ViSig: Automatic Interpretation of Visual Body Signals Using On-Body Sensors

YIFENG CAO, Georgia Institute of Technology, USA
ASHUTOSH DHEKNE, Georgia Institute of Technology, USA
MOSTAFA AMMAR, Georgia Institute of Technology, USA

Visual body signals are designated body poses that deliver an application-specific message. Such signals are widely used for fast message communication in sports (signaling by umpires and referees), transportation (naval officers and aircraft marshallers), and construction (signaling by riggers and crane operators), to list a few examples. Automatic interpretation of such signals can help maintaining safer operations in these industries, help in record-keeping for auditing or accident investigation purposes, and function as a score-keeper in sports. When automation of these signals is desired, it is traditionally performed from a viewer's perspective by running computer vision algorithms on camera feeds. However, computer vision based approaches suffer from performance deterioration in scenarios such as lighting variations, occlusions, etc., might face resolution limitations, and can be challenging to install. Our work, ViSig, breaks with tradition by instead deploying on-body sensors for signal interpretation. Our key innovation is the fusion of ultra-wideband (UWB) sensors for capturing on-body distance measurements, inertial sensors (IMU) for capturing orientation of a few body segments, and photodiodes for finger signal recognition, enabling a robust interpretation of signals. By deploying only a small number of sensors, we show that body signals can be interpreted unambiguously in many different settings, including in games of Cricket, Baseball, and Football, and in operational safety use-cases such as crane operations and flag semaphores for maritime navigation, with > 90% accuracy. Overall, we have seen substantial promise in this approach and expect a large body of future follow-on work to start using UWB and IMU fused modalities for the more general human pose estimation problems.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Hardware → Sensor applications and deployments**.

Additional Key Words and Phrases: visual signalling, on-body sensors, UWB, IMU, body signals, fallback communication, sports automation, postures, gestures

ACM Reference Format:
Yifeng Cao, Ashutosh Dhekne, and Mostafa Ammar. 2023. ViSig: Automatic Interpretation of Visual Body Signals Using On-Body Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 4 (March 2023), 27 pages. https://doi.org/10.1145/3580797

1 INTRODUCTION

Visual body signals, where individuals communicate or broadcast messages using hand gestures or specific body postures, are important for accurate and unambiguous message delivery in a variety of fields, such as sports, construction, and transportation. It might come as a surprise that equipment worth billions of dollars and lives of millions of people directly or indirectly depend on accurate delivery and interpretation of body signals even today.

Authors' addresses: Yifeng Cao, Georgia Institute of Technology, USA, 266 Ferst Dr NW, Atlanta, Georgia, ycao361@gatech.edu; Ashutosh Dhekne, Georgia Institute of Technology, USA, 266 Ferst Dr NW, Atlanta, Georgia, dhekne@gatech.edu; Mostafa Ammar, Georgia Institute of Technology, USA, 266 Ferst Dr NW, Atlanta, Georgia, ammar@cc.gatech.edu.

2474-9567/2023/3-ART4
https://doi.org/10.1145/3580797

Fig. 1. Visual body signals are ubiquitous; examples from (a) construction industry (b) sports (c) naval operations (d) aircraft marshalling.

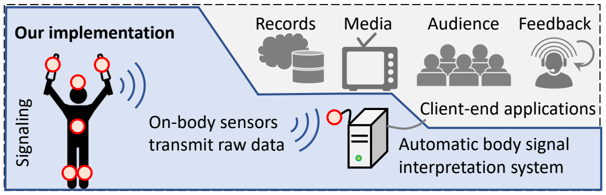

Fig. 2. The overall envisioned ViSig system.

Some examples of body signals are shown in Fig. 1. Visual body signal are directly used by over 600,000 people on a regular basis in the United States alone. According to the 2021 Labor Force Statistics [11], this includes around 360,000 construction equipment operators, around 158,000 aircraft pilots and flight engineers, around 31,000 airfield operations specialists, around 51,000 crossing guards and flaggers, and around 17,000 umpires and referees. There are several reasons why body signals prevail even to this day and age: (1) Body signals are language agnostic and universal; (2) They can be used in very loud environments such as at construction sites and at airports, where audio communication can be challenging; (3) Body signals work even at fairly large distances without requiring a communication channel to be established beforehand, such as in the case of pilots taxiing at an airport.

In this work, we plan to create a system that can automatically interpret the signaller's actions. These interpreted actions can be stored in memory, or used to provide real-time feedback to the signaller, or make the signal interpretation available to the receiver of the signal, depending on the exact use-case. For example, in the case of aircraft marshalling interpreted actions can be stored for record-keeping and auditing purposes which can be particularly helpful during accident investigations. Immediate feedback to the signaller will be helpful in the construction industry to ensure correct signals are being given and to allow quick correction of errors. Finally, broadcasting the signal's interpretation could help automatic score keeping in sports to keep the audience informed based on the umpire signals. While automatic interpretation of body signals would simplify score keeping in sports which might be thought of as a mere convenience, in the context of construction and transportation, it can enable safer operations and avoid catastrophic accidents [1, 23, 45] caused due to signaling errors or misinterpretation. Due to their critical nature and importance, body signals are often standardized by organizations such as the International Civil Aviation Organization [52] and Occupational Safety And Health Administration (OSHA) [65]. While body signals are irreplacable, non-invasive and contact-free technology interventions can increase their effectiveness, and improve record-keeping, accident investigations, and auditing.

We envision a system that captures body signals using on-body distance measurement sensors (see Fig. 2) and transmits their meaning to the intended recipient, be it broadcasting on TV (for sports), or to airplane pilots, or crane operators, or even just as a feedback to the signaling person, as confirmation that their intended signal was indeed conveyed through the action correctly (see Fig. 2). Most existing works in this space attempt to automate interpretation of signals from a *viewer's perspective*, using computer-vision based analysis [30, 80]. In this work, we break with tradition by instead using on-body sensors to interpret the signals—a starkly different approach to solve the signal interpretation problem. We believe, doing so provides significant simplifying benefits: (1) the signals are captured at their source where corruption of the signal is least-likely; (2) the person signaling is already instrumented using on-body sensors, meaning we do not need to identify the signaling person from a crowd as camera-based systems require; and (3) distance from the capturing cameras and occlusions do not hinder detection when signals are captured directly on the person's body.

