Contextra: Detecting Object Grasps With Low-Power Cameras and Sensor Fusion On the Wrist

NATHAN DEVRIO, Carnegie Mellon University, USA
ROGER BOLDU, Meta Reality Labs Research, USA
ERIC WHITMIRE, Meta Reality Labs Research, USA
WOLF KIENZLE, Meta Reality Labs Research, USA

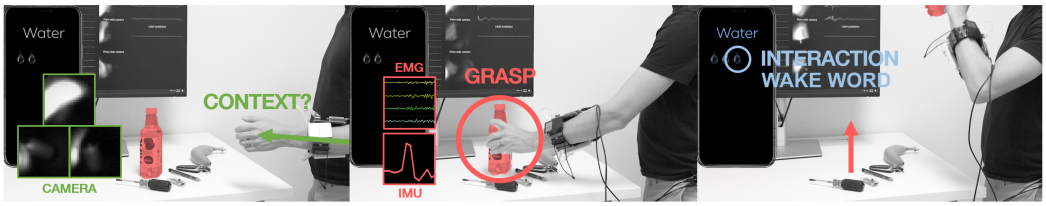


Fig. 1. Contextra is a wrist-worn sensing system that tries to determine a user's context (left) by fusing data from multiple low-resolution IR cameras, EMG electrodes, and an IMU to detect when a users grasps an object (middle). Once grasp is detected, a variety of downstream interactions can be enabled such as identifying held objects and habit tracking (right).

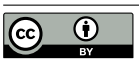
Knowing when a user picks up an object plays a vital role in many context-aware applications. For example, tracking water consumption, counting calories consumed, or reminding you to bring your keys are all context-centered scenarios involving picking up objects. In this project, we propose Contextra, a wrist-worn system that uses sensor fusion to recognize when a user grasps objects. Sensor fusion allows all parts of the grasp to be sensed in ways single channels cannot alone. In our wristband, we fuse EMG and IMU data with video captured from three low-power IR cameras. These cameras maintain privacy by using an active-illumination technique to only capture features close to the sensors. Beyond grasps alone, we see Contextra as playing a foundational role in providing continuous awareness of context triggers to extend the functionality of existing AI devices that cannot run continuously due to power and privacy concerns.

CCS Concepts: • **Human-centered computing** → **Mobile devices**.

Additional Key Words and Phrases: Smartwatch; Sensing; Grasp detection; Sensor fusion; Mobile devices; Interaction techniques

ACM Reference Format:
Nathan DeVrio, Roger Boldu, Eric Whitmire, and Wolf Kienzle. 2025. Contextra: Detecting Object Grasps With Low-Power Cameras and Sensor Fusion On the Wrist. *Proc. ACM Hum.-Comput. Interact.* 9, 5, Article MHCI006 (September 2025), 25 pages. https://doi.org/10.1145/3743741

Authors' Contact Information: Nathan DeVrio, Carnegie Mellon University, Pittsburgh, PA, USA, ndevrio@cmu.edu; Roger Boldu, Meta Reality Labs Research, Redmond, WA, USA, rboldu@meta.com; Eric Whitmire, Meta Reality Labs Research, Redmond, WA, USA, ewhitmire@meta.com; Wolf Kienzle, Meta Reality Labs Research, Redmond, WA, USA, wkienzle@meta.com.



ACM 2573-0142/2025/9-ARTMHCI006
https://doi.org/10.1145/3743741

1 Introduction

Many people carry multiple wearable devices with them wherever they go. Yet despite their omnipresence, these devices still know little information about the settings users are in. As an example, a smartphone may be able to infer that a user is making dinner if they have opened a cooking app and are looking through a recipe, but very rarely is this done automatically. If obtained, contextual information, such as if a user is making a meal, can be used to provide better adaptive and context-specific interfaces and experiences. In this work, we explore how a device on the wrist is uniquely suited to sensing information that can be used to inform context.

Smartwatches are the most commonly worn all-day wearable devices; many users are already acclimated to wearing them. Smartwatches are also worn in close proximity to the hand, the primary body part people use to interact with the world. As a result, they are well-positioned to sense signals from interactions that are too noisy and too difficult to detect from other parts of the body. Past HCI research around wearable sensing on the wrist has actively explored recognizing general context by means of intermediate problems such as pose tracking and activity recognition. In contrast, there has been less exploration of directly recognizing specific moments of micro context, like picking up an object, when just-in-time feedback is most valuable.

In this project, we propose Contextra, a system that uses sensor fusion on the wrist to recognize when a user picks up an object. We reference Wimmer's and Zaiti et al.'s [44, 48, 52] claim that *object grasps* are a foundational component of inferring when a user is in an interesting context that we might want to recognize and use it as the basis for this work's motivation. An "interesting" context relates to the idea that for some objects, if a user picks them up, it may be useful to the user to provide additional information or guidance to help them with their task. Examples include tracking caffeine consumed when picking up a coffee cup or providing explanation on how to use a tool. Wimmer's framework of grasp interactions [47] defines three stages of a grasp: capturing the grasp, identifying it, and interpreting it. In this work we focus on a technical, wearable solution for the first two stages. For the second stage, we focus on identifying whether a grasp occurred or not rather than recognizing the type of grasp from a known grasp signature. For the third stage, we discuss potential applications of Contextra in grasp interpretation, but leave this to future work.

Object grasps are complex actions, made up of multiple stages. Through experimentation we found certain types of sensors were better suited to detecting parts of a grasp than others. Past work has also shown the disadvantages of individual sensing modalities [28, 30, 31, 43]. Thus, using these factors as our motivation, we designed a system that uses a fusion of sensing approaches. On the wristband, we use optical, electromyography (EMG), and accelerometer sensors, each to detect different stages of the grasp. For example, the accelerometer is good at detecting initial movement of the arm and vibrations on object contact, while EMG can detect extension and flexion of the fingers when opening and closing the hand. Sensing the entire grasp process versus only trying to detect no grasp/grasp states has advantages for small objects and for EMG or acceleration whose features do not appear in steady-state signals. To clarify, data from the whole grasp process is used to pick up unique impulse features, but our target is to detect no grasp/grasp state transitions.

While EMG and accelerometers have been used in the past to try to detect grasps [7, 43], our incorporation of multi-view optical data from three low-resolution cameras is novel for such a wearable system. Beyond providing comparable grasp recognition performance on previously tested objects, we believe our camera fusion approach also has the potential to improve reliability and robustness across objects and users. This is because EMG and accelerometer data can be high variance depending on a particular grasp style [14, 32], while the presence of objects in the camera views tended to be a lower variance feature. These advantages show how fusing many sensors

System	Worn position	Sensing method	Power consumption	Weight	Performance (within-session)	Performance (cross-session)	Performance (cross-user)	Objects
Pistohl et al.	Brain implant	Electrocorticography (ECoG)	n/a	n/a	0.89 recall	n/a	n/a	Drink mug
Xiao et al.	Wrist, forearm	Force myography (FMG)	~20 mW (*estimated*)	~40 g (*estimated*)	95% accuracy	n/a	n/a	12 x 6 cm cylinder
Maufroy et al.	Forearm	sEMG (Myo)	34 mW	93 g	96.8% accuracy	88.9% accuracy	82.8% accuracy	Hammer, 1 kg mechanical part, electric screwdriver, ratchet, plastic basket
Berlin et al.	Wrist, back of hand	RFID	210 mW	~35 g	72.8% detected (no ML)	72.8% detected (no ML)	72.8% detected (no ML)	Plastic bottle (L), plastic bottle (S), screwdriver, USB stick, hammer, box cutter, glass bottle, lipstick,
Theiss et al.	Wrist, back of hand	RFID, EMG, IMU	238 mW	~128 g	0.95 recall	n/a	n/a	Plastic bottle (L), plastic bottle (S), screwdriver, USB stick, hammer, box cutter, cardboard box
Cofer et al.	Wrist	Near-IR pair detectors	n/a	~60 g (*estimated*)	98% accuracy	n/a	n/a	Cylinder
DeVrio et al. (Contextra)	Wrist	IR cameras (3), EMG, IMU	255 mW	148 g	n/a	0.87 precision, 0.86 recall	0.82 precision, 0.84 recall	Plastic bottle (L), plastic bottle (S), screwdriver, USB stick, pencil, car keys, banana, notepad

Fig. 2. High-level overview of highly-related worn grasp-detection systems. Comparable numbers drawn from published materials as best possible, but please consult individual papers for specifics.

together is advantageous and how Contextra offers a unique contribution to the space of grasp detection that has not been filled by previous wearable devices.

We emphasize we tackle the problem of detecting grasps from continuous sensor streams, not just classifying the type of a known grasp. Compared to more common problems in prior work of classifying hand pose, grasp type, or what object is held, ours is a unique problem with its own challenges. We call this problem *continuous grasp detection* and believe it is an exciting problem because if solved, a system could be an automatic "wake word" recognizer for contextual AI, saving devices with conventional sensors like head-mounted cameras enormous power. In sum, in this paper we contribute the following:

(1) A multi-view camera sensing approach on the wrist for detecting when object grasp events occur.
(2) A late-fusion model that incorporates EMG and acceleration sensor features from the entire grasping process to overcome camera weaknesses like occlusion.
(3) Our entire system is low-power and privacy-sensitive so that it could run all day and prevent higher-power headset cameras and microphones needed for context-dependent apps from burning power and invading user privacy.

2 Literature Review

In this section, we review past systems that used wearable sensors to look at objects held in a user's hand. We break out our discussion into three primary subsections. These range from (1) using held objects to recognize activities, the least similar to our task, to (2) classifying grasp types, and finally (3) detecting object grasps, the most similar to our task. Although there is overlap between our techniques and those used in grasp research for robotic hands, we omitted these works because of the gap in interaction task and the lack of human physiology to sense. The one exception is recent work on (4) sensor fusion to plan grasps for prosthetic robotic hands which we discuss in a final subsection.

2.1 Object-Centered Activity Recognition

The first category of work we examine involves systems that use only wearable sensors (i.e., no instrumentation of objects) to detect an object a user is holding and/or how is it being used to recognize a particular activity. These sorts of systems are unique from Contextra in that they do not sense grasp directly and are not focused on recognizing grasps, and instead focus on recognizing

activities or use of held objects, which is different from our goal. Despite this, many of the sensing techniques used, particularly in camera-based systems are also applicable for our grasp recognition task and worth explaining because they both involve imaging objects near the hand.

