the weight of the nth feature map. The probability map ${}_{\mathbf{f}}\mathbf{P}_{[\mathbf{12}\times\mathbf{12}\times\mathbf{5}]}$ will be resized to pixel level probability map ${}_{\mathbf{p}}\mathbf{P}_{[\mathbf{362}\times\mathbf{362}\times\mathbf{5}]}$ using cubic spline interpolation. The probability map ${}_{\mathbf{p}}\mathbf{P}_{[\mathbf{362}\times\mathbf{362}\times\mathbf{5}]}$ and colored image $\mathbf{I}_{[\mathbf{362}\times\mathbf{362}\times\mathbf{3}]}$ are input into a CRF algorithm to perform pixel level unsupervised segmentation by minimizing the following energy function [29]:

$$ {}^{c}E(\mathbf{x}) = \sum_{i} U(i) + \sum_{(i,j)} \mathrm{Par}(i,j) \tag{11} $$

Table 1 Results of training on MINC-2500 and testing on the close-view material dataset

Network	MINC-2500 accuracy (%)	Close-view material accuracy (%)	Speed (ms)*
VGG-16	71	22	279
InceptionV3	**83**	21	**215**
VGG-16-N*	68	21	292
InceptionV3-N*	77	22	225
MobileNetV2	75	20	**173**
ResNet50V2	78	**23**	228
ResNet152V2	84	22	487
InceptionResNetV2	81	21	472

[a]Speed*: the inference time is measured by inference of one image on a E5-1260 CPU
[b]-N*: NetVALD layer with 32 clusters is added after the last convolution layer
The selected networks and their performance are highlighted in bold, indicating the reason they were chosen for transfer learning

Table 2 Performance comparison between transfer ImageNet and MINC-2500 weight to the close-view material dataset

Network	Transfer MINC-2500 accuracy (%)	Transfer ImageNet accuracy (%)
ResNet50V2	**79**	76
MobileNetV2	72	71
InceptionV3	75	72

The selected network and its performance are highlighted in bold, indicating the reason it was chosen for this application

where ${}^{c}E(x)$ is the energy function for class c. $\mathbf{x}$ is the set of all pixels in image $\mathbf{I}$. i and j are pixel indexes in set $\mathbf{x}$. i and j control a nested loop to pair each pixel with all other pixels without repetition. $U(i)$ is the unary energy that is the negative log probability of a pixel belonging to class c. $\text{Par}(i, j)$ is the pairwise energy that measures the pixels' spatial and color similarity. The unary and pairwise energy is defined in the following equations:

$$U(i) = -\log({}^{\mathbf{i}}_{\mathbf{p}}\mathbf{P}_{\mathbf{c}}) \tag{12}$$

$$\text{Par}(i, j) = \exp\left(-\frac{|{}^{i}p - {}^{j}p|^2}{2s_p^2} - \frac{|{}^{\mathbf{i}}\mathbf{I} - {}^{\mathbf{j}}\mathbf{I}|^2}{2s_c^2}\right) \tag{13}$$

where ${}^{\mathbf{i}}_{\mathbf{p}}\mathbf{P}_{\mathbf{c}}$ is the pixel level probability of ith pixel in the image belonging to class c. ${}^{i}p$ and ${}^{j}p$ are the position of ith and jth pixels. ${}^{\mathbf{i}}\mathbf{I}$ and ${}^{\mathbf{j}}\mathbf{I}$ are the RGB values of ith and jth pixels. Long-range connections were used in the energy calculation. Thus, the pairwise energy contains only the appearance kernel. s_p and s_c are the position similarity and color similarity parameters, respectively. Parameter values s_p and s_c were chosen to be 60 and 10, respectively, based on Krähenbühl and Koltun. The results of the CRF algorithms will be an updated pixel level probability map ${}_{\mathbf{crf}}\mathbf{P}_{[\mathbf{362\times362\times5}]}$.

The classification results can be found by finding the maximum value of the ${}_{\mathbf{c}}\mathbf{P}$ class probability map. The results can be directly used to estimate the grasp force. The segmentation results can be used to perform pixel-wised classification when the target object contains different materials. The sample segmentation results and classification accuracy are available in Sect. 8.4.

6 Weight estimation

The estimated size and material of the target object can be obtained based on the methods described in the previous sections. However, the information is insufficient to estimate the weight, and some assumptions need to be made in order to calculate the volume of the target object.

The target object in this application can be classified into four different categories according to their required grasp type (shown in 3). The weight of an apple and a cell phone is not affected much based on size; thus, the average weight of an apple and a cell phone can be used as the weight of the target object. Sports balls are usually very light, so it was assumed that a sports ball weighs 20 g if it has a diameter less than 5 cm, weighs 100 g if it has a diameter between 5 and 10 cm and weighs 250 g if the diameter is larger than 10 cm.

The shape of a spoon or fork can be simplified to a plate with a thickness of 0.1 cm. Thus, the weight of a spoon or fork can be estimated using the following:

$$v_{sf} = 0.1wh \tag{14}$$

$$s_{sf} = v_{sf}\rho \tag{15}$$

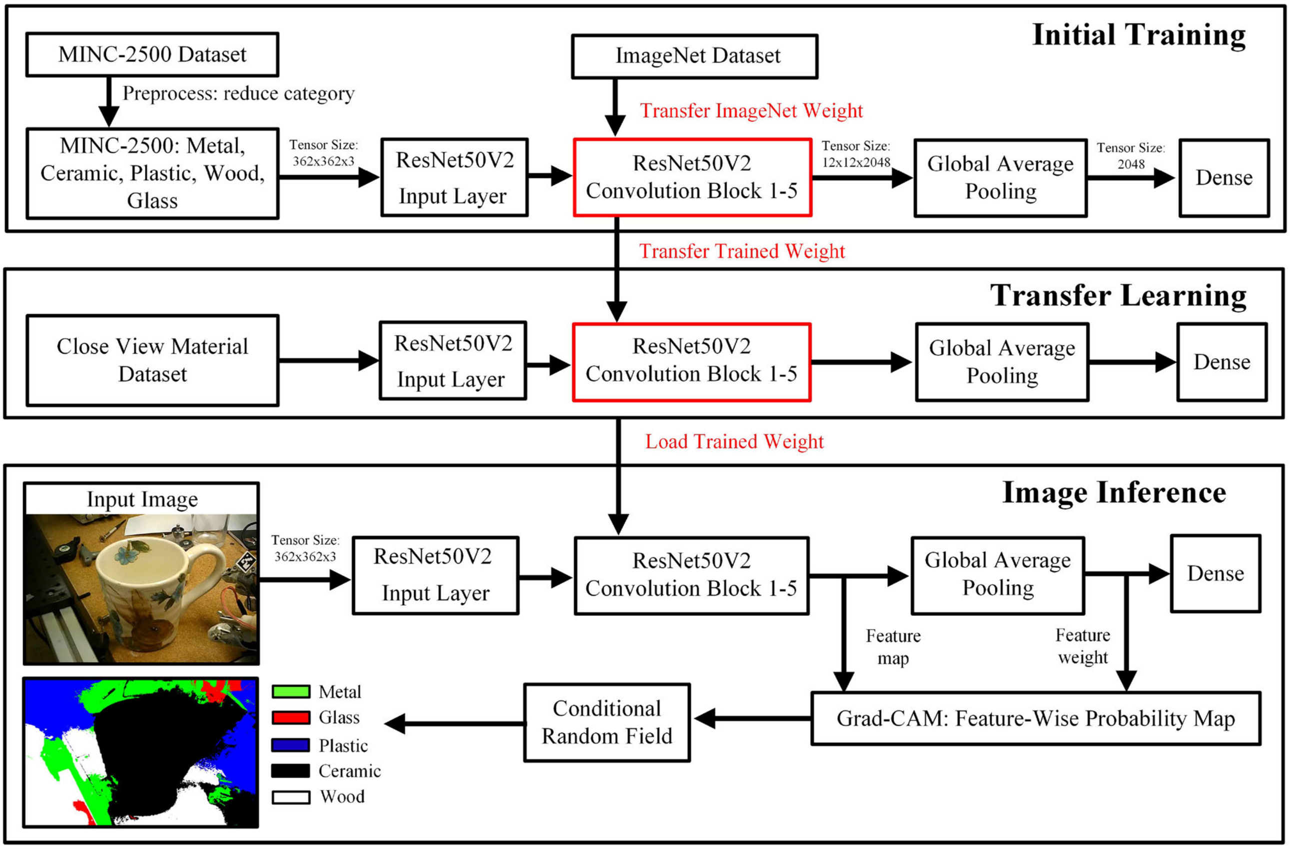

Fig. 5 Training and inference procedure for vision-based material classification and segmentation. Initially, training commences with the MINC-2500 dataset featuring reduced categories, utilizing ImageNet weights as the starting point. Subsequently, transfer learning is applied to the collected close-view material dataset, leveraging the previously trained weights as a foundation. During inference, the network produces two outputs: a probability identifying the primary material and a pixel-wise material probability map

where w and h are the estimated width and height of the target object, respectively. ρ is the density of the material of the target object. v_{sf} is the volume of the object. s_{sf} is the weight of the target spoon or fork.

The shape of a bottle, cup, and wine glass can be simplified to a hollow truncated cone. It is assumed that the truncated cone has $\frac{2}{3}$ of the volume of a cylinder of the same height. The thickness can be assumed to be 0.2 cm. Thus, the weight of a bottle when filled with water can be estimated using the following.

$$\begin{aligned} v_\mathrm{b} &= \frac{2}{3}(v_\mathrm{o} - v_t exti) \\ &= \frac{2}{3}\left(\pi(\frac{w}{2})^2 h - \pi\left(\frac{w}{2} - 0.2\right)^2 (h - 0.4)\right) \end{aligned} \tag{16}$$

$$s_\mathrm{b} = v_\mathrm{b}\rho + v_\mathrm{i}\rho_\mathrm{w} \tag{17}$$

where v_b is the volume of the material to form the bottle. v_o is the outer volume, v_i is the inner volume. s_b is the weight of the bottle. ρ is the density of the material of the bottle. ρ_w is the density of water.

The weight of a cup can be estimated similar to that of a bottle. The only difference is that a cup might have a handle and will make the volume calculation inaccurate. The size of the handle was assumed to be 30% of the weight of the cup w. Thus, the weight of a cup when full of water can be estimated using the following.

if $h \geq w$:

$$\begin{aligned} v_\mathrm{c} &= \frac{2}{3}(v_\mathrm{o} - v_\mathrm{i}) \\ &= \frac{2}{3}\left(\pi\left(\frac{w}{2}\right)^2 h - \pi\left(\frac{w}{2} - 0.2\right)^2 (h - 0.2)\right) \end{aligned} \tag{18}$$

if $w \geq h$:

$$\begin{aligned} v_\mathrm{c} &= \frac{2}{3}(v_\mathrm{o} - v_\mathrm{i}) \\ &= \frac{2}{3}\left(\pi\left(\frac{0.7w}{2}\right)^2 h - \pi\left(\frac{0.7w}{2} - 0.2\right)^2 (h - 0.2)\right) \end{aligned} \tag{19}$$

$$s_c = v_c\rho + v_i\rho_w \tag{20}$$

where v_c is the volume of material to form the cup. v_o is the outer volume, and v_i is the inner volume. s_c is the weight of the bottle. ρ is the density of the material of the cup. ρ_w is the density of water. Wine glass is a special cup with a long leg, so it was assumed that the capacity of the glass is 50% of a normal cup. Thus, the weight of a wine glass when full of water can be estimated using the expression:

$$s_{wg} = v_c\rho + 0.5v_i\rho_w \tag{21}$$

7 Initial grasp force calculation

The initial grasp force is calculated based on the predicted weight and the shape of the standard object. Figure 6 illustrates the coordinate systems for grasping force initialization. The origin of the world coordinates is placed at the center of the object. The coordinates for the exoskeleton glove are positioned at the center of the 9-DoF (Degree of Freedom) MPU-9250 inertia measurement unit (IMU). The IMU undergoes calibration using recursive least squares [42] for magnetometer calibration and extended Kalman filter [43] for sensor fusion to align with the world coordinates. Assuming that there is no torque applied on the object and the contact forces are normal to the last link of each of the exoskeleton fingers, for an arbitrary object, the force equilibrium equation can be expressed as:

$$\sum_i \mu\, {}^{\mathrm{w}}\mathbf{R}_{\mathrm{e}}\, {}^{\mathrm{e}}\mathbf{R}_i\, {}^{\mathrm{e}}\mathbf{F}_i + M\mathbf{g} = \mathbf{0} \tag{22}$$

where $i \in$ {thumb, index, middle, ring, little}, ${}^{\mathrm{w}}\mathbf{R}_{\mathrm{e}}$ is the rotation matrix from the exoskeleton glove coordinates to the world coordinates, which is calculated based on readings from the IMU. μ is the friction coefficient, which is estimated based on the surface material. ${}^{\mathrm{e}}\mathbf{R}_i$ is the rotation matrix from the fingertip i to the exoskeleton glove coordinates, which is calculated based on the forward kinematics of the glove [2]. ${}^{\mathrm{e}}\mathbf{F}_i$ is the vector of the contact force applied on fingertip i, which is measured based on a calibrated SEA [44]. M is the mass of the object, and $\mathbf{g}$ is the vector of gravitational acceleration.