Others have also proposed on-body sensor solutions to tackle the action recognition problem, most of which concentrate on inertial and magnetic sensors (IMU) [32, 39, 51, 62, 86]. IMUs are indeed useful to obtain orientation of body joints which results in a skeletal estimation of the human pose. However, using IMUs alone is inherently limiting since they can only capture *orientation* of individual skeletal points and not *where* these points are with respect to the rest of the body. For example, an IMU on the wrist cannot distinguish between parallel positions of the hand which are similarly oriented. Furthermore, orientation estimation can be erroneous when the person is moving or near magnetic materials (a common occurrence in locations where body signals are used, e.g., construction sites). Such errors occur because orientation calculations depend on two external forces—the earth's acceleration due to gravity, and the earth's magnetic field—whose measurements can be influenced by the user's movements and other magnetic fields in the vicinity. Therefore, IMUs alone cannot provide a robust solution to identify body signals. To make fundamental innovations in this field, it is important to take a step back and rethink the problem space of visual body signal identification from first principles.

At a high level, body signal interpretation can be thought of as being a subset of the more general human activity recognition (HAR) problem. In fact, identifying body signals is a simpler problem since only a small set of well defined gestures are performed when signaling, and inter-individual variations are minimal. However, *this does not mean that automatic interpretation of body signals is trivial.* We show in this work that several challenges must be overcome to achieve our goal and, in doing so, we embark on exploring new ways of capturing a person's pose. **Challenge 1:** Human signals comprise many different actions made using a combination of independent movements of appendages including hands, legs, fists, and fingers. How can we use a small number of sensors to capture this multi-dimensional action space? **Challenge 2:** All sensors have shortcomings; some sensors allow fast interpretation of the signal, while others are robust over a long period of time, some are affected by the environment while others are affected by the body itself. How can we then produce a solution that identifies signals in real-time, but also works over a substantial amount of time without drifting or accumulating errors? **Challenge 3:** Applications differ in the number and shape of body signals. Can we devise a general software pipeline? How can we enable numerous applications without having to redesign the entire software solution stack from scratch for every application?

We believe the answer to the first challenge lies in a careful choice of on-body sensors. We take inspiration from the simplicity of stick-figure icons (see Fig. 3 for some examples) expressing the signals we wish to interpret, and enlist the broad distinguishing factors between different signals: (1) the *distance* of the user's hands and feet from the torso, (2) the *dynamic motion* performed in some of the signals, (3) the *orientation* of the hand and (4) the *finger configuration.* Taking a first-principles approach we explore the appropriate sensors to detect these distinguishing factors.

To measure the *distance* between various parts of the body, we employ ultra-wideband (UWB) radios, performing wireless ranging with each other at strategic points on the body. This covers the first two distinguishing factors: body-distances and dynamic motions. However, we observe that distance measurements alone are not sufficient to distinguish between certain signals. Consider for example the difference between "raise boom" and "lower boom" in Fig. 3 (crane operator signals). The only difference is the orientation of the hand. We therefore employ

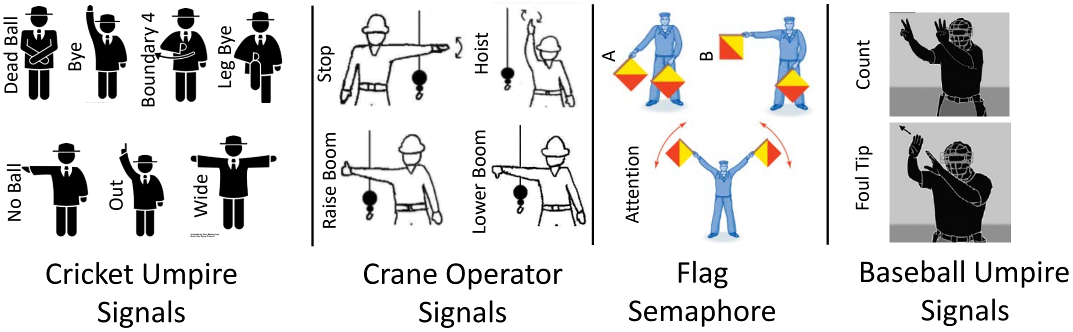

Fig. 3. Example body signals used in sports and the construction industry.

IMUs to obtain the orientations. Finally, gestures of the fingers are not captured by any of the other sensors. We employ photodiodes at the base of the fingers to distinguish between different finger configurations. Together, these sensors address Challenge 1. It is natural to wonder, *will it be cumbersome to wear these on-body sensors?* Paying attention to the *domain* where these signals are used is important to answer this question. The signallers are often already wearing specially designed gear including heavy-duty shoes, helmets, high-visibility vests and so on. Therefore, we believe that the additional light-weight on-body sensors will not be an increased burden for the users.

The combination of sensors we employ work in concert to achieve our desired goals. IMUs and UWBs complement each other when used together. While IMUs are affected by the magnetic noise in the environment, UWB distance measurements remain stable. Similarly, in cases where UWB signals are blocked by the human body and produce erratic distance reading, IMUs can verify that the user is actually static. Finger configurations obtained from photodiodes need to be assessed only in specific poses. UWB and IMUs together act as a filter to enable processing of the photodiode signals only when relevant poses are detected. Further, UWB radios serve a dual-purpose: sensing and communication. UWB wireless packets simultaneously measure distances while also communicating IMU and photodiode values to the outside world. Thus together, UWB, IMU, and photodiode sensors complement each other's shortcomings and address Challenge 2.

Finally, we develop an integrated software pipeline that first sanitizes UWB distance measurements through guidance from IMU sensors, and then extracts a set of features that describe both static poses and dynamic movements with a neural network. Our learner trains on domain specific body signals to appropriately weigh input features, resulting in a class label. Challenge 3 is addressed by decoupling the domain specific class labels and training, from the overall data sanitization and feature extraction. Further, the wearable sensors that we have designed can be thought of as personal belongings. While the underlying mechanisms for interpretation of body signals must be general, the user's sensors can be trained to interpret signals from just their own user, just like personal assistance tools such as Amazon Alexa, or Google Home, are routinely trained to interpret the home owners' voices [42, 56].

We call our on-body system for interpreting visual body signals, ViSig. Our contributions are:

(1) A unique sensor system that captures body signals.
(2) A fast software pipeline to obtain pair-wise distance measurements and transmit them along with IMU and photodiode data to the outside world.
(3) A machine learning based fusion algorithm to detect and identify the obtained body signals based on UWB, IMU, and finger configuration data.

The rest of this paper is organized as follows. We first present a primer that describes background material on IMU orientation and its issues, and presents a background on UWB ranging. We then discuss ViSig's system design starting with a design overview, followed by the specifics. We then describe our implementation using custom built UWB + IMU sensors and flexible PCBs for finger signals, followed by our experimental setup and evaluation through an IRB approved user study.

2 PRIMER: IMU, UWB BACKGROUND

To achieve the desired performance in body signal interpretation through wearable sensing, it is important to fully understand how various body signals are differentiated and the limitations of current solutions.