Cameras are one of the most powerful sensors available for sensing object-based activities. This is because they are able to directly image if there is something being held in or near the hand. Maekawa et al. conducted investigations through multiple works showing how a camera on the wrist could effectively be used in combination with IMU and microphone data to capture information what object a user is holding to classify activities of daily living [29, 30]. Independently, IMUs, particularly the accelerometer component have also been popular for recognizing object-based activities [8, 25, 34]. Common techniques when using IMUs involve sensing grasping posture and motion with held objects to get clues about the activity [8, 34]. Other less common sensing schemes that have been used to recognize object-based activities include magnetic sensors [26, 27, 46], capacitive sensors [38], and measuring vibrations from a voice coil actuator [35].

2.2 Grasp Type Classification

There is a separate body of work focused on the problem of classifying what type of grasp is being used. Knowing what type of grasp a user is using is often important because it can clue the identity of the object or what context or activity the object might be used in. Many different sensing approaches have been used on wearable devices to classify grasp types without instrumenting objects including EMG [1, 16], force myography [22], cameras [9, 24, 49], and inference directly from hand pose [54].

The approaches in these works are in general broken into two parts. First, sensing is used to detect what class of grasp is being used, such as a pinch, palmar, or cylindrical [1, 9, 16, 22, 49]. In some cases such as Fan et al's work using EMG, they also examined how performance varied for objects of different sizes and weights as these characteristics are directly correlated with the strength of EMG signals [16].

The second part of this category involves taking information about the grasp type and combining with prior knowledge about a specific context or object set to identify what object the user is holding [9, 16]. In Section 6, we describe our vision for how Contextra could be combined with a cameras on a headset to recognize objects in a power-efficient manner. Additionally, we discuss how objects could be directly identified from our camera data given prior information, similar to the works discussed in this section.

2.3 Object Grasp Detection

The last category of systems are focused on the same interaction problem as our system. Namely, given a continuous sequence of sensor data, can we identify at which moments, if at all, a user grasps objects? Systems in this category are most related to Contextra. As such, to show a direct comparison of Contextra to prior work, we provide a table in Figure 2 that provides a high-level overview of technical properties and performance for systems of this category.

Early on and to this day, the most popular approach for detecting grasps in this way has been using RFID. These systems instrument objects with passive RFID tags and place an RFID reader on the user's body to detect (via proximity) when objects are picked up. Schmidt et al. [39] pioneered this technique with a more abstract goal focused around interactions with objects once held and used RFID reader prototypes integrated into clothing. Since then, many systems have iterated on the idea, compacting the prototype into a glove [18] or wristband [7, 17, 43].

While reliable, a major drawback with using RFID is that it requires all objects to be instrumented with tags ahead of time. To overcome this, many other sensing techniques have been used to detect grasps either instead of or in addition to RFID. The most popular of these involve sensing biosignals

generated from muscles or neurons that are characteristic of the arm grasping an object. Example biosensing techniques that have been explored in the past include EMG [31, 43], force myography [50], and electrocorticography [37]. While biosignals can generate strong features and avoid object instrumentation, their practical use is limited by the high variance of their signals across different user populations.

One of the most promising ways to retain the benefits of biosignals like EMG while overcoming their limited robustness is to use data from additional sensing channels such as accelerometers [43]. Outside of EMG, tiered sensing has also shown promise when combining channels such as accelerometers and near-infrared detector pairs [11]. Two of the systems just mentioned are the closest to Contextra's approach, thus we describe them in more detail here.

Theiss et al. [43] created a wearable system for predicting grasps using a tiered approach to sensing to minimize power consumption. By default, an accelerometer was running, then when motion was detected, the EMG system was woken up. Then if a grasp was detected, RFID was used to recognize what object was picked up. Their grasp detector is notable because of its goal to detect grasps by using a fusion of features derived from accelerometer and EMG data. We note a critical difference that Theiss et al. used an EMG (Myo) band placed on the upper forearm while ours was worn at or close to the wrist. Sensed EMG signals are much stronger where muscle is thicker near the elbow, however this location is not practical for a system like ours that envisions itself as a part of a future smartwatch.

In Cofer et al.'s [11] work on the other hand, they used optical sensor data from 20 near infrared emitter-detector pairs. The NIR sensors were placed on a wristband and aimed inwards towards the wrist to measure tissue displacement and detect when objects were touched (picked up) by users. When a grasp was detected, and IMU was used to classify the grasp type. Note the NIR sensors only collect a single brightness value each, thus are much lower resolution than the cameras we use. Although the sensors used in Cofer et al.'s system were less like Contextra than Theiss et al.'s, their 1D-CNN model is more similar to ours than Theiss et al.'s decision-tree based model.

Compared to these past systems, we believe Contextra presents a novel innovation in fusing IMU and EMG data with feature-rich optical data from multi-view cameras. No prior system has attempted to use wearable camera data (single or multi-view) alongside EMG and IMU data for grasp detection. For the problem of binary grasp detection, Contextra is the first system to fuse data from different sensor types together. Prior systems that employed multiple types of sensors only used them in tiered approaches where one sensor detected an event that was used to trigger another [11, 43]. Finally, to combine our sensing channels, we took advantage of a multi-layer, late fusion deep learning model that stabilized performance across objects.

In practice, this means our system is more accurate at detecting when objects are picked up than prior work, and can be expected to work more reliably across a wider variety of different object shapes and sizes. This is the primary advantage of our technique over those compared against in Figure 2. While our grasp performance is superior to some other systems, for those with comparable performance Contextra demonstrated an ability to work across a wider set of objects. For example, Xiao et al. [50], Maufroy et al. [31], and Theiss et al. [43] relied heavily on an EMG-based technique, thus performance was only strong for larger and heavier objects with required more muscle contraction. We believe our improvements in grasp detection versatility pave the way for a general-purpose context-aware agent that can provide feedback to a user based on what they are holding in a way not possible with previous techniques.

2.4 Sensor Fusion for Planning Prosthetic Hand Grasps

Robots always know when they are initiating a grasp. Thus, instead of recognizing when grasps occur, work in the robotics field is typically centered around separate problems of determining the

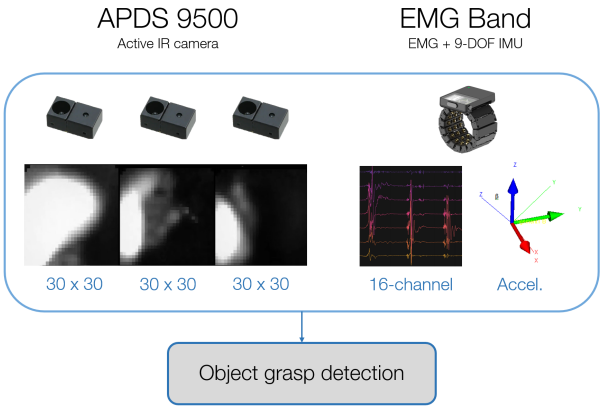


Fig. 3. Contextra system diagram. We take 30 × 30 optical data from three mini active IR cameras and fuse it with 16 EMG channels and IMU acceleration. These sensing streams are combined in our model to perform object grasp detection.

correct pose to execute the grasp and planning the motion [23]. However, with recent advances in prosthetics, there have new avenues of robotics research focused on using wearable sensors to initiate and plan grasps for prosthetic hands. These works differ in that there is a human wearing the sensors who expresses an intent to grasp. The primary sensing modality used is EMG to allow users to use natural muscle contractions to control the hand. Secondary sensing modalities such as eye-tracking, IMUs, and cameras have also been used in sensor fusion which is highly relevant to our system.

Deshmukh et al. [12] and Zandigohar et al. [53] used a fusion of EMG and a head-mounted camera to plan grasps. In both they created a camera-EMG fusion model. Deshmukh et al. focused on classifying grasp types from EMG data [12] and Zandigohar et al. segmented continuous EMG data into four grasp motion phases [53]. On the other hand, Shi et al. used a model combining eye tracking and a camera from a headset and EMG and an IMU from the arm to control a prosthetic and plan the pose of grasps [41].

While many of the technical techniques used overlap, there are key differences between these works and Contextra. The first key difference between these works in ours is the difference in task. Namely, these works focused on planning grasp hand poses rather than using existing hand poses and objects to detect grasp moments. Second, our technique has the novel advantage of using multiple lower resolution cameras worn on the wrist instead of a single high resolution camera on the head. This change allowed Contextra to consume much less power while achieving strong performance.

3 Implementation

Our work on Contextra involved developing both novel sensing hardware and the software to process sensor data. When combined, these components could automatically detect when the wearer picked up objects. In this section we detail their technical background and implementation.

3.1 Sensor Selection

One of the most important questions in the initial design of our system was what sensing modalities to use. To decide, we broke down the anatomy of a grasp to identify opportunities for sensing. A typical grasp begins with the user reaching for an object by moving their arm. An IMU sensor,

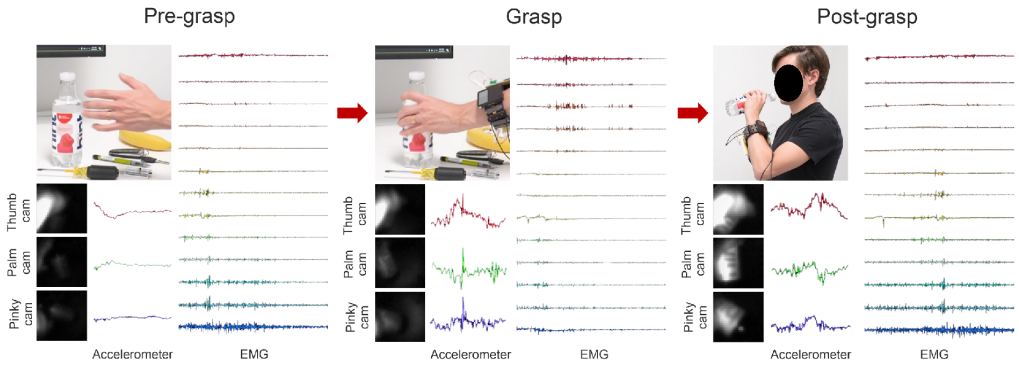


Fig. 4. Timeline of a user grasping an object, along with demonstrative traces of sensor data from our different sources. In each frame, we highlight how different sensing channels (optical, EMG, and acceleration) can contribute by sensing different features of the grasp. For example, motion of reaching for an object can be sensed by the accelerometer, then the object can be imaged by the cameras when opening the grip, and finally EMG can sense muscle contraction when lifting the object.

particularly an accelerometer, is suited for detecting gross motion such as this. Next, a user opens their hand (if not already), makes contact with the object, and closes their hand around it. An accelerometer may be useful again for detecting vibrations from contacting the object, but the extension and flexion of the fingers may be better sensed using EMG. In addition, cameras could be used to detect objects that have come into close contact with the hand and may be likely picked up. Finally, a user tightens their grip on the object and picks it up to complete the grasp. Depending on the weight of the object, EMG can be useful here to detect muscle strain during the lift. Otherwise, gross arm lift movement can be sensed using the IMU again.