For the cylinder grasp and the tip grasp, the direction of the friction force on each fingertip is always opposite to gravity. Therefore, the above equation can be simplified to $\sum_i \mu F_i = Mg$.

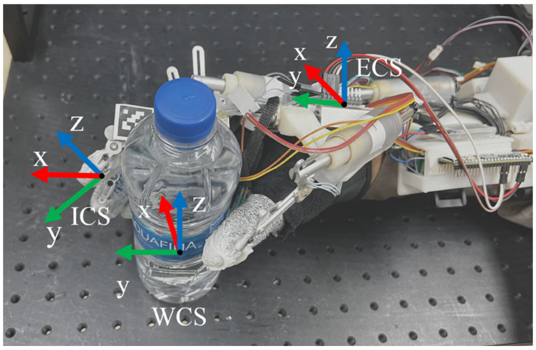

Fig. 6 The coordinate systems for initial force estimation. WCS, world coordinate system; ECS, exoskeleton glove coordinate system; ICS, i-th fingertip coordinate system

8 Experimental results

The experiment section encompassed three primary components. Initially, the collected datasets used in this research were introduced. Subsequently, the performance of object detection, size estimation, and material classification was assessed within these datasets. Finally, vision-based HMI was incorporated into the slip-grasp force planning approach and coupled with a voice-activated system. The experiments were structured to contrast the combined approach of vision and the slip-grasp method against the exclusive use of the slip-grasp force planning method. The experimental procedure involving human subjects received approval from the Carilion Clinic Institutional Review Board (IRB-19-330).

8.1 Datasets

This application utilized two collected datasets: the FPV Grasp dataset and the close-view material dataset.

The FPV Grasp dataset is used for validating the vision-based grasp force planning method and comprises 30 images captured from a 1080P SVWSUN video glass worn by a user of an exoskeleton glove. The Video Glass undergoes calibration through a standard chess board calibration method to rectify lens distortion. Each grasp object is annotated with a bounding box. Sample images are depicted in Fig. 3.

The close-view material dataset is designed to enhance the material classification performance with close-view context. The dataset encompasses five labels: ceramic, plastic, metal, wood, and glass. Each class includes a training set of 119 images, a testing set of 30 images, and a validation set of 20 images. The material of the object's center serves as the basis for labeling each image. This dataset amalgamates images sourced from online searches, the FMD dataset, and images captured specifically for this research on grasp objects. Sample images are illustrated in Fig. 7. Notably, the images in

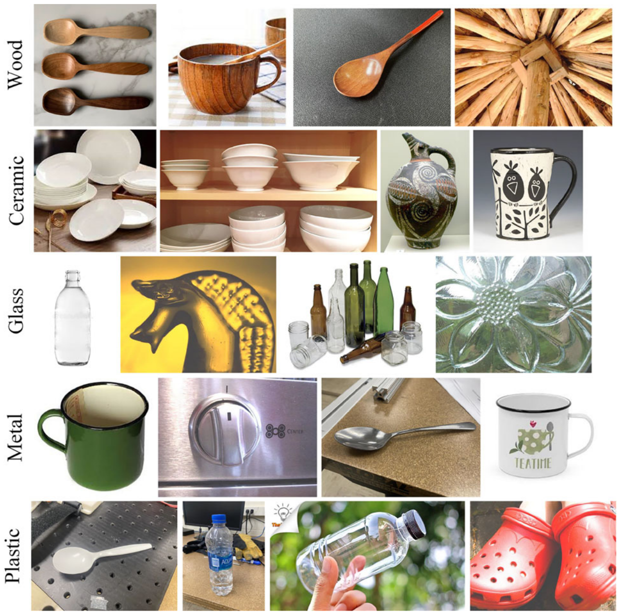

Fig. 7 Sample images used in the close-view material dataset

this dataset offer more intricate details and fewer contextual elements compared to those in MINC-2500.

8.2 Object detection and ARUCO marker detection

The FPV Grasp dataset was used to test the performance of different networks trained on the COCO dataset. The mean average precision (mAP) at 50% intersection over union (IOU) was used to quantify the object detection performance. The speed was measured based on the average inference time of 10 images using the E5-1260 CPU. The results are shown in Fig. 8. Multiple networks were tested, and YOLOV4 with a 0.75 threshold was selected based on mAP and speed.

The successful detection rate R_s of object detection and ARUCO marker detection can be calculated using the following equation:

$$R_s = \frac{\text{TP} - \text{FP}}{n} \tag{23}$$

where TP is true positive, which means that the ARUCO API detection successfully detects the marker, and the object detection successfully identifies the center object. FP is false positive, which means that the marker detection recognized the wrong marker or the object detection detects the wrong center object. n is the total number of test images. The experiments' successful detection rate was 90% in the collected FPV Grasp dataset.