In the field of anthropometry, a visual body signal is uniquely defined by state parameters including angles, velocity, acceleration, rhythmic pattern, space envelope, etc., of joints, bones and muscles of the human body [25]. Because of the complexity of obtaining all the state parameters, current solutions simplify it to only observing the orientation of a few skeleton joints of the human body by mounting IMUs at strategic points on the body. Of course, it is not desirable to exhaustively mount sensors on every joint since it makes the system quite cumbersome. Therefore, one of the research focuses in this area is to reduce the number of mounted on-body sensors. For instance, [33, 78] have made it possible to recover simple human poses with only 6 IMUs on body, improving upon another work [61] that uses 17 IMU sensors. However, we argue that with sparse sensors, an IMU-only solution is not sufficient to perform visual signal interpretation effectively. To verify this claim, it is essential to understand how IMUs help to understand the human pose.

2.1 Inertial Measurements

A 9-DOF IMU includes accelerometer, gyroscope, and magnetometer which measure acceleration, angular velocity, and magnetic strength in its local reference frame (LRF). Because the measured acceleration is influenced by gravity, i.e., a static IMU should observe an acceleration equal to the earth's acceleration due to gravity (g_0), the direction of g_0 can function as an external anchor to calculate the orientation of IMU in the global reference frame (GRF). Similarly, as the magnetic strength is influenced by the geomagnetic field m_0, the direction of m_0 functions as the second anchor. With two anchors, the orientation of IMU in GRF can be easily obtained through vector orthogonalization. IMU-only based signal interpretation derives orientation via the above process for each mounted IMU, and uses the obtained orientation to infer the signal label. However, we find that such inferring process faces two issues: orientation error, and body signal ambiguity.

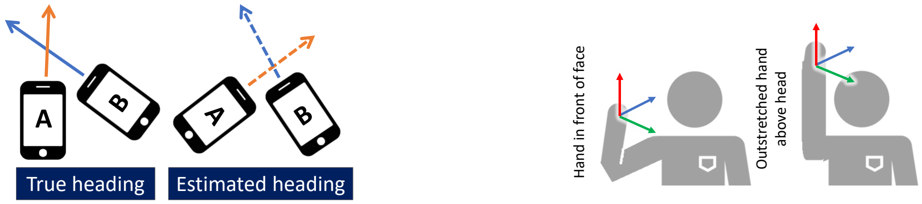

Fig. 4. (a) Wrong heading because of the magnetic field. (b) An IMU-only approach to pose estimation can lead to ambiguity.

2.1.1 *Orientation Error.*

The above process for calculating orientation in GRF is based on the theoretical assumption that the measured acceleration and magnetic strength is only influenced by gravity and geomagnetic field, which does not hold in practice. Due to environmental influence on magnetometers, the calculated orientation has a non-negligible errors, especially at construction sites or at airports where heavy machinery is commonplace. ViSig ignores magnetometer readings and only relies on the gravity vector to avoid corruption due to magnetic fields.

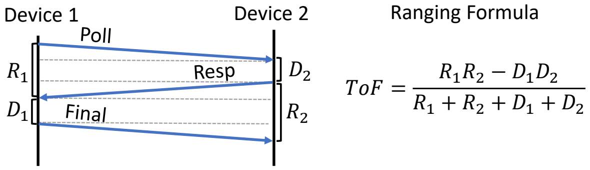

Fig. 5. Two way ranging protocol [34, 47].

2.1.2 Signal Ambiguity. Another issue of IMU-only approaches is the inherent ambiguity in describing the human pose with sparse IMUs. An IMU-only approach assumes the position of body joints can be uniquely determined by the IMUs, which does not hold when IMUs are sparse. For example, with a single IMU mounted at the wrist, one cannot tell the difference when raising the hand overhead and placing the hand beside the torso if the hand is upright in both cases (see Fig. 4(b)). To overcome these hurdles in use of IMUs, ViSig couples acceleration from IMU with distance measurements using UWB. We describe UWB ranging next.

2.2 UWB Ranging

A pair of UWB nodes can measure the in-air wireless time of flight (ToF) and multiply it by the speed of light to determine the distance between them. Benefiting from its 1GHz bandwidth, UWB generates extremely narrow pulses which can measure ToF at a sub-nanosecond level, making it possible to range with only decimeter-level error [36]. However, in spite of the high precision in measuring ToF, the inherent clock difference between two UWB transceivers would yield a $10ms$–$100ms$ error, unless the clock offset and drifts are compensated. Intelligent ranging schemes are devised to perform this compensation, which we describe next.

To remove the effect of clock difference, a pair of UWB nodes perform asymmetric double-sided two way ranging (TWR) as specified in IEEE 802.15.4z [3, 34]. The mechanism of TWR is shown in Fig. 5. Specifically, Device 1 will send a Poll message to initiate a ranging request. When Device 2 receives the Poll message, it will reply with a Resp message to complete one round of ranging. In this process, Device 1 records the time duration R_1 between Poll sending and Resp reception, and Device 2 records the time duration D_2 between Poll reception and Resp sending. Finally, Device 1 will reply with a Final message to complete a second round of ranging. Device 1 records the time D_1 between Resp reception and Final sending, and Device 2 records the time R_2 between Resp sending and Final reception. Together the two rounds eliminate clock offsets and clock drifts [47] using the equation shown in Fig. 5.

When more than one UWB device pair is involved, this process must be repeated for each pair. In ViSig, we use an improved approach to significantly increase the ranging update rate in ViSig.

3 SYSTEM DESIGN

3.1 Design Overview

ViSig aims at *using a small number of on-body sensors that capture sufficient information to understand the position and the orientation of a few key body joints.* These positions and orientations lead to identification of the body signals. The core innovation in ViSig is to estimate the body pose through direct distance measurements between a few points on the body and fusing these with inertial orientation and light sensing. The distance measurements help localize the hands (2 body-joints at the wrists), feet (2 body-joints at the ankles), and the head (1 end-point on top of the head), with respect to the torso (1 body-joint on the waist) representing an approximate frame of the body through a total of 6-points. The choice of these 6 points is not arbitrary, and neither is the number of points we choose to capture—5 of the points capture the extremities of the human body with 1 central reference.