While EMG and an accelerometer are powerful on their own, they do have certain disadvantages that make it hard to detect some grasps. A characteristic example is picking up objects slowly or lightly. A prominent EMG feature useful in grasp detection is the transient change in signal power. When picking up a light object, there is less strain put on the arm muscles so signal power changes less, to the point where in some cases it is barely noticeable from an EMG trace if an object is being held or not. Similarly, a prominent accelerometer feature is looking for signal spikes produced by vibrations when making contact with an object or surface. When objects are picked up quickly or aggressively, these spikes are prominent. On the other hand, when objects are picked up slowly or gently, the spikes can be nonexistent.

As a result of these disadvantages, we were motivated to pursue a sensor fusion approach where we combined data from multiple sensing modalities. Beyond EMG, and an accelerometer, which each have their own advantages for sensing different parts of the grasp, we searched for an additional sensing channel we could use to recognize objects held in the hand. In particular, we wanted a sensor that could use a static frame to detect grasps since transient features were already well covered by EMG and the accelerometer and could be diminished depending on the grasp. A promising candidate was to use optical data, such as from a depth sensor or camera, to see if an object was being held in that moment. However, optical sensors often have their own disadvantages, such as being vulnerable to occlusion, betraying user privacy, and consuming more power.

With all this in mind, we decided to take a balanced approach and used a series of three infrared (IR) cameras which were low-power and low-resolution and put together could prevent occlusion

Component	Power consumption (mW) @ 3.3 V
Cameras (3 x APDS9500)	33 mW
EMG band (EMG + IMU)	150 mW
ML model (A15 Bionic)	72 mW
	Total = 255 mW

Table 1. Power consumption of each system component used for grasp detection.

with multiple views. In more detail, the reason we used three cameras as opposed to one was because each camera can capture a different part of the hand and grasp that may be occluded in a single, static view. For example, a camera near the palm can see pinches of the index finger and thumb, but not objects that are approached from the side. This approach was inspired, in part, by other recent work taking advantage of multiple, smaller camera views for hand pose tracking [13, 19, 51]. We decided upon the number of cameras (not more or less) through an initial exploration in our simulator (detailed in subsection below) that showed three could well capture all visual aspects of the grasp without redundancies.

A diagram visualizing our chosen sensing channels can be seen in Figure 3. We note that despite our decision to use multiple sensors at the same time (mini IR cameras, EMG, and accelerometer), compared to a higher-res camera, total power consumption is still less and we still do not collect any data more privacy-compromising than 30 × 30 IR images (with background removed).

3.2 Camera Sensor Details

Contextra uses the APDS9500 optical sensors. These sensors are marketed to detect gestures using built-in blob detection algorithms. In our work, we used them as mini IR cameras by reprogramming the firmware to output raw camera frames over SPI. Notably, beside the imaging sensor, each component also includes an IR LED that can be used for active illumination of the target. Cameras were chosen as opposed to just photodiodes because of the number and quality of features cameras provide. Photodiodes may detect the presence of a target, but a camera can better see its shape to discern the target is a finger, not a held object.

When taking a picture of a target, we capture a sequence of three frames while toggling the LED from OFF-ON-OFF. Once complete, the two LED-OFF frames are averaged and subtracted from the LED-ON frame to produce, using in-situ hardware processing, a single 30 × 30 frame. In this frame, only the objects illuminated by the LED are visible (e.g., those close to the sensor), while all of the background is removed. An example an output frame can be seen in Figure 5. The advantage of this approach is that the most prominent features in the output images are those which are salient to the grasp detection problem, namely the hand and objects near the hand. This makes the downstream recognition task easier by obviating the need to perform separate object segmentation on the camera frames to find the hand and objects. In addition, by removing the background, coupled with the fact that the cameras are already low-resolution, we betray significantly less privacy than higher-resolution cameras.

The goal of the camera sensors we used was to image objects held in the hand, thus one potential confounding factor is the presence of other, non-grasped objects nearby. While certainly a challenge, this is a situation that highlights the benefits of fusing other sensing modalities. For example, IMU and EMG would not trigger when just hovering near, but not grasping other objects. In addition, because of their limited field-of-view and proximity to the hand, many objects are naturally occluded and only those directly in or next to the hand are visible.

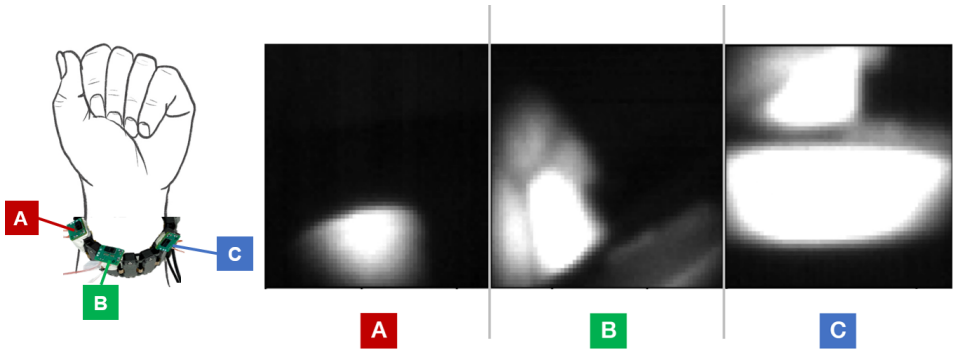


Fig. 5. Example data from the three wrist cameras and mappings to their locations on the wrist; captured while picking up a paper cup from a table. These views show what kind of features can be picked up by each camera.

3.3 Power Consumption

Each camera outputs frames at 280 FPS. By reprogramming registers, we were able to lower the LED peak current to 56.7 mA and the exposure time the LED was on to 216 us. This led to only a 6% duty cycle for the LED and 3.4 mA average LED current draw which made each sensor only consume 11 mW during continuous operation. Considering that comparable small-package depth sensors and higher-res cameras can consume in the range of hundreds of mW [15, 42], using these sensors could result in an up to 10× savings in power. The low-power consumption of our optical sensing stack was one of the driving factors behind Contextra's overall design and purpose. Since it runs at such a low power, Contextra can run continuously to recognize grasps where other more powerful but higher-power systems cannot. In this way, even when Contextra cannot sense all parts of a context, it can provide a trigger to wake up other devices in a low-power sleep, such as smartglasses or a mixed reality headset, to perform deeper recognition.

Many existing commercial VR and XR headsets have built-in cameras used for pose tracking, however running and processing these cameras is a significant reason why they do not possess "all day" battery life. The Meta Quest 3 and Apple Vision Pro, two popular headsets with such capabilities, only have battery lives of 2.2 and 2.0 hours, respectively [3, 33]. In addition, although such headsets already use cameras (thus camera-grasp detection would not increase power overhead), these devices are also typically used only for short sessions of entertainment or productivity. Our target for Contextra is instead to augment devices like AR smart glasses that do not constantly run cameras and could have an "all day" runtime. This is the reason we advocated for using less powerful but lower-power camera sensors.

Outside of the cameras, our band that collects EMG and IMU data consumes 150 mW and our estimate for the power consumption of our model on a mobile chipset is 72 mW (details in subsection 3.7). A breakdown of the power consumption of the entire system by component is shown in Table 1.

3.4 Comparison to Other Camera Sensors

When selecting the camera sensor, we sought a camera that possessed a small package size, low power, active illumination, high-frame rate, and a resolution large enough to sense object and hand features. The APDS9500 fit all of these criteria and thus we chose it because it was easy to procure

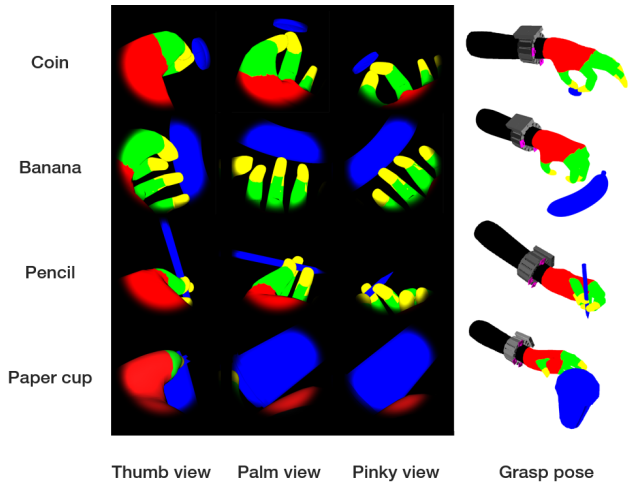


Fig. 6. Views from the final camera positions of the thumb, palm, and pinky during a run of our simulator. On the right we show a rendering of the grasp pose looked like for each of the four test objects shown.

and interface with. However, there are alternatives used by other systems with similar properties or minor trade-offs.

Maekawa et al. used an unspecified camera which captures 352 x 288 pixel images at 6 Hz (no power specified) [30] OptiRing from Waghmare et al. used the OVM6948 and a red LED [45]. This sensor had a 200 x 200 pixel resolution, but was limited to 30 FPS and required 39 mW to run due to extra analog processing required of its signal. CyclopsRing by Chan et al. used a SONY CCD sensor that captured 640 x 480 images at 30 FPS [10] No power was listed but the similar OV7670 consumes 60 mW continuously [6] General trends from these systems and other popular sensors (*OV2640*: 1600 x 1200 px, 15 FPS, 125 mW [5], *IMX708 Pi Camera 3*: 4608 x 2592 px, 10 FPS, >750 mW [4]), are that within this form factor, resolution can be increased, but this trades off frame rate and power consumed. These were two properties we wanted to optimize for, which justified our balanced choice of resolution, frame-rate, and power.

One of the most competitive alternatives is the Himax HM01B0. This sensor, used by Maruki et al. in their IRIS ring-worn camera, runs up to 60 FPS at 4 mW and is ~2x as large as ours. There is no active illumination however; pulsing an external IR LED would result in a total power similar to ours and add size. Notably, the HM01B0 has a 320 x 320 pixel resolution, much higher than ours. While more resolution may help discern tiny objects better, many of the contextual features we detect (e.g., presence of hand and object) are detectable with a lower resolution.