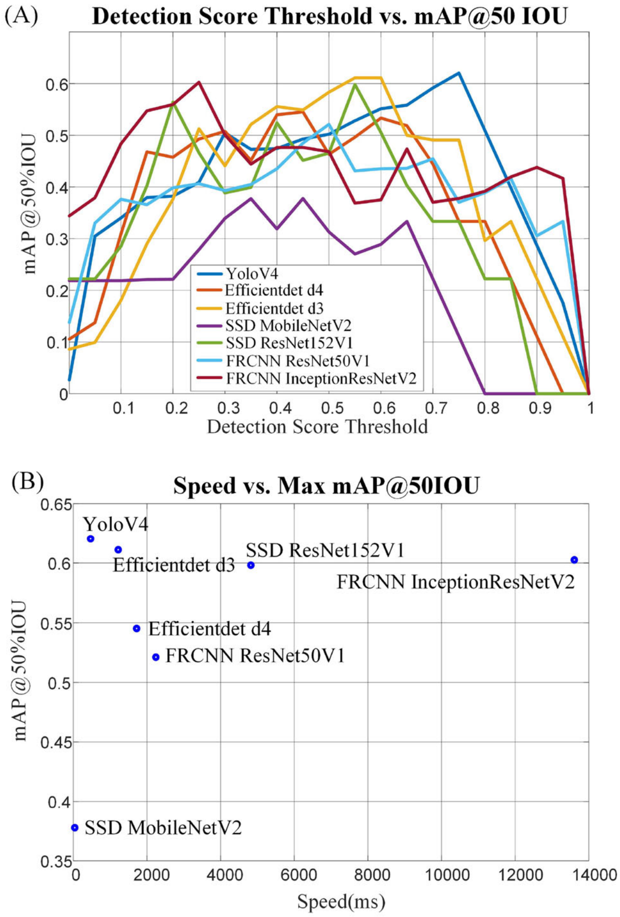

Fig. 8 Object detection results. **A** mAP at 50% IoU of 7 different state-of-the-art neural networks. **B** mAP vs. average inference time of each neural network

8.3 Object size estimation

The experiment involved evaluating the FPV Grasp dataset by comparing the detected target object's size with the ground truth sizes. For this purpose, images successfully detected by both the YoloV4 object detector and the ARUCO marker detector were utilized. To obtain the predicted size for each object, the average of the estimated sizes from different angles was taken. The ground truth sizes were determined based on the width and height of the orthographic projection, as illustrated in Fig. 9.

The obtained results are presented in Table 3. To quantify the difference between the predicted and actual object sizes, the percentage difference between the products of width (w) and height (h) was calculated. This evaluation metric is termed the mean absolute percentage error (MAPE). The MAPE difference between the predicted and actual object sizes was found to be 26.9%. The main source of this error was identified as the estimation process, particularly when utilizing the bounding box to estimate the object's dimen-

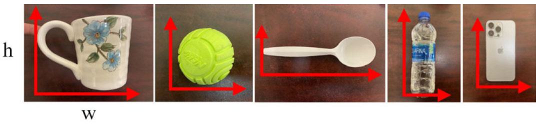

Fig. 9 Examples of size measurements

Table 3 Size estimation experimental results

Object	Actual* (cm)	Predicted* (cm)	Diff %
Plastic bottle A	6.5×11	8×13	45.4
Plastic bottle B	7×20	8×21	20
Plastic spoon	13×4	14×3.5	5.8
Plastic fork	14×4	14×3.5	12.5
Plastic cup	12×10	14×11	28.3
Plastic ball	7×7	6.8×6.8	5.6
Metal spoon	18×3.5	14.1×6.7	49.9
Metal fork	18×2.5	12×6	60
Metal cup	14×9	13.4×10.6	12.7
Wood spoon	16.5×4	11×7.8	30
Glass cup	12×14	12.9×15.1	15.9
Wine glass	20×5.5	19.4×7.3	28.7
Ceramic cup	19×11.5	18.8×15.2	30.8
Ceramic bowl	17.5×17.5	17×10	44.5
Cell phone	15×7.5	12×8.2	12.5
MAPE	–	–	26.9

[a] Actual*: The actual size is defined by the width times height in centimeters

[b] Predicted*: The predicted size is defined by the width times height in centimeters

sions. This error tends to occur when the object is placed at an angle during detection.

8.4 Object material detection

The training and testing results in the collected close-view material dataset are shown in Table 2. According to the accuracy and speed of classification, the material classification network used is ResNet50V2. The weight is transferred from the MINC-2500 dataset.

Material classification validation was also performed on the FPV Grasp dataset. The material classification accuracy for all detected objects was 96%. In addition to material classification, material segmentation is performed using the CRF method to visualize the result of material classification. Sample images of material segmentation are shown in Fig. 10. Because of the restricted data used for training, particularly the absence of pixel-wise labeled data, the accuracy of material segmentation is suboptimal and can only serve as a visualization tool for our material classification network. Our approach involves detecting the material situated at the center of the object and assumes that the target object has uniform material. In the future, as the availability of pixel-wise labeled material images increases, it will become possible to conduct end-to-end training. This can possibly improve segmentation results and further refine material detection.

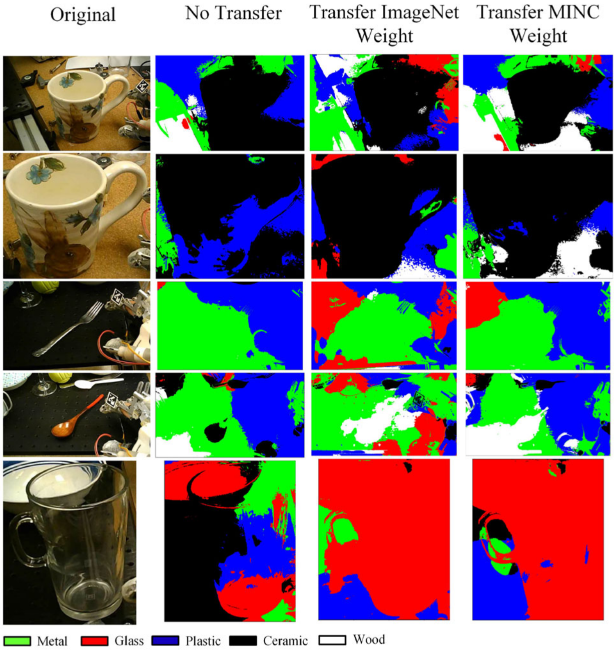

Fig. 10 Sample material segmentation results

It is essential to acknowledge that material detection relies solely on visual information, and appearances can be deceiving even for humans. This implies that lighting conditions, object colors, or reflections can all influence the effectiveness of the system. For instance, reflective objects may be recognized as metal objects due to lighting conditions.

8.5 Object weight estimation

The experiments on the FPV Grasp dataset involved comparing the weight of the target object with the weight of the corresponding ground truth. The materials used in the objects had different densities: plastic (0.92 g/cm^3), metal (7.85 g/cm^3), glass (2.7 g/cm^3), ceramic (6 g/cm^3), and wood (0.9 g/cm^3).