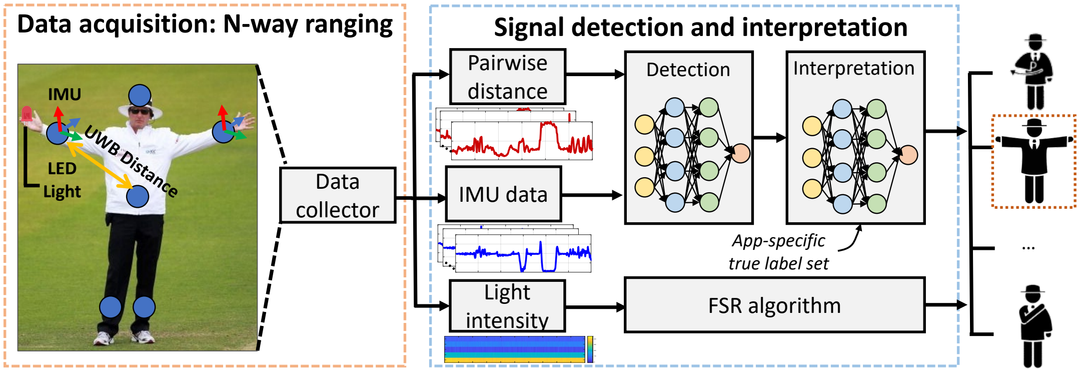

Fig. 6. A system overview of ViSig.

Through pair-wise ranging, we obtain 15 non-symmetric pair-wise distance measurements providing sufficient information to localize all the points with respect to the central reference in 3D.

All distances are measured using UWB-based wireless ranging. We thus obtain a distance matrix which measures the inter-node distances. Our choice of UWB allows ViSig to accurately calculate the inter-node distance at decimeter-level precision [53], owing to the large bandwidth signals employed by UWB radios. Since the inter-node distances and per-joint positions are highly correlated, the distance matrix serves as a descriptive feature to understand the per-joint position. The introduction of the distance matrix makes ViSig inherently different from existing solutions, which rely solely on IMUs for determining human pose. Compared to IMU-only solutions, ViSig addresses the ambiguity introduced when the same orientation occurs at different positions, and is robust to the error caused by magnetic field variations. Of course, whereas ViSig offloads the *position* inference from IMUs to UWB, IMU is still useful in determining the *orientation* of the joint. Hence we mount one IMU on each wrist. Note that IMUs and UWBs on the wrist are co-located hence no additional space is required for mounting IMUs. When using IMU, ViSig **does not use the magnetometer for computing wrist orientation** since the magnetometer can be easily corrupted by nearby ferromagnetic materials, particularly in the transportation and construction domains, due to prevalence of large metal objects, motors, and electromagnets. Finally, neither UWB nor IMU can estimate finger positions, which are important in certain applications, such as differentiating between "Bye" and "Out" in cricket umpire signals. To interpret finger signals, we mount photodiodes on fingers which capture varying amounts of ambient light depending on the finger positions.

Fig. 6 shows data acquisition and processing pipeline of ViSig. At a high level, the design of ViSig can be split into data acquisition phase and signal interpretation phase. In the data acquisition phase, ViSig fuses the raw data of different modalities and transmits it to the processing computer in real-time. However, streaming data in real-time is non-trivial because the standard TWR used by UWB incurs significant delay. While the individual UWB packet rate can be as high as 300 Hz, delay increases quadratically as the number of nodes increase. To tackle this issue, ViSig adopts N-way ranging [28] to re-use the timings of previously received packets for ranging. Such a protocol is particularly applicable for ViSig since the number of UWB devices performing all-to-all ranging is fixed and these devices are all nearby. In the body-signal interpretation phase, the distance matrix and the IMU data are first fed into a body signal detection module. This module functions as an on/off switch or a filter which judges if there is a meaningful body signal in the current data stream. If this module returns "True", the distance matrix and the IMU data will be further fed into a model consisting of LSTM layers and fully-connected layers for primary signal interpretation. Of course, as the fingers cannot be sensed by the UWB or the IMU, there remains ambiguity in the finger-related signals. To remove this ambiguity, ViSig adopts a finger signal recognition (FSR) algorithm to distinguish the signals with the same body pose but different finger states. Finally, the signal interpretation phase generates a vector of probabilities that the given data corresponds to a certain signal and returns the signal with the highest probability as the output.

3.2 Fast Data Collection: N-way Ranging

Real-world signal interpretation applications have a strict real-time demands. ViSig combines together the distance from UWB transceivers, IMU data and light intensity from photodiodes as data streams, in which UWB communication is the bottleneck for time efficiency. In this section, we first discuss the time efficiency issue of standard TWR for multiple UWB pairs. Then, we show how ViSig employs the idea of N-way ranging, briefly introduced[1] by [28], as an N-way time transfer method (NWTT), to significantly reduce the time delay.

As discussed in Section 2, the standard TWR requires 3 messages to perform ranging between one device pair. ViSig requires the computation of the 3D locations of 6 on-body UWB devices. Performing all-to-all distance measurements creates 15 independent equations allowing us to locate all devices in 3D space with respect to a central node on the waist. Standard TWR would require a total of $3n(n-1)$ messages to obtain a fully-filled distance matrix for n nodes. For a 6-node UWB transceiver system, the average time it takes to obtain this distance matrix with standard TWR is $\approx 316\, ms$ $(3\, Hz)$; too slow to extract motion features of some dynamic body signals.

To increase the update rate, we require a ranging protocol that will reduce the number of messages exchanged to perform all-to-all ranging. While this is a difficult problem in general (due to wireless reachability, collisions, etc.), in our case, where the number and IDs of nodes are fixed, and nodes are close to each other on the user's body, we can exploit time division multiplexing (TDM) allowing UWBs take turns in transmitting the data. Given the feasibility of TDM, we adopt NWTT to significantly reduce the ranging time. NWTT makes two key improvements: (1) Instead of specifying a certain receiver ID, each UWB node broadcasts its message to all other nodes in its turn. (2) When transmitting the message, NWTT does not explicitly specify the message type (Poll, Resp, Final). The reception and sending time of every message will be recorded and communicated in the ranging system. Fig. 7 shows a typical timing diagram of NWTT. We use $Msg(SrcId, Seq)$ to identify a message in the system. Node 1 broadcasts a message $Msg(1,1)$ to all the nodes. When the other nodes receive $Msg(1,1)$, the next node 2 will take its turn to broadcast a message $Msg(2,1)$. Note that $Msg(2,1)$ will serve as both a reply to the received $Msg(1,1)$ to node 1 (like a Resp), and also a new Poll message to other nodes. When all the nodes send out a message, one round is complete at which time each pair should complete Poll-Resp round of ranging. Then node 1 starts a new round by sending a new message $Msg(1,2)$. Similar to the previous round, $Msg(1,2)$ serves as both the Final message to previously received $Msg(2,1)$, $Msg(3,1)$,$\cdots$ and a new Poll message. By now, node 1 has completed the full two way ranging with all the other nodes. For example, $Msg(1,1)$, $Msg(2,1)$ and $Msg(1,2)$ completes the two way ranging between node 1 and 2.