While Contextra could operate with higher-res camera sensors, there are three main drawbacks to consider. First, high-res cameras (with some exceptions such as the HM01B0) tend to consume significantly more power [4–6]. Second, the higher the resolution, the greater the bandwidth burden, which can slow overall system speed. Finally, increasing resolution increases the risk to invade user privacy. Down-sampling a high-res camera could be used to alleviate this, but the other two drawbacks remain. Additionally, a lens with a short focal length could help to preserve privacy when designing a future high-res camera sensor, however using external lenses with existing sensors adds bulk.

3.5 Camera Position Optimization

In Contextra, we take three of the camera sensors and place them at different positions around the wrist. The positions, which are shown in Figure 5, include (A) beside the thumb, (B) in front of the palm and ring finger, and (C) beside the pinky finger. These positions were found, after experimenting with live data, to provide the most information on object grasps across all camera views.

We target the camera sensors to be placed on the front of a smartwatch band, in a normal worn position on the wrist. From experimenting with sensor data, we knew the rough positions we wanted the three sensors to be located on the wrist. However, to refine further and determine the exact position, pitch angle, and tilt angle of each sensor, we made a simulator that could take in 3D models of a grasp pose and an object and output the best sensor configurations.

The simulator modeled each sensor as a directional light source and camera with the correct field-of-view (45°) to match the real-world sensor. To ensure that the simulated cameras did not all converge the same "best" view, we implemented strict boundaries on how far the sensors could move from their starting positions (thumb, palm, pinky), and how much they could tilt. This resulted in views which could image the same target, but retained the unique perspective from their respective wrist position.

With camera sensors, placing them too close to the hand can lead to occlusion by the palm [13]. This is part of the reason why our simulation was so important and we experimented with changing the sensor tilt angle. In our simulation, we assumed the smartwatch band was placed on the wrist, ~3.5 cm from the base of the palm. During our Evaluation, we asked users to wear the real band in this same position, however due to a fixed minimum band size and differences in users' wrist sizes, the band was worn further up the forearm for some users (such as the model in Figure 1). Thus, although we did not formally evaluate band placement on the wrist/forearm, the wrist was the primary worn position and variations were represented in our data.

In the simulation all parts of the hand and object that could be imaged were painted a unique color according to their region of interest. The regions we painted included the palm (red), fingers (green), fingertips (yellow), and object (blue). The purpose of the simulator was to take input on whether it should reward having more or less of certain regions in each image and find sensor configurations that maximized this reward function. During operation, the following process was used:

(1) Instantiate the real-world experimented sensor positions as origin points.
(2) Define a range of possible sensor translations and rotations (from user input).
(3) Iterate over all possible sensor configurations and capture a 2D image at each.
(4) Use a weighted function to calculate a score for each image by looking at each pixel, assigning a score (0: black, 1: red, 2: green, 3: yellow, 4: blue) based on what color that pixel is, and summing all pixel scores.
(5) Average the best angle and position configurations across different grasp poses to get the best overall configuration.

In our analysis, we picked a range of translations and rotations to match the physical constraints of the band. In our color-weighted function, we assigned highest weights to the fingertips and object (most important for grasp) and lowest to the palm (least important for grasp). After running this simulation on four different objects, while varying approach and wrist angles, we averaged the properties of the best configurations. From here we made minor tweaks to account for outlier views, and produced a "simulated optimal" sensor configuration, shown for different objects in Figure 6.

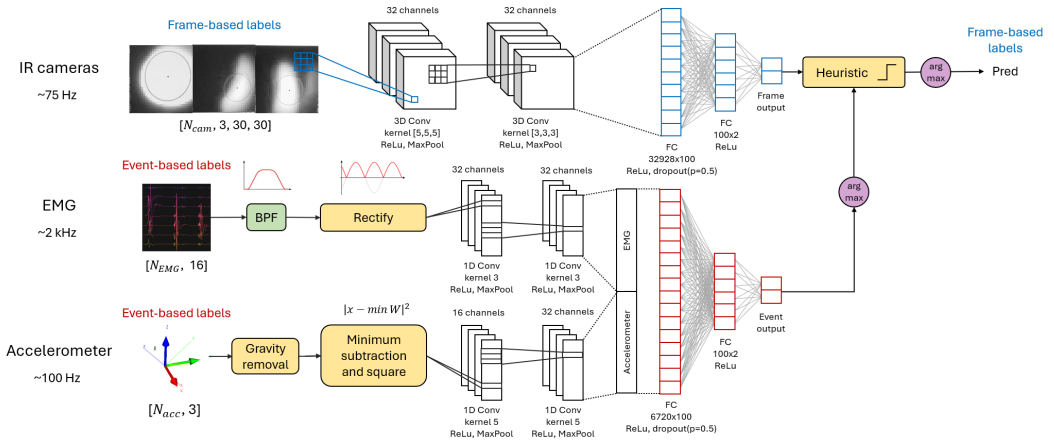


Fig. 7. A diagram of our model architecture. Our model was split into two branches based on ground truth label structure. For camera data, we used frame-based labels which indicated a touch or no touch state and for EMG and IMU data, we used event-based labels which indicated when grasp events (pick up or set down) occurred. EMG and IMU data was combined using a middle-fusion approach and then passed through a heuristic boost function to convert its output to match the frame-based labels. Once in a common format, the outputs from both branches were combined and returned with the frame-based labels.

The resultant configuration produced by our simulator is what we used to guide the mounting of our real-world sensors on the EMG band. Anecdotally, we noticed an immediate improvement in the features we were able to image in live camera data after switching to simulation-verified sensor positions. Outside of Contextra, we believe there is potential in this approach for other wrist-based optical sensing systems and believe it could be improved by using motion capture data to pose the 3D models.

3.6 Data Processing

A challenge when processing our data was that each sensor stream came from a different source, each with a different data rate, and with frame timing that could not be synchronized. This was a problem because our fusion model needs fixed-size CNN inputs. Without interpolation, if we received 26 EMG frames for 1 camera frame and 27 EMG frames for the next camera frame, our fused CNN input would not be fixed size. To manage this, we employed a buffered window approach. In this approach, data from all sources was buffered for one second. Each second, a timer was called that emptied these buffers and interpolated them to the expected FPS of each stream to account for possible dropped or extra frames. For example, for one camera, we expected 75 new frames each second. This was downsampled from the 280 FPS output of the sensor to improve model latency with negligible performance impact. If 74 or 76 were received instead, we interpolated *in time* to get exactly 75 frames. The effect of this on the final data was almost undetectable. Afterwards, the one-second chunks collected could be rearranged and overlapped to produce fixed-sized, windowed inputs to the model. We used a 400 ms window with 20% overlap, which has been found to encapsulate the duration of a grasp [50].

The one second buffer we used in our implementation meant in live operation, predictions would take one second to arrive plus the inference time of the model (4.4 ms). We chose this buffer size to minimize interpolation needed to batch data across sensor streams. However, with a minor increase

in interpolation, buffer size can be reduced down to 400 ms, the size of our model's window. This would cap the end-to-end latency of the system at ~405 ms.

Before being passed into our grasp detection model, some of the data steams required extra processing. Aside from background subtraction and a normalization step the camera frames were not pre-processed and were passed directly into the model. On the other hand, EMG and accelerometer data benefited from small transformations. For EMG data, we applied a bandpass filter (cutoff frequencies of 20 Hz and 800 Hz) and performed full-wave rectification. The filtering helped remove baseline noise and motion artifacts while rectification improved interpretability of the signals.

Past EMG-based systems have found success with extracting features such as mean absolute value (MAV), RMS, waveform length, and covariance between channels and running these features through a random forest or SVM classifier [1, 16, 43]. We tried this approach but found better success by keeping the EMG signals as a time series and processing them with a 1D-CNN. This may be due to the fact that we were classifying EMG using event-based grasp labels that prioritized transient changes in the data.

For accelerometer data, we used a similar approach to EMG. First, we removed the gravity component from the accelerometer vector. We then took each window and subtracted the min of the window from the entire window and squared the result. The purpose of this transformation was to emphasize transient spikes in the data which is the main accelerometer feature we used for grasp detection.

3.7 Model Architecture

A diagram of our complete system architecture can be seen in Figure 7. This diagram includes the data processing steps described in the previous subsection and the machine learning models described in this one. Contextra's approach to recognizing grasps uses data fusion. In our model we decided to use two layers of fusion. First, the data from the cameras was separated from the EMG and accelerometer data. At their core, the data in these two groups looked at grasp from two different angles. The features from the camera data tried to answer the question "is there currently an object held in the hand?", for which a label is assigned to each frame to indicate if a grasp is occurring or not. On the other hand, for the time-series data, the features tried to answer the question "did a grasp just occur?", for which a label is assigned to an event in time. To simplify processing, we treated both the moment of picking up an object (lift off from rest) and setting down an object (touch down at rest) the same as "grasp events". As a result of this label distinction, one model with frame-based labels was used for camera data while a second model with event-based labels was used for EMG and accelerometer data.

Inside the frame-based model, a 3D-CNN was used. This was found to have better results than a 2D-CNN since it can capture temporal changes in image data, and was found to train more stably than a convolutional LSTM model. Inside the event-based model, separate 1D-CNNs were used for EMG and accelerometer to capture transient temporal changes in the signal traces. After this, the outputs of the 1D-CNNs were jointly fused before additional fully-connected layers. The output from both models were combined using a late-fusion algorithm.

In the algorithm, we use the raw output of our frame model as our base for prediction (before taking its argmax). We then look for transitions in the predicted output of event model from 0 to 1 (we predict an event just started). When found, we add a boost B_e to the index of the frame output corresponding to the class predicted *least* in the last 5 frames. The purpose of this boost is to weight transitions in the frame model predictions when events are predicted by the event model. On top of this, a boost B_m is also given to the index of the frame output corresponding to the class predicted in the last frame. This second boost serves to give momentum to state predictions and

prevent bouncing output. In our evaluation we used boost values of $B_e = 20$ and $B_m = 10$ which were found by experimentation on pilot data.

We used this heuristic because we found (A) the frame-based model was generally more reliable and (B) that transitions in the frame and event-based models were difficult to exactly match in time. As such it made sense to use the weaker event model to influence the transitions of the frame model but ultimately use the frame model for prediction outputs. In addition, the output frame-based predictions were converted to transitions again in our evaluation metric (see Results section) for a similar problem of aligning predicted transitions with ground truth transitions which using an event-based metric allowed us to overcome.