The results of these experiments are presented in Table 4. However, it is worth noting that the weight of the containers varied due to differences in the fluid level. For consistency, it was assumed that all containers were full. To assess the accuracy of the weight estimation, MAPE was employed as the evaluation metric. The MAPE between the predicted and actual object weights was found to be 59.8%. The relatively large weight estimation error can be attributed to the following factors. First, weight estimation is heavily influenced by size estimation, which in turn can be affected by the angle at which the object appears in the camera. Sec-

Table 4 Weight estimation experimental results

Object	Actual(g)	Predicted(g)	Diff (%)
Plastic bottle A	12–512	698	36.3*
Plastic bottle B	207	432	108.7
Plastic spoon	3	5	66.7
Plastic fork	3	5	66.7
Plastic cup	11–502	881	75.5*
Plastic ball	69	100	45
Metal spoon	48	74	54.2
Metal fork	22	57	159.1
Metal cup	172–576	619	7.5*
Wood spoon	7	8	14.3
Glass cup	358–779	1459	89.5*
Wine glass	188–369	399	8.1
Ceramic cup	480–1059	1756	65.8*
Ceramic bowl	315	596	89.2
Cell phone	222	200	10
MAPE	–	–	59.8

*: Containers have various weight due to the content. During weight estimation, we assume all containers are full of water

ond, the assumption of standard shapes for all objects, such as cylinders or boxes, may not hold true for most cases, where cups might have handles, and wine glasses may have long legs, leading to deviations from the standard shapes used in the estimation process. Furthermore, despite some instances of substantial percentage errors, the overall weight difference remains acceptable. For instance, the metal fork experienced a weight estimation error of 35 g, representing a 159.1% overestimation compared to its actual size. The average weight difference across all objects is 173 g, which still provides meaningful information for several reasons. Firstly, we deliberately overestimate the object's weight to prevent slipping, for instance, assuming a cup always has a full water level. While the output force may not be optimal, more force is applied and effectively reduces the risk of grasp failure. Secondly, in cases where the estimated weight is low, our slip-grasp force control program can make adjustments based on slippage. The weight estimation serves as an initial reference point for our slip-grasp force control program.

8.6 Grasp experiments

Given the preliminary stage of the proposed method, a single healthy male participant is involved in conducting the grasp experiments. Due to the nature of the exoskeleton glove used in this research, which is a rigid linkage exoskeleton, the user cannot apply any force to the fingertips of the exoskeleton linkages when wearing it.

The grasp procedure is as follows: The user initiates the system using a personalized voice command system [6] to capture a 1280×760 pixel image using the first-person view SVWSUN video glass. By employing the methods proposed in previous sections, the size and weight of the grasped object can be calculated. The 9-DoF MPU-9250 IMU detects the pitch, yaw, and roll of the exoskeleton glove. Using the weight of the object and the IMU data, the initial grasp force is computed, and the exoskeleton glove applies this force to each fingertip [44]. The slip-grasp system is subsequently employed to fine-tune the grasp force, ensuring stability during grasping.

During the experiment, each of the 15 objects present in the FPV Grasp dataset was subjected to 2–6 grasping attempts from various angles and fluid levels, resulting in a total of 64 grasp trials. Among these trials, 6 experienced failure of object detection, while 5 encountered errors in material detection. The grasp success rate is defined as the success in picking up the target object. The overall grasp success rate using vision-based HMI combined with the slip-grasp method was 87.5%.

The failure in grasping is attributed to the combination of the workspace limitation of the exoskeleton glove and mismatched grasp force. For instance, the exoskeleton utilized in this research exhibits coupled finger motion. When performing cylinder grasp and tip grasp, this coupled motion prevents the exoskeleton from grasping at the ideal contact angle and grasp trajectory. However, deviating from the ideal grasp angle or trajectory does not always result in failure but necessitates a more accurate initial grasp force. As depicted in Fig. 11, cylinder grasp and tip grasp exhibit a higher failure rate compared to other grasps. When paired with a vision system, both methods show significant improvement. A more detailed comparison is discussed in the next section.

8.7 Comparison between vision-based force estimation and slip grasp force planning

To illustrate the efficacy of the vision-based initial grasp force estimation method, we conducted 64 experiments solely employing the slip-grasp force planning approach, achieving a grasp success rate of 71.9%. However, when combining the slip-grasp method with the vision-based approach, the success rate increased to 87.5%. The success rates for each grasp category are depicted in Fig. 11.

The comparison experiment reveals that utilizing a combination of vision-based force estimation with the slip-grasp system leads to a higher success rate compared to using only the slip-grasp system. To demonstrate the benefits of utilizing the vision-based initial force estimation technique, we carried out an additional set of 20 grasp trials involving four distinct items: a plastic bottle, a wine glass, a plastic spoon, and a metal spoon. These particular objects were chosen based on their notable performance in previous grasp experiments.

Table 5 Comparison between vision-based force estimation and slip grasp force planning

Object	Slip-grasp (succ/total trials)	Vision (succ/total trials)	Slip-grasp (index force/thumb torque)	Vision (index force/thumb torque)
Plastic bottle	3/6	5/6	3.67N / 367Nmm	2.73N / 459Nmm
Wine glass	6/6	6/6	2.67N / 267Nmm	1.59N / 267Nmm
Plastic spoon	2/4	4/4	2N / 200Nmm	0.75N / 31.6Nmm
Metal spoon	3/4	4/4	2N / 200Nmm	1.2N / 50.8Nmm

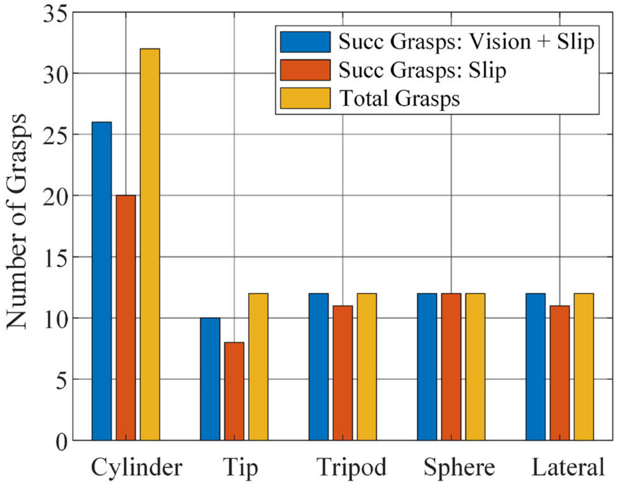

Fig. 11 Experimental result of grasping daily used objects. Blue: number of successful grasps performed using the vision-based initial force estimation with slip-grasp method. Red, number of successful grasps performed using only the slip-grasp method. Yellow, the total number of grasps for each individual method