The time complexity of ranging with NWTT is $O(n)$ as it takes only n messages to perform all-all ranging on average, far more efficient than naively performing TWR between every device-pair, which has $O(n^2)$ time complexity. By applying the N-way time transfer method, the theoretical update rate increases from $3\, Hz$ to $45\, Hz$. This enables collecting sufficient data to accurately classify body signals in real-time. In our experiments, we set data rate to 16Hz to balance real-time performance, power consumption, and compensate for the longer UWB packets when they carry IMU data.

3.3 Body Signal Detection

We now present the system design for *automatic* detection of body signals, which demarcates the time a body signal occurs. In many application domains, additional mechanisms such as whistle blowing in football, or a

[1] [28] provides a sketch of NWTT in slide format. To fully deploy the protocol, we had to fill in the gaps in its description. We describe details of NWTT implementation for completeness. While we do not claim NWTT as our contribution, details of the deployment we provide are not available elsewhere in the literature as far as we can tell.

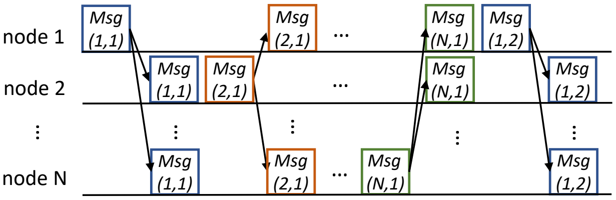

Fig. 7. N-way time transfer method for ranging.

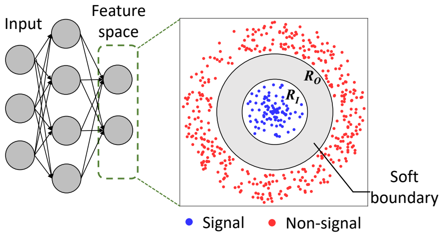

Fig. 8. ViSig uses a soft boundary buffer to separate signals and non-signals.

button press is already used for demarcating a body signal. Those remain applicable in ViSig as well, but we wish to explore the harder problem of automatic body signal detection, only based on the sensors we use.

The distance matrix, IMU data, and light intensity collected by each sensor are streamed to a processor continuously as time-series data. As the raw data are collected continuously, the data stream may contain many irrelevant non-signal actions; we need to first determine the existence of a valid body signal and extract it from the data stream. However, such signal detection is challenging as the *space of irrelevant actions is unbounded*. It is extremely difficult, if not impossible, to include every possible non-signal action in the dataset to make the model robust. Simple binary classification with a learner leads to high false positive rate. One set of solutions to address infinite space of unknown non-signal actions is novelty detection techniques, such as one-class SVM [66] and support vector data description (SVDD) [63, 72]. They employ a hyperplane or a hypersphere to make the space of positive class as compact as possible. One issue in employing these techniques is that they train on only positive classes. However, in some body signal applications, signals can have variance. Therefore, using one class to train the model in body signals produces significant false negatives. For instance, to signal "out" in cricket, the umpire needs to raise an arm and stretch out one finger, but there is no strict rule about how high the hand should be raised. An unsupervised novelty detection model can easily mis-classify such variations.

To address this issue, ViSig employs a supervised neural network to train a binary classifier to perform signal detection. We follow the idea of SVDD to use a hypersphere to separate different classes. The time-series distance matrix and IMU data are firstly fed into an encoder which learns useful feature representations of the raw physical data. This encoder tries to separate signals and non-signals based on their distances to the origin: signals are closer to the origin while non-signals are farther. To avoid high false positive rate in unseen non-signal actions, ViSig creates a soft boundary buffer between body signals and non-signal actions in the encoded feature space (as shown in Fig. 8). The purpose of this soft boundary buffer is to leave space for unseen non-signal actions or body signal variations. The model is trained to keep all the body signal samples inside the inner boundary and all the non-signal samples outside the outer boundary. Specifically, for input time-series data x in the input space $\mathcal{X}$ and its label y indicating whether this is a signal, let $\phi(x; w, y) : \mathcal{X} \rightarrow \mathcal{F}$ be a temporal neural network which encodes the raw data to some output feature space. The objective of ViSig's signal detection is

$$\min_{w} \sum_{i=1}^{n} (\max\{\phi^2(x; w, y = 1) - R_I^2), 0\}) + (\max\{R_O^2 - \phi^2(x; w, y = 0)), 0\}), \tag{1}$$

where R_I, R_O (satisfying $R_I < R_O$) are the inner radius and outer radius of the soft boundary buffer. Once the model is trained, signals will mostly aggregate inside the inner boundary while non-signal actions will be distributed outside the outer boundary. ViSig uses an application-specific threshold ρ to judge whether a time-window sample incorporates a valid signal.

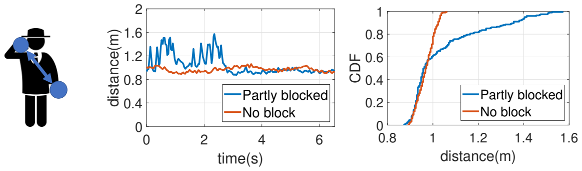

Fig. 9. "short run" cricket signal (left); distance measurement: Time-series (middle), and CDF (right).

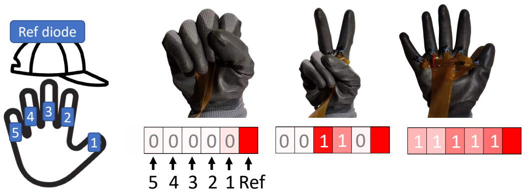

Fig. 10. Light intensity of different finger signs.

3.4 Body Signal Classification

After ensuring the existence of a body signal in the current data stream, ViSig will segment it into fixed-length sliding time windows, and feed them into a classification model. However, *should we feed all the data (distance matrix, IMU, light intensity on finger) directly to one classification model?* In an application-oriented view, the interpretation of body signals is frequently a two-part process: (i) determine the position and the orientation of main body joints and (ii) track finger pose to resolve the ambiguity if two signals are only different in finger configuration. Thus, on-finger light intensity is separate from other inputs. In this section, we will firstly focus on how to obtain a classification with only UWB and IMU data.