During testing, we found our model's performance improved if pre-trained on a larger corpus of pilot data. This pilot data was separate from the main training data collected on participants during the study. The size of this pilot data was roughly 4 hours of grasp events collected on data from one of the authors and two additional users (not in user evaluation). Our model was pre-trained for 10 epochs on the pilot data then fine-tuned for 40 epochs with user data using the ADAM optimizer and a learning rate of 1e-4.

In total, the model has 613 M FLOPs and 4.07 M parameters. When tested on a Core i7-10750H laptop CPU the model achieved an inference time of 4.4 ms (227 FPS output). For reference, new frames arrive only every 13.3 ms which still gives a 8.9 ms buffer for real time inference. For a mobile deployment we would expect our model to target a SoC such as the Apple Watch S9 which contains a 4-core neural engine. Exact specs for the S9 are not available, but its design is based on the A15 Bionic which has benchmarked an average latency of 9.44 ms (106 FPS) at 72 mW for a 67 M parameter CNN-based model 16x larger than ours [21]. Thus, we are encouraged our model could run at speed and remain low power on a mobile chipset.

3.8 Ground Truth Data

To collect ground truth data during our study, we used a Sensel Morph touch pad [40]. The pad detected when there was a pressure applied to it, which we used as a signal for if the object was currently set down. For the event-based model, we used transition points in the Sensel data (e.g., from pressure to no pressure) to signal when a grasp event occurred. In our study design, we considered a potential issue in that by detecting pressure, the Sensel would not detect the exact moment of the grasp but rather when the object was picked up and set down. Although an important distinction, this ended up not mattering with the technique for evaluating our data on the labels we adapted from Cofer et al. [11] and describe in Section 5. This is because a Sensel pick up transition always happens after a correctly predicted grasp up transition and vice versa for the set down transitions.

We also encountered a challenge when we discovered the outline of the pad (made of reflective metal) was clearly visible in the optical camera data. This worried us that our model may attempt to train on this feature instead of looking at the shape and outline of objects. We fixed the issue by laying a cloth over the pad and table underneath which provided a uniform and undistictive visual pattern.

4 Evaluation Procedure

To evaluate the performance of the system we developed, we conducted a user study on grasp detection accuracy. The goal through the study was to investigate the following research questions:

(RQ1) How well does Contextra recognize grasps of a diverse set of objects?
(RQ2) How much does Contextra's performance degrade if it is used by a user not included in the training data?

Fig. 8. The set of eight objects used in our user study. Objects were chosen in part to parallel prior work and in part to offer a diversity of sizes, weights, and reflectances.

(RQ3) Are there certain types of objects that Contextra performs better or worse on than others?

Our study was run on a group of 8 participants (8 male, 2 female, mean age 27). The study lasted approximately 45 minutes in length and was broken up into three identical sessions. Each session began with the participant putting on the Contextra device. We did not measure participant's hands explicitly, however anecdotally, hand sizes varied widely (large to small) as well as large variations in wrist size and hair thickness on the back of the arm (which could be seen in camera images).

Participants then stood in front of a table on which was a Sensel touch pad and a collection of eight objects (shown in Figure 8). Half of the objects in this set were chosen to overlap with a common set from prior work [43]. The other half was consisted of new objects designed to both provide more challenging targets and align with our envisioned example uses of Contextra.

During the session, a researcher placed an object on top of the Sensel pad, began recording data, and asked the participant to pick up and set down the object 15 times. We did not explicitly instruct users to perform challenging cases such as grasps with no object in the hand or resting the hand on a surface. Nevertheless, participants were asked to experiment with different typical grips they might use to pick up the object, and incorporate natural motions and actions while holding or not holding an object. For example: a participant is not holding an object and starts with their hand at their side. Upon being prompted to pick up a water bottle, they grab it around its base and hold it up to their mouth before setting it back down. Once complete, they return their hand to their side. The purpose of this was to capture grasp poses that users would use in the real-world and not just repetitive artificial grasps.

The goal of this instruction was to incorporate variety in the sensor data that would be more typical of how our system might be used in real-world interactions. While doing so makes the sensing task more challenging, it reflects the reality that users do not always grasp the same objects with the same grip or intensity. Further, a peak in EMG or accelerometer data on its own cannot reliably be assumed to be a grasp when working outside a lab setting.

Once 15 graps were complete, this process was repeated for the remaining seven objects. At the end of the session, participants removed the Contextra band before re-wearing it for the next session. The purpose of this was both to give participants a break and simulate different worn positions of the band typical of normal use.

In our Evaluation, because of the nature of the ground truth sensor used, we only tested for cases where the object was picked up immediately after grasp. While sensor information from the entire grasp sequence (in a 400 ms window) was available to our model, the models were trained

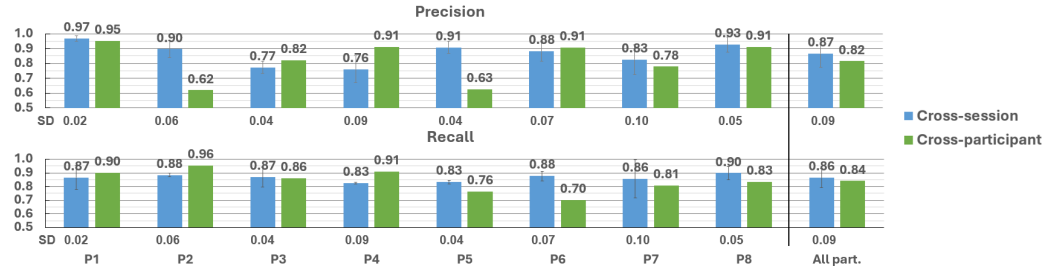


Fig. 9. Average precision and recall were 0.87 (SD=0.09) and 0.86 (SD=0.07) for cross-session and 0.82 (SD=0.12) and 0.84 (SD=0.08) for cross-participant, respectively. Recall tended to be better and lower variance. P2 and P5 had drops in precision in the cross-participant case, potentially from unique physiology that highlights need for larger and more diverse training data.

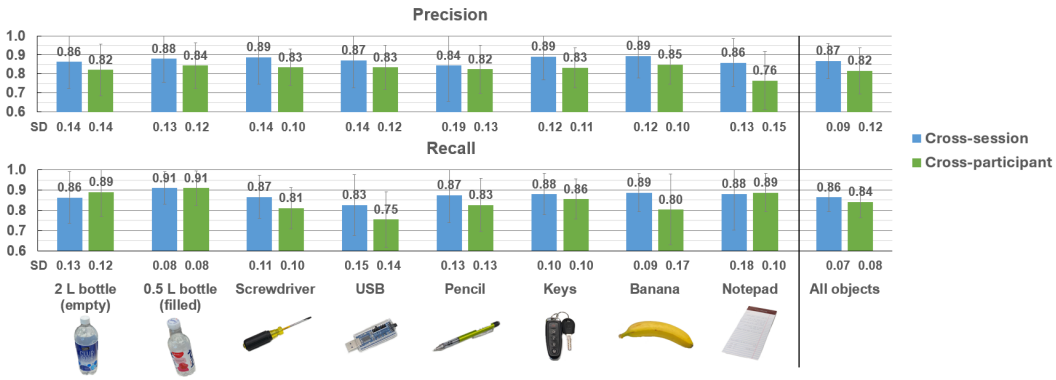


Fig. 10. Precision and recall were fairly consistent across objects, with precision varying only 0.03 from the mean for cross-session and recall only 0.04. Performance was loosely correlated with the size of objects because small ones were harder to image. Small objects like the USB and pencil performed worst while large ones like the bottles and notepad performed best.

to detect ground truth transitions, in other words, the exact moment in time of the grasp. We did not explicitly test cases were objects were grasped but not picked up. These cases may be more challenging due to less IMU motion data, however the IMU acceleration and EMG spike features would still be present and could help maintain performance.

5 Results

In this section we detail the results from our evaluation. For each study, we also examine whether the research questions of interest we outlined in the previous section were met. To conclude the section, we provide a breakdown of performance with respect to different objects tested.

5.1 Calculation of System Performance

In our system, we calculate performance metrics for the percentage of correctly identified grasps, missed grasps, and spuriously detected grasps. We chose this approach because of the fact that our system continuously outputs predictions for a grasp state while our objective is to detect finite grasp events in time, thus measuring accuracy on a frame-by-frame basis is less useful. Before

calculating metrics on our model's outputs, we first pass the predictions through a filter stage that looks for short oscillations between the low (no touch) to high (touch) states. If any state change is detected lasting less than 0.24 s, we determine it to be spurious and filter it out. This cleans up noisy oscillations in our output predictions. The chosen value (0.24 s, 3 window overlaps) was taken from Cofer et al.'s assumption of 0.16 s for minimal press length [11], where we added an extra window overlap of a window (0.08 s) after finding it increased stability.

When a change from the low to high state is detected in either label or prediction outputs, we check to see if the other time series also changed from low to high before the original series returns to the low state. If such a change is found, the event is correctly detected and marked a true positive. If not, the event is marked a false negative if it is initially detected in the label series, and a false positive if detected initially in the prediction series. This is the same technique used by Cofer et al. [11]. Improving their method, we also added extra logic to detect when low states are missed to ensure false negatives and false positives are correctly categorized.

5.2 Cross-Session Performance

To evaluate the results from our study, we performed two analyses. The first of these involved looking at performance across sessions. In this analysis, for each participant, we used a model trained on data from two out of the three sessions and tested on the third session. This process was then repeated for the other two sessions and results were averaged to get final results for each participant.

In our results, we tabulated the number of true positives (a grasp occurred and we predicted a grasp), false negatives (a grasp occurred and we did not predict it), and false positives (we predicted a grasp that did not occur). From these values, we calculated metrics for precision and recall to measure the predictive performance of our system. These results are shown in Figure 9. We chose to use these two metrics because of their widespread usage for comparability and their ability to show how well our system recognizes relevant grasp events. We also did not chose accuracy because we're detecting entire grasp events, not classifying individual frames, and accuracy would be diluted because there are many more no grasp states. Between precision and recall, we are most interested in high recall because our priority is to detect grasp events when they occur. Since we're trying to detect "wake gestures", recall measures how often the "wake" is not missed.

In our cross-session results, we calculated an average precision and recall over all users of 0.87 (SD=0.09) and 0.86 (SD=0.07), respectively. These results confirm RQ1. While recall was fairly consistent across participants, precision varied more widely with some participants having a much greater number of false positive triggers. We believe this behavior may be due to the volume of data we collected in our study. Curious, we tested using a smaller subset of session data and without pre-training and found that our model's precision improved greatly (>15%) for all participants when the size of the training corpus was increased. In the future, we believe this could be improved by collecting more training data of that user, and also pre-training with a larger corpus of data collected by many users. This is in line with past work showing the benefit of large training corpuses for EMG [36].