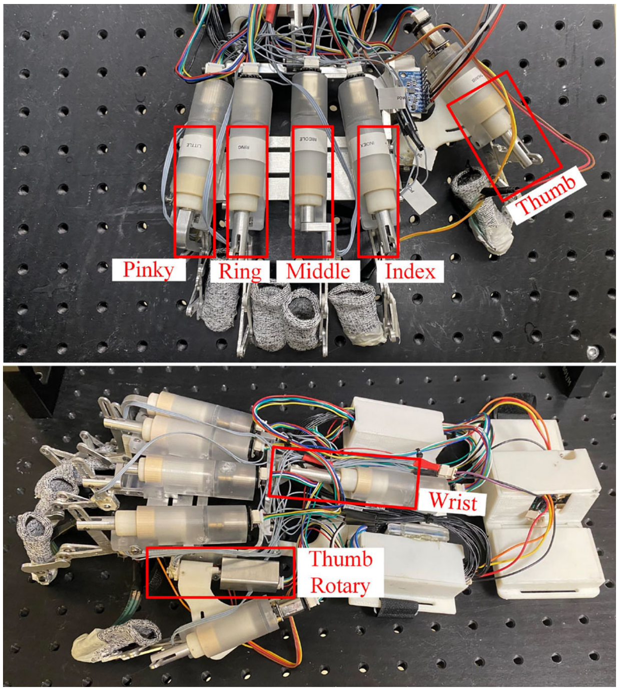

Fig. 12 The joint configuration of the exoskeleton glove in the grasp experiment. The thumb rotary joint of the exoskeleton facilitates movement of the thumb carpometacarpal joint in the human hand

For the vision-based method, the initial grasp force was determined using the vision-based force estimation system, and the slip-grasp method was not utilized in this experiment. For the slip-grasp method, a predefined initial grasp force of 2N and 200Nmm is used. This method adjusted the grasp force based on slippage to achieve a stable grasp (details can be found in paper by Xu et al. [23]).

The grasping process was facilitated by 7 SEAs as depicted in Fig. 12. The force and torque output of the index finger and thumb rotary joints, which are the most critical actuators during grasping, were measured and reported in Table 5.

The results from the additional 20 grasp experiments are presented in Table 5 and Fig. 13, demonstrate that the vision-based force estimation system can produce adequate initial grasp forces for various objects. This offers three main advantages during grasping. First, the initial grasp force estimate helps prevent the application of insufficient thumb torque, which can result in slippage. For example, in Fig. 13B, the plastic water bottle could not be lifted by the slip-grasp method due to the insufficient predefined thumb torque. Second, the initial grasp force can prevent the application of excessive force and torque. For example, in Fig. 13F, the plastic spoon could not be lifted by the slip-grasp method due to excessive fingertip force and thumb torque. Third, even for objects that can be successfully lifted by the slip-grasp method, incorporating a vision-based force estimation system allows for a reduction in the applied force (as shown in Table 5), thereby optimizing the grasping process.

8.8 Vision-based HMI system latency

The image processing is running on a desktop server with an E5-1260 CPU, and there is no GPU involved. The estimated size, weight, and surface friction coefficient are sent to the exoskeleton's onboard microcontroller, which generates the initial grasp force using IMU data and operates the exoskeleton. The computation time for processing a single image is around 700 ms. The processing time meets this application's requirements as only one image needs to go through the complete processing per grasp. The time consumption