3.4.1 Avoiding Distance Errors Due to Body Blocking. The typical short-distance ranging precision of UWB in a line-of-sight (LOS) scenerio is about 10cm [36]. For our application domain, in theory, as most body signal sets ensure sufficient inter-class pose difference, 10*cm* error in the distance should suffice for the classification task. However, in practice, the human body can cause strong radio-shielding [54], effectively blocking the direct signal path between a pair of UWBs. When blocking occurs, the measured distance is actually the in-air distance of *first multipath*, which is larger than the ground-truth direct distance. For instance, in the "short run" cricket signal (see Fig. 9), the distance between left wrist and right wrist is occasionally blocked by the torso. This leads to an overestimation of the distance since UWB receivers detect a later path and treat it as the first path.

We ask the question: *How can we correct these intermittently incorrect distance measurements in our specific setting?* Two key observations are likely to help perform these corrections: (1) any real change in distance can only result from physical movement which should be corroborated by the IMU sensors; (2) the measured distance can never be lower than the direct path, since this is the fastest that wireless signals can travel. These observations lead to a detect-and-then-correct approach. We observe that a key feature indicating signal blocking is intermittent distance overestimation as shown in Fig. 9, which results from the near-threshold direct path power and slight joint motion when performing a ranging measurement. To avoid confusing these intermittent distance overestimations with a dynamic signal which inherently has distance variance (e.g., "deadball" in cricket umpire signal), ViSig first checks the acceleration in the IMU data to classify the body signal as static or dynamic. Then ViSig calculates the distance variance, $\mathrm{var}(d_{ij})$, between UWB nodes i, and j, in a fixed time window for static body signals. If $\mathrm{var}(d_{ij})$ is larger than a threshold, ViSig flattens the top 50% samples to the mean distance in this time window, based on the key observation (2) above that direct path-length is always smaller than any multi-path.

3.4.2 Signal Interpretation Model. Then UWB distances and IMU are fed into signal interpretation model. Only the streaming data that have been classified as containing a body signal by the signal detection module, will be considered for classification. As mentioned before, one of the key benefits of fusing distance and IMU features is that joint features expose more details to understand the skeletal pose than using either of the modalities alone. Therefore, compared to existing work [14, 29, 51], ViSig uses a much more simplified architecture to extract features. Specifically, ViSig employs a neural network composed by a Long Short-term Memory (LSTM) layer followed by two fully-connected layers. The hidden dimension size of the LSTM layer is set to 128. This LSTM layer is used to extract temporal features. The outputs of the LSTM layer are then fed into two fully-connected layers to be encoded into feature space. A softmax layer is concatenated at the end to output a k-dimension vector which gives the probability that the given input describes a certain body signal (k signals in all). We use mean-square-error (MSE) of the predicted class as the loss function.

3.5 Finger Signal Recognition

Apart from the body signals identifiable through the 6 key body-points, many body signals also use different finger signals to further differentiate signals. For instance, a baseball umpire will use the fingers to count the number of balls and strikes pitched for the current batter. In the crane signal, one will open and close the four fingers to additionally indicate "lower/raise the load". However, finger configuration is not covered by either the UWB or IMU modalities. *How can we correctly detect the motion of the fingers without incurring too much physical overhead?* Recent advances in wearable finger tracking includes embedding IMU, flex or EMG sensors [6, 8, 10, 43] to sense the motion of the fingers. While existing solutions are successful in tracking fine-grained finger motion, they rely on complicated models for training and regression on large amounts of observed data, or they need expensive on-finger sensors to detect the state of the fingers.

To simplify the process of finger signal recognition, we make an observation that finger signs that we are concerned with can be reduced to understanding the “stretched” and “closed” states of each finger. For example, finger sign “two” is usually represented by stretching out index finger and middle finger. Based on this observation, we mount 5 tiny photodiodes at the root of each finger. Due to its miniaturized size, the photodiodes can be either mounted on gloves or embedded on accessory rings without affecting the user experience. The principle behind using photodiodes is quite intuitive: when a finger is stretched out, the photodiode will receive more ambient light than when a finger is closed. By establishing a threshold on light intensity, we can judge whether a finger is closed or not. Of course, different lighting conditions can cause any fixed threshold to fail. Therefore, we also attach an additional photodiode on the head [2] which is used as the reference ambient light level (I_{ref}). As the head area is less likely to be blocked, the on-head photodiode can reliably track current ambient light. Fig. 10 shows the light intensity of each photodiode (I_{fin}) when performing "zero", "two" and "five" with finger signs. Evidently, photodiodes receive little light when the fist is closed for signalling zero. For “two” and “five”, we can observe an increase in the light intensity of corresponding stretched fingers.

With the on-finger and on-head photodiodes, ViSig performs finger signal recognition as follows: ViSig firstly takes the label output from the process in Section 3.4. If multiple body signals correspond to this label, ViSig calculates relative light intensity $\frac{I_{fin}}{I_{ref}}$ for each finger to form a quintuple. Then ViSig discretizes each finger’s state into a binary array $L = \{0, 1\}_{i=1}^{5}$ with empirically-set threshold and determines the final output based on the minimum hamming distance from known finger configurations. Note that here the threshold varies under different ambient light. We empirically calculate the optimal threshold setting under different light intensity in Section 5.5. For the tested signal sets in this paper, binary finger state is sufficient to distinguish different visual signals. Of course, signal sets heavily dependent on finger poses (e.g., ALS sign language [92]) may require more sophisticated models for classification. Such applications are out of the scope of this work. One final challenge remains due to the large dynamic range of lux values in full sun compared to those indoors. We tackle it by simply reducing the incident light using a translucent tape on the photodiode when used in bright sun.

4 IMPLEMENTATION

Our real-world prototype comprises 6 UWB DWM1000 nodes [4], each connected to a Cortex M0 microcontroller. These 6 nodes are mounted on the waist, left wrist, right wrist, left ankle, right ankle, and the head, as shown in Fig. 11. The two nodes on the wrist also host an ISM330DHCX 6-DOF inertial sensor (accelerometer+gyroscope). We calibrate the distances for UWB device by minimizing the Euclidean Distance Matrix error [5]. For IMUs, we

[2] This additional photodiode can be packaged with the head-mounted UWB/IMU sensor required for body signal classification.

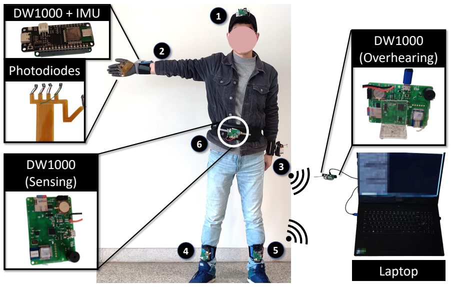

Fig. 11. The ViSig prototype.