5.3 Cross-Participant Performance

In our second analysis, we used the same methods as the first one, however this time we were interested in examining how well the system performed when tested on new users it had never seen before. To do this, for a particular participant, we trained our model on data from all three sessions of all other participants and tested the model on all three sessions of the participant in question. The results from this analysis are also shown in Figure 9.

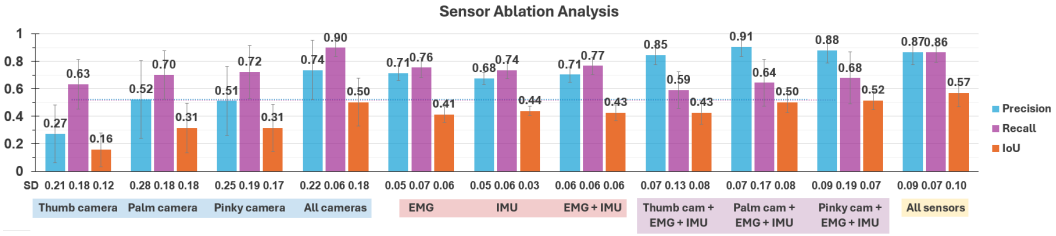


Fig. 11. Ablation analysis for the different sensing channels of Contextra. Results are shown for mean precision, recall, and IoU from the cross-session participant data.

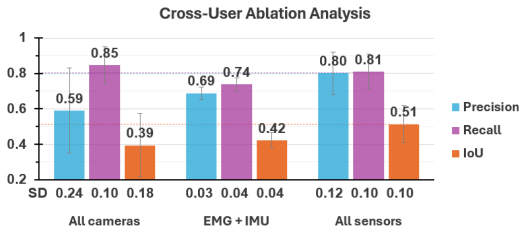


Fig. 12. A subset of the ablation analysis specifically on cross-user data to highlight the advantage of adding cameras to EMG. Results are shown for mean precision, recall, and IoU.

In our cross-user results, we calculated an average precision and recall over all users of 0.82 (SD=0.12) and 0.84 (SD=0.08), respectively. This result is lower than our system's performance for the cross-session case, highlighting the fact that due to the unique physiology of each user, the features they create are slightly different (answering RQ2). Thus, including data from the current user in training will improve the performance for that user. This result is also consistent with prior work using EMG sensor data which showed that although powerful, they tend to have a higher variance across users as compared to alternative sensing methods [14, 32].

5.4 Per-Object Performance

Next, in this analysis we break down performance for the different objects tested. In Figure 10 we show precision and recall for each of the eight objects. To answer RQ3, results were fairly consistent across all objects. For precision, all objects were within 0.03 of the mean for the cross-session case, with the lowest value for the pencil (0.84, SD=0.19) and highest for the banana (0.89, SD=0.12). For recall, all objects were within 0.04 of the mean for the cross-session case and variance was slightly higher with the USB performing the worst (0.83, SD=0.15) and the 0.5 L bottle the best (0.91, SD=0.08). These results are notable because as compared to prior work [7, 43], our system did not have a significant fall off in performance for smaller or lighter objects, indicating more robustness for recognizing grasps of diverse object types.

5.5 Sensor Ablation Analysis

In our final analysis we performed an ablation study on the contribution of each sensing channel to the results of our entire system. For this analysis we used our cross-session data and trained and ran separate models for all major configurations of sensors. Specifically this included breaking out

results for: only one camera, all three cameras, only EMG, only IMU, EMG + IMU, EMG + IMU + one camera, and all sensors.

5.5.1 Precision and Recall Analysis. Mean results for precision and recall are shown in Figure 11. Overall, results were the worst when using one camera. Of the cameras, the thumb camera was the worst performing with a low recall (0.27, SD=0.21) while the pinky camera performed best, albeit with a still low recall (0.51, SD=0.25) but the highest precision (0.72, SD=0.19). This is logical given that the pinky camera has the best view of objects as they are approached while the thumb camera may only seem them, if at all, once picked up. On the other hand, combining all cameras provided a significant boost to precision (+0.22) and recall (+0.18) versus the best single camera (pinky).

For the other sensors, the cases for just EMG, just IMU, and EMG + IMU all performed very similarly (both metrics within 0.03). Of these, EMG + IMU performed best for precision (0.71, SD=0.06) and recall (0.77, SD=0.06). Compared to the all cameras case, EMG + IMU had a slightly worse precision (-0.03), and a much worse recall (-0.13).

We also examined the cases of adding just a single camera to the EMG + IMU case. By doing so, precision significantly increased (+0.20 at best) while recall decreased (-0.18 at worst). The best camera sensor to combine with EMG + IMU was the pinky cam which had nearly the best precision (0.88 SD=0.09) and clearly the best recall (0.68, SD=0.19).

Finally, the best combined average score for precision (0.87, SD=0.09) and recall (0.86, SD=0.07) occurred when all sensors were used together. Compared to the best combined scores for using a single camera (pinky) with EMG + IMU, precision was only slightly worse (-0.01), but recall was significantly better (+0.19).

Comparing other cases, using all sensors significantly boosted precision (+0.13) while only slightly reducing recall (-0.04) versus only all cameras. Compared to EMG + IMU, using all sensors significantly boosted precision (+0.16) and recall (+0.10). These results show adding EMG and IMU to a camera-only model boosted precision and most importantly, adding cameras to a model that uses only EMG + IMU significantly boosted precision and recall. Given prior works [43] use only EMG and IMU to predict grasps, this result confirms the benefit of sensor fusion and shows Contextra's novel contribution of fusing multi-view cameras provides demonstrable benefits for grasp detection.

5.5.2 Intersection-Over-Union Analysis. In addition to precision and recall, we also include a third metric, intersection-over-union (IoU) in Figure 11. This metric is typically used in object recognition but we thought it would be appropriate for examining how well our model overlaps in time predictions of grasp events with ground truth labels. Typically, a value above 0.5 is considered good, however for grasp events of multiple seconds, higher is better. In our ablation analysis we saw IoUs above 0.5 when using all cameras, and EMG + IMU with one or all cameras. In general results for IoU followed the same pattern as the precision and recall ablation results with the best value (0.57, SD=0.10) achieved by fusing all sensors.

5.5.3 Cross-User Analysis. It is well known that EMG sensors have a high variance across users due to differences in physiology. For this reason, we also thought it would insightful to conduct a subset of our ablation analysis on cross-user data to see the advantage of adding cameras to EMG. The results for this analysis are shown in Figure 12. We discussed earlier how recall was the most important metric for grasp detection to not miss true positives. In this analysis we can see clear patterns, mirroring the cross-session analysis, that show using cameras, particularly adding them to EMG + IMU sensors, helps to significantly boost recall along with all other metrics.

In cross-user analysis, we acknowledge a limitation of our evaluation on a small set of 8 participants. We believe the large volume of data (x3-x20 more than related studies [11, 43]) and positive

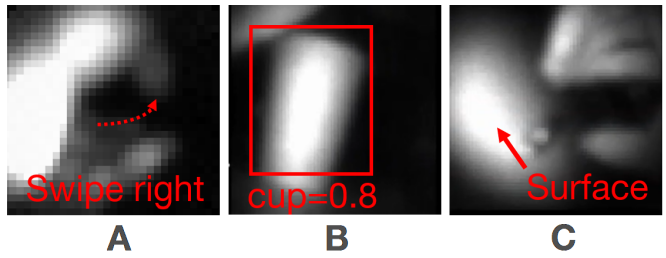


Fig. 13. Examples of future applications for data collected from Contextra along with demonstrative example camera frames. Contextra could be used to detect micro gestures on the hand or objects (A), the identity of a particular object (B), or when a user's hand is near a surface for surface-based interactions (C).

results from these participants were strong enough to show the validity of Contextra's method. However, ultimately we were not able to capture the most diverse set of possible hand physiologies and grasp poses. As such, additional testing on more users would be necessary to fully evaluate the generalizability of this approach.

Comparing Contextra to prior work, we reference the performance columns from the table in Figure 2. Direct comparisons between systems is difficult due to the differing performance metrics, experimental conditions, and tested objects. Nevertheless, focusing on the cross-session and cross-user cases, Contextra exceeds or matches performance of prior systems and uses what we believe to be a more challenging object set (smaller and lighter-weight).

5.5.4 Performance Trade-offs. For full clarity, we highlight some of trade-offs made for this performance improvement. To recap, for power, the EMG and IMU sensors from our band collectively consume 150 mW and each camera sensor consumes 11 mW. For bandwidth, the EMG and IMU sensors collectively produce 126 kB/s of data and each camera sensor produces 66 kB/s. This means by adding cameras precision is boosted by 23% and recall by 13%, but power is increased by 22% and bandwidth by 157%. Different embedded systems may have different budgets, but we believe for the target application of using grasps as a wake gesture for contextual AI, a significant increase in recall is well worth these trade-offs.

6 Discussion

6.1 Future Work for Grasp Applications

In this subsection, we provide discussion of extended applications where Contextra could be used in the future. To recap our rationale behind focusing on grasp, we see detecting grasps as a key building block to enable deeper and more powerful contextual experiences. The ability to continuously and reliably detect when objects are picked up is what makes the following experiences possible.

6.1.1 All-Day Context Sensing for AI. The vision behind AI hardware devices, like the Humane AI Pin [20] and features of the Apple Vision Pro [2], is to eventually have a device that can learn about a user's behavior all day and use that knowledge to provide just-in-time contextual assistance. However, present limitations in the power consumption of such devices prohibit their runtime and privacy concerns may restrict constant use of prominent wearable cameras. Contextra on the other hand, can continuously run because of its privacy-aware and low-power sensors. By using the two devices in conjunction, Contextra can solve problems of current devices by continuously sensing at the wrist to keep up with context info, then waking up the headset's sensors for AI assistance when ready.

6.1.2 Using Objects to Infer Context. Using external devices in a constellation with Contextra has additional benefits for context recognition. Our wristband can know how you are grasping and interacting with an object, while the headset can know what object you are interacting with. More concretely through an example: (1) Contextra detects the user is holding something, (2) a headset is woken up and its cameras are used to determine the user is holding a water bottle, (3) EMG and optical sensors from the wristband are used to determine how the user is holding the bottle: they are opening it then drinking from it. The context is the user is taking a drink of water. Once context is determined, relevant interactions can be enabled, such as having a habit app update a daily hydration tracker.