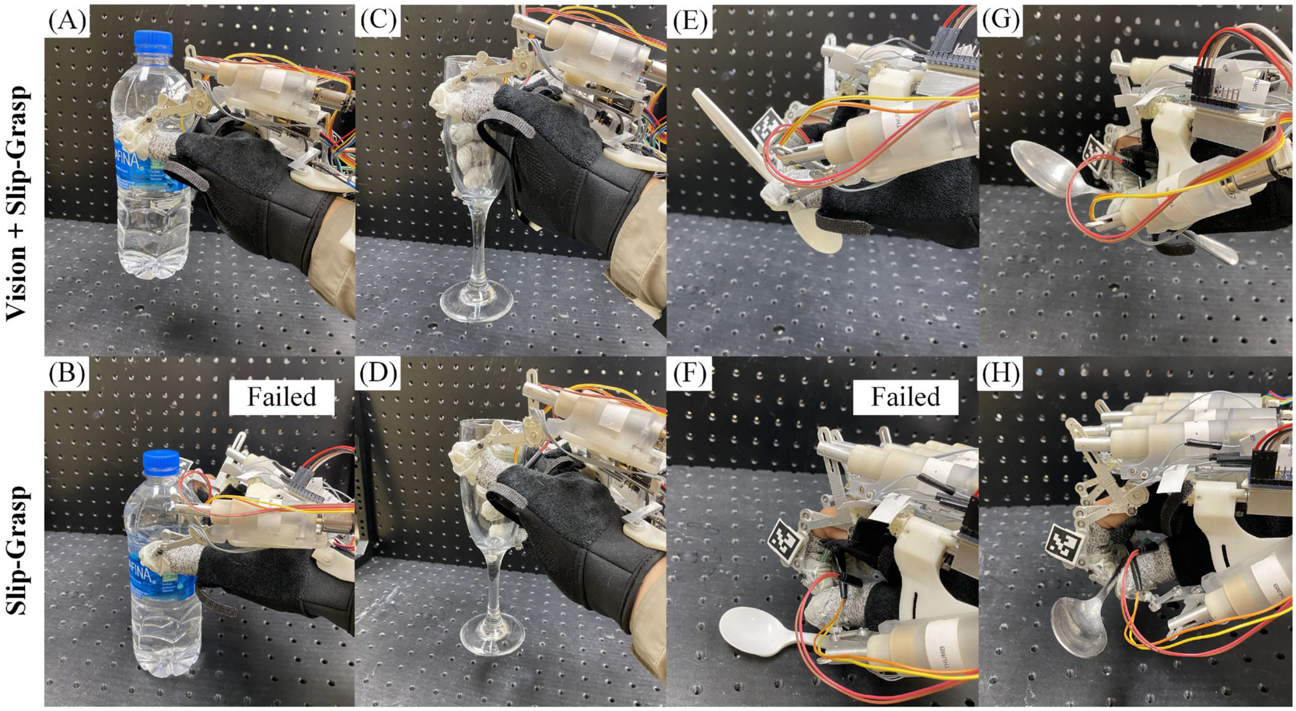

Fig. 13 Demonstration of grasping daily used objects using vision-based initial grasp force prediction method and slip-grasp method. **A** Successfully grasp a 512 g water bottle with vision system. **B** Failed to grasp a 512 g water bottle using the slip-grasp method due to inadequate thumb torque. **C** and **D** Successfully grasp an 188 g wine glass with both the vision system and the slip-grasp method. **E** Successfully grasp a 3 g plastic spoon with vision system. **F** Failed to grasp a 3 g plastic spoon using the slip-grasp method due to excessive force and torque. **G** and **H** Successfully grasp a 48 g metal spoon with both the vision system and the slip-grasp method

Table 6 Inference speed of one 1280×760 pixel image using the vision-based HMI

Section	Speed*(ms)
ARUCO marker detection	7
Object detection	470
Material classification	228
Size and weight estimation	3
Total	708

Speed*: the inference time is measured by averaging the inference time of ten images on a E5-1260 CPU

for processing one image is shown in Table 6. While the 700 ms delay is less than ideal, it does not present a significant problem in our system. It is important to note that our current system lacks a GPU, and the processing speed could be readily enhanced by incorporating one.

9 Conclusion

This paper presented a novel vision-based HMI aimed at estimating the initial grasp force required to manipulate a target object using an automated exoskeleton glove designed for patients with BPI.

The proposed approach employed object detection and material classification techniques to predict the initial grasp force, using information about the weight, size, and material of the object. In the FPV Grasp dataset collected for validating the HMI system, the object size estimation yielded a MAPE of 26.9%, while the object weight estimation exhibited a MAPE of 59.8%. Despite the relatively high MAPE for weight and size estimation, vision-based initial grasp force estimation still yielded a significant result, aiding in the grasping process.

The vision-based HMI successfully distinguished between different materials and accurately predicted the initial grasp force for objects of varying weights. When integrated with the slip-grasp method, the combined approach attained a 87.5% success rate, outperforming the standalone slip-grasp method (71.9%). These results highlighted the importance of estimating the initial grasp force to prevent slippage caused by inadequate or excessive application of force and torque.

The proposed work can be enhanced in several ways. Firstly, by expanding the close-view material dataset to include pixel-wise labeling, it opens up the possibility of end-to-end pixel-wise material segmentation. Secondly, incorporating image segmentation into the object detection process can lead to more accurate shapes and reduce errors in size and weight estimation. Additionally, conducting clinical experiments to test the proposed system could yield valuable insights for further improvement.

In conclusion, the proposed vision-based HMI demonstrated the potential to enhance the grasping capabilities of an automated exoskeleton glove, contributing to improved functionality and usability for patients with BPI. The findings of this experiment pave the way for future advancements in assistive technologies, facilitating more effective and reliable interactions between users and robotic systems.

Funding Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R21HD095027. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Human or animal rights The experimental procedure involving human subjects in this paper was approved by the Carilion Clinic Institutional Review Board (IRB-19-330).

References

1. Midha R (1997) Epidemiology of brachial plexus injuries in a multitrauma population. Neurosurgery 40(6):1182–1189
2. Xu W, Pradhan S, Guo Y, Bravo C, Ben-Tzvi P (2020) A novel design of a robotic glove system for patients with brachial plexus injuries. In: International design engineering technical conferences and computers and information in engineering conference, vol 83990. American Society of Mechanical Engineers, pp 010–10042
3. Jian EK, Gouwanda D, Kheng TK et al (2018) Wearable hand exoskeleton for activities of daily living. In: 2018 IEEE-EMBS conference on biomedical engineering and sciences (IECBES). IEEE, pp 221–225
4. Ge L, Chen F, Wang D, Zhang Y, Han D, Wang T, Gu G (2020) Design, modeling, and evaluation of fabric-based pneumatic actuators for soft wearable assistive gloves. Soft Rob 7(5):583–596
5. Xu W, Guo Y, Bravo C, Ben-Tzvi P (2023) Design, control, and experimental evaluation of a novel robotic glove system for patients with brachial plexus injuries. IEEE Trans Rob 39(2):1637–1652. https://doi.org/10.1109/TRO.2022.3220973
6. Guo Y, Xu W, Pradhan S, Bravo C, Ben-Tzvi P (2020) Integrated and configurable voice activation and speaker verification system for a robotic exoskeleton glove. In: International design engineering technical conferences and computers and information in engineering conference, vol 83990. American Society of Mechanical Engineers
7. Guo Y, Xu W, Pradhan S, Bravo C, Ben-Tzvi P (2022) Personalized voice activated grasping system for a robotic exoskeleton glove. Mechatronics 83:102745
8. Cutkosky MR, Howe RD (1990) Human grasp choice and robotic grasp analysis. In: Venkataraman ST, Iberall T (eds) Dextrous robot hands. Springer, New York, pp 5–31
9. Bicchi A, Kumar V (2000) Robotic grasping and contact: a review. In: Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065), vol 1. IEEE, pp 348–353
10. Bronks R, Brown J (1987) IEMG/force relationships in rapidly contracting human hand muscles. Electromyogr Clin Neurophysiol 27(8):509–515
11. Artemiadis PK, Kyriakopoulos KJ (2008) Estimating arm motion and force using EMG signals: on the control of exoskeletons. In: 2008 IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 279–284
12. Paek AY, Gailey A, Parikh P, Santello M, Contreras-Vidal J (2015) Predicting hand forces from scalp electroencephalography during isometric force production and object grasping. In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 7570–7573
13. Araujo RS, Silva CR, Netto SPN, Morya E, Brasil FL (2021) Development of a low-cost EEG-controlled hand exoskeleton 3D printed on textiles. Front Neurosci 15:626. https://doi.org/10.3389/fnins.2021.661569
14. Li M, He B, Liang Z, Zhao C-G, Chen J, Zhuo Y, Xu G, Xie J, Althoefer K (2019) An attention-controlled hand exoskeleton for the rehabilitation of finger extension and flexion using a rigid-soft combined mechanism. Front Neurorobot 13:34. https://doi.org/10.3389/fnbot.2019.00034
15. Wang X, Tran P, Callahan SM, Wolf SL, Desai JP (2019) Towards the development of a voice-controlled exoskeleton system for restoring hand function. In: 2019 international symposium on medical robotics (ISMR), pp 1–7. https://doi.org/10.1109/ISMR.2019.8710195
16. Kim YG, Little K, Noronha B, Xiloyannis M, Masia L, Accoto D (2020) A voice activated bi-articular exosuit for upper limb assistance during lifting tasks. Robot Comput Integr Manuf 66:101995. https://doi.org/10.1016/j.rcim.2020.101995
17. Kim D, Kang B, Kim KB, Choi H, Ha J, Cho K-J, Jo S (2019) Eyes are faster than hands: a soft wearable robot learns user intention from the egocentric view. Sci Robot. https://doi.org/10.1126/scirobotics.aav2949
18. Pham T-H, Kheddar A, Qammaz A, Argyros AA (2015) Towards force sensing from vision: observing hand-object interactions to infer manipulation forces. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2810–2819
19. Calandra R, Owens A, Jayaraman D, Lin J, Yuan W, Malik J, Adelson EH, Levine S (2018) More than a feeling: learning to grasp and regrasp using vision and touch. IEEE Robot Autom Lett 3(4):3300–3307
20. Yamaguchi A, Atkeson CG (2017) Grasp adaptation control with finger vision: verification with deformable and fragile objects. In: Proceedings of 35th annual conference robotics society of Japan.(RSJ), pp 1–301
21. Takamuku S, Gomi H (2019) Better grip force control by attending to the controlled object: evidence for direct force estimation from visual motion. Sci Rep 9(1):1–12
22. Stone K, Gonzalez C (2015) The contributions of vision and haptics to reaching and grasping. Front Psychol 6:1403. https://doi.org/10.3389/fpsyg.2015.01403
23. Xu W, Guo Y, Bravo C, Ben-Tzvi P (2022) Development and experimental evaluation of a novel portable haptic robotic exoskeleton glove system for patients with brachial plexus injuries. In: 2022 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 11115–11120. https://doi.org/10.1109/IROS47612.2022.9981468