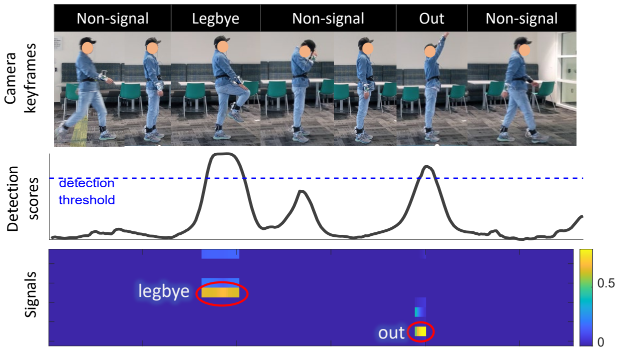

Fig. 12. ViSig's processing pipeline in action on a real-world like umpire movements scenario.

calibrate the gyroscope readings by removing the zero bias when the sensor is static. Additionally, we mount 5 photodiodes on a pair of gloves to detect stretched or closed state of the fingers. The photodiodes are connected to the on-wrist sensor through a custom manufactured flexible PCB specifically designed for this work (will be open-sourced).

UWB nodes, IMU, and photodiodes simultaneously collect and stream data to the microcontroller. Because IMU and photodiodes stream much faster than UWB, we first perform resampling to align the data rate for each modality. UWB nodes have the slowest data rate due to the overhead in performing NWTT. We set the data transmission rate to 110Kbps, and center frequency to $4\,GHz$. Under this setting, the time to transmit a packet is approximately $3ms$. In the practical implementation, we additionally insert a guard time interval of $7ms$ to ensure transmission reliability and allow for processing delays. Hence, it takes $6 \times (3 + 7) = 60ms$ to perform a full all-all ranging with our NWTT protocol, resulting in a practical data collection rate of 16Hz. For IMU, we set the sampling rate to $80Hz$. We embed 5 consecutive IMU entries in a single UWB packet so that the data rate can be matched. For the photodiodes, we embed only one entry in a UWB packet for each photodiode. All the data are collected by the M0 microcontroller and communicated through UWB to each other. These packets are also overheard by a nearby UWB eavesdropper device which does not participate in the all-to-all ranging, but just collects data. This eavesdropper device is connected to a laptop to stream received data and feed through the ViSig pipeline. This way, all on-body sensors send their information to a single overhearing UWB device in the vicinity; UWB is both a communication vehicle as well as a sensing signal performing the ranging operation. While this eavesdropper is placed at about $2m$ from the user in our implementation, it can also be placed on body in which case it can forward received packets to an edge device over wireless links with longer range. Our overall setup is pictured in Fig. 11. We intend to open source body pose data in this work.

5 EVALUATION

We evaluate ViSig on 5 popular body signal applications:

- **Cricket umpire signal:** Cricket is an outdoor bat-and-ball game prevalent in several Commonwealth countries [82]. The umpire of a cricket game uses arm motion to signal key events in the game. There are 11 major signals in the official cricket signal set [69].
- **Baseball umpire signal:** Baseball is an outdoor bat-and-ball game with wide popularity all over the world. A baseball umpire is responsible for giving signals to indicate the start and the end, enforce the rules, and handle unsportsmanlike conduct in a game. We use the signals in NFHS [7] in our experiment, which includes 13 signals. Umpires also signal using the fingers in this sport.
- **Crane signal:** Crane signals are widely used in construction to give instructions to the crane operator. There are 20 signals in the crane signal set [65], standardized by Occupational Safety and Health Administration (OSHA).

Table 1. Signal detection accuracy in non-signal-well-defined datasets and non-signal-undefined datasets.

Signal Set	Softmax classifier (WD)	Deep-SVDD (WD)	ViSig (WD)	Softmax classifier (UD)	Deep-SVDD (UD)	ViSig (UD)
Cricket umpire signal	99.0%	80.0%	**98.5%**	79.7%	85.7%	**91.9%**
Baseball umpire signal	99.3%	84.9%	**99.7%**	99.1%	88.9%	**99.3%**
Crane signal	98.1%	77.2%	**98.7%**	96.7%	90.9%	**97.3%**
Flag semaphore	97.7%	89.3%	**98.7%**	85.4%	75.0%	**95.1%**
Football official signal	90.6%	76.0%	**87.2%**	88%	88.2%	**88.7%**

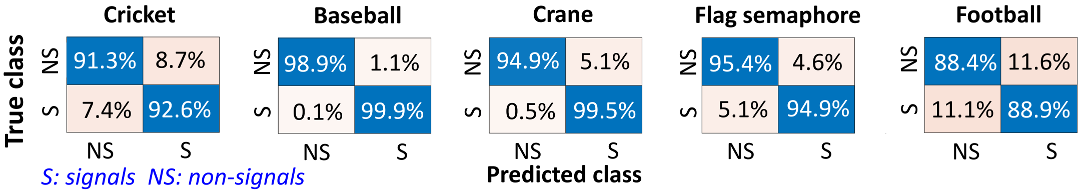

Fig. 13. Confusion matrices of signal detection performance on UD-dataset in each application.

- **Flag semaphore:** Flag semaphores were used to convey information, particularly by navies world-wide [81], and are still used today for underway replenishments [83] and emergency communication, (and in a humorous April 1st RFC by IETF [31]). Most signals in the flag semaphore are static and do not involve the fingers. There are 30 signals in the contemporary flag semaphore system [81].
- **American football official signal:** American football is the most popular sport in the US. An official needs to use body signals to indicate the start and the end of a game, fouls, illegal action, and a wide range of other incidents. According to NFHS, there are 47 different signals in football [9]—this large number makes classification challenging.

In our evaluation, we first present an end-to-end overview of ViSig performance in Section 5.1. Then, we evaluate the performance of ViSig in body signal detection (Section 5.2), body signal classification (Section 5.3), and finger signal recognition (Section 5.4). Then we provide microbenchmark evaluation in Section 5.5. Finally, we compare ViSig with state-of-the-art visual systems in Section 5.6.

5.1 ViSig Pipeline Overview: Signal Detection and Classification

We first show an example case of how ViSig performs body signal detection and classification in a realistic setup. A volunteer is asked to wear the device and perform actions as if he/she is a cricket umpire. This includes two body signals (leg-bye and out), interspersed with other random non-signal actions. In this test, ViSig first uses the signal detector to determine whether a signal exists in the current time window. Once a potential signal is detected, the data will be fed into the classifier for interpretation. Fig. 12 shows a layered view of ViSig's activities. On the top layer, camera key-frames as observed by a camera are shown for easy interpretation. In the middle layer, we have plotted the likelihood that the current time-window contains a body signal as determined by the signal detector module. On the bottom layer, at time windows where the detection threshold is exceeded, ViSig classifier module produces a probability of the actual body signal detected. Both the "leg-bye" and "out" signals are correctly identified in this example. Of course, some non-signal action such as touching the hat also increases the detection score, but it stays below the detection threshold, demonstrating the value of tuning the R_0 and R_1 boundaries per application. For more such examples, we refer the reader to view our anonymized video here: https://drive.google.com/file/d/1qnEoSr3OhQ5i63blyIqUlq1XnEe30JzA/view?usp=share_link.