6.1.3 Additional Uses of Camera Data. When experimenting with the output from our camera setup, we hypothesized that it may be possible to track other interactions that build upon grasp. Examples, along with relevant camera data, are shown in Figure 13.

One of these is micro gestures on the fingers or held objects. Thumb gestures such as pinches, taps, and swipes are a potential source of input for a wearable that does not have traditional buttons or touch screens. Once an object is detected as grasped, Contextra may also be able to track thumb gestures on said object to make the object interactable. Returning to the banana example, if a user picks it up and performs swipes on it, they could cycle through information on a heads-up display like nutrition facts or recipes.

Other potential interactions include performing object recognition directly from the Contextra camera data or identifying micro-contexts of the hand. Such micro-contexts include if it is on a surface, in a pocket, or near the head. All together, these future directions hint at potential for Contextra's sensing stack to be taken beyond grasp and enable new context-centered experiences.

6.2 Limitations

We believe our proof-of-concept prototype demonstrates the feasibility of our multi-view optical fusion method for detecting grasps. Nevertheless, there are limitations to consider. First among these are characteristic limitations of wrist-based optical sensing. When placed on the wrist, cameras can occasionally suffer from occlusion. Two common [11, 13, 19, 43] techniques for offsetting this issue, namely using multiple camera views and fusing data from sensors not affected by occlusion like (EMG and acceleration), were both used in this work. Contextra also has the advantage that even if the fingers are not in view, the object may be able to be imaged. Still, in some rare cases, such as a tiny, light, object that is gently picked up with the fingers straight ahead, detection accuracy may drop.

Unique to our optical sensing technique is also sensitivity to IR backgrounds with a high dynamic range. A reminder: our camera processing uses a subtraction step between the scene illuminated by the IR LED and background IR illumination. Although we only conducted our study indoors, our sensor has an auto-gain feature, and we tested that clear images can still be obtained in sunlight. Challenging cases occur when the user is in a dark environment, which the auto-gain tunes to, but sees a light source or a bright window in the camera view. This can produce small artifacts in the subtracted image which might be mistaken for objects. However, if a future software model was developed to perform background subtraction on raw 60 x 60 camera frames, this algorithm could also take care of these IR artifacts.

This point connects to the final issue on limitations of our camera sensors. First, our cameras are susceptible to occlusion such as long or baggy shirt sleeves. In such cases detection may would still work, albeit with reduced performance, by using only EMG and IMU. Second, some objects are too small or have a surface so dark and matte that they are not well captured by the IR cameras. At their core, the camera sensors work by looking for differences in reflectivity between the objects and

background that correspond to object faces and contours. As such, cases such as a dark, matte object on a highly reflective background could be recognized, but a matte object on a matte background would be more challenging. By looking at the raw 60 x 60 outputs from our image sensors, we know the features of such objects can be made out better when more pixels are available. As such, we expect that if the higher-resolution data were made available, grasp detection of these types of objects would improve, however there would be trade-offs in privacy versus our current low-resolution approach.

6.3 Implications For Practitioners

Many of the insights from this research may have practical implications for practitioners designing context-dependent applications. For one, we have demonstrated the ability to provide always-on context detection at a low power (255 mW) and that conserves privacy for users versus constantly running headset cameras. For practitioners this means they can develop apps that treat event detection as a black box and can focus instead on recognizing events that have been detected by our system.

Our quantitative results showed that grasp detection performed best on larger objects with a larger profile to image while held in the hand. Exploration and testing of our camera sensors also showed us better optical performance on glossy objects and those with high contrast from the background. As such, designers should structure the entry points to their contextual apps around objects with these properties. For example, if tracking calories consumed from a bag of popcorn, reliability would be better for detecting grasp of the bag rather than individual kernels.

Finally, our ablation results showed that while performance was best when fusing all sensor streams, acceptable performance could still be achieved when using a subset of sensors. This is valuable information for practitioners who may want to design apps that could change their wake up expectations dynamically based on sensors available. For instance if a user puts on a sweatshirt with long sleeves that cover their wrist, practitioners could detect the cameras are blocked, switch to EMG+IMU-only mode and lower the grasp confidence threshold outputted from Contextra to wake up their app. This design would ultimately lead to more wake ups than may be needed, but would help to maintain a consistent end-user experience even when some sensors are obstructed.

7 Conclusion

In this work we presented Contextra, a wrist-worn system that uses sensor fusion of EMG, accelerometer, and multi-view IR camera data to detect when users grasp objects. Our approach builds upon prior systems for grasp detection by incorporating a novel late-fusion model that takes event-based processing of EMG and IMU signal spikes and adds frame-based processing of low-resolution IR camera data from three different sensors positioned around the wrist. For the first time, we can detect grasp in a wrist form factor while still being privacy-aware, low-power, and enabling reliable detection across objects. Through a user study involving picking up a diverse set of objects and incorporating natural grasp poses and motions, our system yielded a grasp detection recall of 0.86. Interactions with objects form the foundation of many of the contexts we are in and activities we perform. By creating a device that is able to reliably detect object grasps, our hope is to provide a key component to enable an exciting variety of context-aware interactions.

References

[1] Farshid Amirabdollahian and Michael Walters. 2017. Application of support vector machines to detect hand and wrist gestures using a myoelectric armband. (July 2017). http://uhra.herts.ac.uk/handle/2299/19095 Accepted: 2017-07-26T10:22:36Z.

[2] Apple. 2024. Apple Vision Pro. https://www.apple.com/apple-vision-pro/

[3] Apple. 2024. Connect and charge Apple Vision Pro battery. https://support.apple.com/en-us/117740

[4] Arducam. 2023. Arducam 12MP IMX708 Fixed Focus HDR High SNR Camera Module for Raspberry Pi. https://www.arducam.com/product/arducam-12mp-imx708-fixed-focus-hdr-high-snr-camera-module-for-raspberry-pi/

[5] Arducam. 2023. Arducam OV2640 Camera Module, 2MP Mini CCM Compact Camera Modules Compatible with Arduino ESP32 ESP8266 Development Board with DVP 24 Pin Interface. https://www.arducam.com/product/arducam-ov2640-camera-module-2mp-mini-ccm-compact-camera-modules-compatible-with-arduino_m0031esp32-esp8266-development-board-with-dvp-24-pin-interface_/

[6] Arducam. 2023. Arducam OV7670 Camera Module, VGA Mini CCM Compact Camera Modules Compatible with Arduino ARM FPGA, with DVP 24 Pin Interface. https://www.arducam.com/product/arducam_ov7670_camera_module_vga_mini_ccm_compact_camera_modules_m0030/

[7] Eugen Berlin, Jun Liu, Kristof van Laerhoven, and Bernt Schiele. 2010. Coming to grips with the objects we grasp: detecting interactions with efficient wrist-worn sensors. In *Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction (TEI '10)*. Association for Computing Machinery, New York, NY, USA, 57–64. https://doi.org/10.1145/1709886.1709898

[8] Thisum Buddhika, Haimo Zhang, Chamod Weerasinghe, Suranga Nanayakkara, and Roger Zimmermann. 2019. OSense: Object-activity Identification Based on Gasping Posture and Motion. In *Proceedings of the 10th Augmented Human International Conference 2019 (AH2019)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3311823.3311841

[9] Minjie Cai, Kris M. Kitani, and Yoichi Sato. 2018. Understanding hand-object manipulation by modeling the contextual relationship between actions, grasp types and object attributes. *ArXiv* (July 2018). https://www.semanticscholar.org/paper/Understanding-hand-object-manipulation-by-modeling-Cai-Kitani/9c1b526dc9adba029809e45693f758c2af913f60

[10] Liwei Chan, Yi-Ling Chen, Chi-Hao Hsieh, Rong-Hao Liang, and Bing-Yu Chen. 2015. CyclopsRing: Enabling Whole-Hand and Context-Aware Interactions Through a Fisheye Ring. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. Association for Computing Machinery, New York, NY, USA, 549–556. https://doi.org/10.1145/2807442.2807450

[11] Savannah Cofer, Tyler N. Chen, Jackie Junrui Yang, and Sean Follmer. 2022. Detecting Touch and Grasp Gestures Using a Wrist-Worn Optical and Inertial Sensing Network. *IEEE Robotics and Automation Letters* 7, 4 (Oct. 2022), 10842–10849. https://doi.org/10.1109/LRA.2022.3191173 Number: 4 Conference Name: IEEE Robotics and Automation Letters.

[12] Shrivatsa Deshmukh, Vitthal Khatik, and Anupam Saxena. 2023. Robust Fusion Model for Handling EMG and Computer Vision Data in Prosthetic Hand Control. *IEEE Sensors Letters* 7, 9 (Sept. 2023), 1–4. https://doi.org/10.1109/LSENS.2023.3301837 Conference Name: IEEE Sensors Letters.

[13] Nathan Devrio and Chris Harrison. 2022. DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3526113.3545634

[14] Ethan Eddy, Erik J Scheme, and Scott Bateman. 2023. A Framework and Call to Action for the Future Development of EMG-Based Input in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–23. https://doi.org/10.1145/3544548.3580962

[15] ELP. 2020. ELP High Speed USB3.0 USB Camera 2MP USB Camera Module with IMX291 Image Sensor, 1920 * 1080@50fps Webcam with 100 Degree No Distortion Lens for Android Windows Linux,Plug&Play Webcam [ELP-SUSB1080P01-LC1100] - $60.90 : ELP USB Webcam. http://www.webcamerausb.com/elp-high-speed-usb30-usb-camera-2mp-usb-camera-module-with-imx291-image-sensor-1920-108050fps-webcam-with-100-degree-no-distortion-lens-for-android-windows-linuxplugplay-webcam-p-249.html

[16] Junjun Fan, Xiangmin Fan, Feng Tian, Yang Li, Zitao Liu, Wei Sun, and Hongan Wang. 2018. What is That in Your Hand? Recognizing Grasped Objects via Forearm Electromyography Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (Dec. 2018), 161:1–161:24. https://doi.org/10.1145/3287039 Number: 4.