24. Lee BJB, Williams A, Ben-Tzvi P (2018) Intelligent object grasping with sensor fusion for rehabilitation and assistive applications. IEEE Trans Neural Syst Rehabil Eng 26(8):1556–1565. https://doi.org/10.1109/TNSRE.2018.2848549
25. Romeo RA, Zollo L (2020) Methods and sensors for slip detection in robotics: a survey. IEEE Access 8:73027–73050. https://doi.org/10.1109/ACCESS.2020.2987849
26. James JW, Lepora NF (2020) Slip detection for grasp stabilization with a multifingered tactile robot hand. IEEE Trans Rob 37(2):506–519
27. Bell S, Upchurch P, Snavely N, Bala K (2015) Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
28. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International conference on computer vision (ICCV), pp 618–626. https://doi.org/10.1109/ICCV.2017.74
29. Krähenbühl P, Koltun V (2011) Efficient inference in fully connected crfs with Gaussian edge potentials. Adv Neural Inf Process Syst 24. https://proceedings.neurips.cc/paper_files/paper/2011/hash/beda24c1e1b46055dff2c39c98fd6fc1-Abstract.html
30. Guo Y, Xu W, Pradhan S, Bravo C, Ben-Tzvi P (2022) Personalized voice activated grasping system for a robotic exoskeleton glove. Mechatronics 83:102745
31. Haarman CJ, Hekman EE, Rietman JS, Kooij H (2023) Feasibility of reconstructing the glenohumeral center of rotation with a single camera setup. Prosthet Orthot Int 47(2):218–224
32. Yuan T, Song Y, Kraan GA, Goossens RH (2022) Identify finger rotation angles with ArUco markers and action cameras. J Comput Inf Sci Eng 22(3):031011
33. Chen S, Hong J, Zhang T, Li J, Guan Y (2019) Object detection using deep learning: single shot detector with a refined feature-fusion structure. In: 2019 IEEE international conference on real-time computing and robotics (RCAR), pp 219–224. https://doi.org/10.1109/RCAR47638.2019.9044027
34. Girshick R (2015) Fast R-CNN. In: 2015 IEEE international conference on computer vision (ICCV), pp 1440–1448. https://doi.org/10.1109/ICCV.2015.169
35. Tan M, Pang R, Le QV (2020) Efficientdet: scalable and efficient object detection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079
36. Gao C, Cai Q, Ming S (2020) Yolov4 object detection algorithm with efficient channel attention mechanism. In: 2020 5th international conference on mechanical, control and computer engineering (ICMCCE), pp 1764–1770. https://doi.org/10.1109/ICMCCE51767.2020.00387
37. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision. Springer, pp 740–755
38. Zhang H, Xue J, Dana K (2017) Deep ten: texture encoding network. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2896–2905. https://doi.org/10.1109/CVPR.2017.309
39. Zhao C, Sun L, Stolkin R (2017) A fully end-to-end deep learning approach for real-time simultaneous 3D reconstruction and material recognition. In: 2017 18th international conference on advanced robotics (ICAR), pp 75–82. https://doi.org/10.1109/ICAR.2017.8023499
40. Siddique N, Paheding S, Elkin CP, Devabhaktuni V (2021) U-net and its variants for medical image segmentation: a review of theory and applications. IEEE Access 9:82031–82057. https://doi.org/10.1109/ACCESS.2021.3086020
41. Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J (2016) NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5297–5307
42. Islam SAU, Bernstein DS (2019) Recursive least squares for real-time implementation [lecture notes]. IEEE Control Syst Mag 39(3):82–85
43. Kushner H (1967) Nonlinear filtering: the exact dynamical equations satisfied by the conditional mode. IEEE Trans Autom Control 12(3):262–267. https://doi.org/10.1109/TAC.1967.1098582
44. Guo Y, Xu W, Pradhan S, Bravo C, Ben-Tzvi P (2021) Data driven calibration and control of compact lightweight series elastic actuators for robotic exoskeleton gloves. IEEE Sens J 21(19):21120–21130

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.