[17] A. Feldman, E.M. Tapia, S. Sadi, P. Maes, and C. Schmandt. 2005. ReachMedia: On-the-move interaction with everyday objects. In *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*. IEEE, Osaka, Japan, 52–59. https://doi.org/10.1109/ISWC.2005.44

[18] K.P. Fishkin, M. Philipose, and A. Rea. 2005. Hands-On RFID: Wireless Wearables for Detecting Use of Objects. In *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*. IEEE, Osaka, Japan, 38–43. https://doi.org/10.1109/ISWC.2005.25

[19] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (June 2020), 71:1–71:24. https://doi.org/10.1145/3397306

[20] Humane. 2024. Humane Ai Pin | See the World, Not Your Screen. | Humane. https://humane.com/

[21] Apple Inc. 2022. Deploying Transformers on the Apple Neural Engine. https://machinelearning.apple.com/research/neural-engine-transformers

[22] Xianta Jiang, Lukas-Karim Merhi, and Carlo Menon. 2018. Force Exertion Affects Grasp Classification Using Force Myography. *IEEE Transactions on Human-Machine Systems* 48, 2 (April 2018), 219–226. https://doi.org/10.1109/THMS.2017.2693245

[23] Hakan Karaoguz and Patric Jensfelt. 2019. Object Detection Approach for Robot Grasp Detection. In *2019 International Conference on Robotics and Automation (ICRA)*. 4953–4959. https://doi.org/10.1109/ICRA.2019.8793751 ISSN: 2577-087X.

[24] David Kim, Otmar Hilliges, Shahram Izadi, Alex D. Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *Proceedings of the 25th annual ACM symposium on User interface software and technology (UIST '12)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/2380116.2380139

[25] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. Association for Computing Machinery, New York, NY, USA, 321–333. https://doi.org/10.1145/2984511.2984582

[26] Takuya Maekawa, Yasue Kishino, Yasushi Sakurai, and Takayuki Suyama. 2011. Recognizing the Use of Portable Electrical Devices with Hand-Worn Magnetic Sensors. In *Pervasive Computing (Lecture Notes in Computer Science)*, Kent Lyons, Jeffrey Hightower, and Elaine M. Huang (Eds.). Springer, Berlin, Heidelberg, 276–293. https://doi.org/10.1007/978-3-642-21726-5_18

[27] Takuya Maekawa, Yasue Kishino, Yutaka Yanagisawa, and Yasushi Sakurai. 2012. Recognizing Handheld Electrical Device Usage with Hand-Worn Coil of Wire. In *Pervasive Computing (Lecture Notes in Computer Science)*, Judy Kay, Paul Lukowicz, Hideyuki Tokuda, Patrick Olivier, and Antonio Krüger (Eds.). Springer, Berlin, Heidelberg, 234–252. https://doi.org/10.1007/978-3-642-31205-2_15

[28] Takuya Maekawa, Yasue Kishino, Yutaka Yanagisawa, and Yasushi Sakurai. 2012. WristSense: Wrist-worn sensor device with camera for daily activity recognition. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. 510–512. https://doi.org/10.1109/PerComW.2012.6197551

[29] Takuya Maekawa, Yasue Kishino, Yutaka Yanagisawa, and Yasushi Sakurai. 2012. WristSense: Wrist-worn sensor device with camera for daily activity recognition. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, Lugano, Switzerland, 510–512. https://doi.org/10.1109/PerComW.2012.6197551

[30] Takuya Maekawa, Yutaka Yanagisawa, Yasue Kishino, Katsuhiko Ishiguro, Koji Kamei, Yasushi Sakurai, and Takeshi Okadome. 2010. Object-Based Activity Recognition with Heterogeneous Sensors on Wrist. In *Pervasive Computing (Lecture Notes in Computer Science)*, Patrik Floréen, Antonio Krüger, and Mirjana Spasojevic (Eds.). Springer, Berlin, Heidelberg, 246–264. https://doi.org/10.1007/978-3-642-12654-3_15

[31] Christophe Maufroy and Daniel Bargmann. 2018. CNN-Based Detection and Classification of Grasps Relevant for Worker Support Scenarios Using sEMG Signals of Forearm Muscles. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Miyazaki, Japan, 141–146. https://doi.org/10.1109/SMC.2018.00035

[32] J. A. Mercer, N. Bezodis, D. DeLion, T. Zachry, and M. D. Rubley. 2006. EMG sensor location: Does it influence the ability to detect differences in muscle contraction conditions? *Journal of Electromyography and Kinesiology* 16, 2 (April 2006), 198–204. https://doi.org/10.1016/j.jelekin.2005.07.002

[33] Meta. 2023. Meta Quest 3: Mixed Reality VR Headset - Shop Now. https://www.meta.com/quest/quest-3/

[34] Alessandra Moschetti, Laura Fiorini, Dario Esposito, Paolo Dario, and Filippo Cavallo. 2017. Daily activity recognition with inertial ring and bracelet: An unsupervised approach. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 3250–3255. https://doi.org/10.1109/ICRA.2017.7989370

[35] Seungjae Oh, Gyeore Yun, Chaeyong Park, Jinsoo Kim, and Seungmoon Choi. 2019. VibEye: Vibration-Mediated Object Recognition for Tangible Interactive Applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300906

[36] Angkoon Phinyomark and Erik Scheme. 2018. EMG Pattern Recognition in the Era of Big Data and Deep Learning. *Big Data and Cognitive Computing* 2, 3 (Sept. 2018), 21. https://doi.org/10.3390/bdcc2030021 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[37] Tobias Pistohl, Thomas Sebastian Benedikt Schmidt, Tonio Ball, Andreas Schulze-Bonhage, Ad Aertsen, and Carsten Mehring. 2013. Grasp Detection from Human ECoG during Natural Reach-to-Grasp Movements. *PLOS ONE* 8, 1 (Jan. 2013), e54658. https://doi.org/10.1371/journal.pone.0054658 Publisher: Public Library of Science.

[38] Julius Cosmo Romeo Rudolph, David Holman, Bruno De Araujo, Ricardo Jota, Daniel Wigdor, and Valkyrie Savage. 2022. Sensing Hand Interactions with Everyday Objects by Profiling Wrist Topography. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, Daejeon Republic of Korea, 1–14. https://doi.org/10.1145/3490149.3501320

[39] A. Schmidt, H.-W. Gellersen, and C. Merz. 2000. Enabling implicit human computer interaction: a wearable RFID-tag reader. In *Digest of Papers. Fourth International Symposium on Wearable Computers*. IEEE Comput. Soc, Atlanta, GA, USA, 193–194. https://doi.org/10.1109/ISWC.2000.888497

[40] Sensel. 2021. Sensel Morph. https://morph.sensel.com/

[41] Chunyuan Shi, Le Qi, Dapeng Yang, Jingdong Zhao, and Hong Liu. 2019. A Novel Method of Combining Computer Vision, Eye-Tracking, EMG, and IMU to Control Dexterous Prosthetic Hand. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, Dali, China, 2614–2618. https://doi.org/10.1109/ROBIO49542.2019.8961582

[42] STMicroelectronics. 2022. VL53L8CX - Low-power high-performance 8x8 multizone Time-of-Flight sensor (ToF) - STMicroelectronics. https://www.st.com/en/imaging-and-photonics-solutions/vl53l8cx.html

[43] Marian Theiss, Philipp M. Scholl, and Kristof Van Laerhoven. 2016. Predicting Grasps with a Wearable Inertial and EMG Sensing Unit for Low-Power Detection of In-Hand Objects. In *Proceedings of the 7th Augmented Human International Conference 2016 (AH '16)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2875194.2875207

[44] Radu-Daniel Vatavu and Ionuţ Alexandru Zaiţi. 2013. Automatic recognition of object size and shape via user-dependent measurements of the grasping hand. *International Journal of Human-Computer Studies* 71, 5 (2013), 590–607. https://doi.org/10.1016/j.ijhcs.2013.01.002

[45] Anandghan Waghmare, Roger Boldu, Eric Whitmire, and Wolf Kienzle. 2023. OptiRing: Low-Resolution Optical Sensing for Subtle Thumb-to-Index Micro-Interactions. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*. ACM, Sydney NSW Australia, 1–13. https://doi.org/10.1145/3607822.3614538

[46] Edward J. Wang, Tien-Jui Lee, Alex Mariakakis, Mayank Goel, Sidhant Gupta, and Shwetak N. Patel. 2015. MagnifiSense: inferring device interaction using wrist-worn passive magneto-inductive sensors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 15–26. https://doi.org/10.1145/2750858.2804271

[47] Raphael Wimmer. 2010. Grasp sensing for human-computer interaction. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction (TEI '11)*. Association for Computing Machinery, New York, NY, USA, 221–228. https://doi.org/10.1145/1935701.1935745

[48] Raphael Wimmer and Sebastian Boring. 2009. HandSense: discriminating different ways of grasping and holding a tangible user interface. In *Proceedings of the 3rd International Conference on Tangible and Embedded Interaction (TEI '09)*. Association for Computing Machinery, New York, NY, USA, 359–362. https://doi.org/10.1145/1517664.1517736

[49] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M. Kitani. 2020. Back-Hand-Pose: 3D Hand Pose Estimation for a Wrist-worn Camera via Dorsum Deformation Network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 1147–1160. https://doi.org/10.1145/3379337.3415897

[50] Zhen Gang Xiao and Carlo Menon. 2017. Counting Grasping Action Using Force Myography: An Exploratory Study With Healthy Individuals. *JMIR rehabilitation and assistive technologies* 4, 1 (May 2017), e5. https://doi.org/10.2196/rehab.6901 Number: 1.

[51] Chenhan Xu, Bing Zhou, Gurunandan Krishnan, and Shree Nayar. 2023. AO-Finger: Hands-free Fine-grained Finger Gesture Recognition via Acoustic-Optic Sensor Fusing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 306, 14 pages. https://doi.org/10.1145/3544548.3581264

[52] IonuŢ-Alexandru ZaiŢi, Radu-Daniel Vatavu, and Ştefan-Gheorghe Pentiuc. 2013. Exploring Hand Posture for Smart Mobile Devices. In *Human Factors in Computing and Informatics*, Andreas Holzinger, Martina Ziefle, Martin Hitz, and Matjaž Debevc (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 721–731.

[53] Mehrshad Zandigohar, Mo Han, Mohammadreza Sharif, Sezen Yağmur Günay, Mariusz P. Furmanek, Mathew Yarossi, Paolo Bonato, Cagdas Onal, Taşkın Padır, Deniz Erdoğmuş, and Gunar Schirner. 2024. Multimodal fusion of EMG and vision for human grasp intent inference in prosthetic hand control. *Frontiers in Robotics and AI* 11 (Feb. 2024). https://doi.org/10.3389/frobt.2024.1312554 Publisher: Frontiers.

[54] Qian Zhou, Sarah Sykes, Sidney Fels, and Kenrick Kin. 2020. Gripmarks: Using Hand Grips to Transform In-Hand Objects into Mixed Reality Input. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376313