5.2 ViSig Pipeline Internals: Body Signal Detection Accuracy

Next, we evaluate ViSig's performance on signal detection in each application, i.e., the ability to correctly identify a signal or a non-signal action.

5.2.1 Experiment Setup. In this experiment, volunteers are asked to wear ViSig devices and perform arbitrary signals or non-signal actions. As mentioned before, the key challenge of non-action signals is the unbounded nature and unpredictability of the non-action signal set. Since it is infeasible to train on all possible non-signals, we design two kinds of tests:

- Test on well-defined (WD) non-signal actions. In this evaluation, non-signal actions in both training/test datasets include a few pre-defined common actions such as walking, scratching, natural hand motion, etc.
- Test on undefined (UD) non-signal actions. In this evaluation, non-signal actions in the training set include the pre-defined common actions in WD above. However, non-signal actions in the testing set are not covered in the training set (bending, arm waving, arbitrary poses, etc.). This test evaluates the generality of ViSig's signal detection when samples in the test dataset follows a different distribution from the training dataset.

In both datasets, the samples of non-signals and signals are balanced to avoid training bias. ViSig trains the model following the steps in Section. 3.3 and identifies a signal or a non-signal action by feeding the input data into this model. In our experiment, ViSig employs an LSTM with 128 hidden units followed by a dense layer with 2 units. We set $R_I = 1$ and $R_O = 2$ (see Fig. 8 from Section 3.3). The thresholds are set to 0.8 for cricket umpire signals, and flag semaphore, 1.3 for baseball umpire signals and crane signals, and 2.2 for football official signals, allowing the least body signal deviation of body signals from the training set in cricket and flag semaphore, while allowing for highest deviation in football.

5.2.2 Baseline. We implement two baselines for comparison with ViSig: softmax classifier and deep SVDD [63].

- Softmax classifier: The softmax classifier employs an LSTM followed by fully-connected layers to propagate the input data. The output layer is a softmax classification layer which outputs a number between [0, 1] representing the probability of being a body signal.
- Deep SVDD. Deep SVDD takes only positive samples (i.e., signals) in the training process. It encodes the data with a deep network and then performs minimum volume estimation in the encoded space by finding a data-enclosing hypersphere of smallest size. In the evaluation, a sample is identified as a body signal if and only if it is inside the hypersphere. In our implementation, we employ an LSTM followed by fully-connected layers as the encoding layer.

5.2.3 Accuracy. We compute the accuracy of softmax classifier, deep-SVDD and ViSig on the two different datasets (WD and UD) via 3-fold cross-validation. The results of signal detection are shown in Table. 1. In the WD-dataset, since non-action signals are well defined, softmax classifier and ViSig both achieve high accuracy. In contrast, the accuracy of deep-SVDD is low as it is a model trained with only positive samples, which generates many false negatives. In the UD-dataset, ViSig achieves an overall improvement over other two approaches. Such improvement comes from the soft buffer boundary between signals and non-signal actions in the training dataset. One can decide whether the model prefers identifying input data as signal or non-signal via customizing the threshold ρ with application-specific knowledge. For example, body signals in flag semaphore are expected to be more precise compared to signals in football, hence ViSig uses a small ρ in flag semaphore and a large ρ in football. We also provide the confusion matrix of ViSig signal detection on UD-dataset in Fig. 13. The false positives and false negatives are overall balanced under the specified detection thresholds. We observe that mis-classified samples mainly occur at signals/non-signals which have an ambiguous boundary. For instance, in

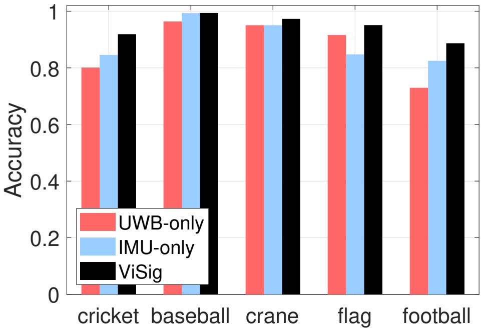

Fig. 14. ViSig outperforms UWB-only and IMU-only features in signal detection.

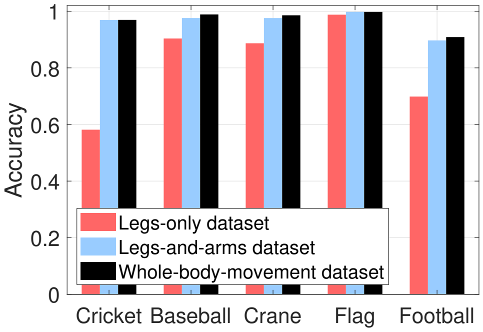

Fig. 15. Signal detection accuracy increases with data augmentation.

the cricket umpire signals, many mis-classified labels happen between “legbye (signal)" and “walking naturally". In the football official signals, "blocking below waist (signal)" is sometimes confused with “bending (non-signal)".

5.2.4 Ablation Study. We also perform an ablation study on the UD dataset to understand the importance of fusing UWB and IMU features. We employ the same network architecture for each case, only varying the input dimension to fit with the raw data dimension after removing UWB/IMU features. Fig. 14 shows that ViSig outperforms UWB-only and IMU-only approaches in all applications. This demonstrates that UWB-IMU fusion is capable of solving ambiguity that exists in the single-modality signal detection.

5.2.5 Effect of Expanding Dataset. Of course, current non-signal dataset is a subset of the whole non-signal space. *How will the performance change if the training set consists of various types of actions?* To understand the effect of an expanding dataset, we design an experiment where progressively more types of actions from the well-defined set are used from train the model. Specifically, we prepare three training datasets with increasing diversity in the non-signal actions:

- Legs-only dataset: Only incorporates standing still and walking as non-signals.
- Legs-and-arms dataset: Additionally incorporates natural actions such as touching face/head, crossing hand, etc.
- Whole-body-movement dataset: Additionally incorporates strenuous actions like running, jumping, etc.

We use the above three datasets and test on the UD dataset which includes hand-waving, bending, etc. Results are shown in Fig. 15. We observe a prominent improvement in detection accuracy from the legs-only dataset to the legs-and-arms dataset whereas the improvement from the legs-and-arms to the whole-body-movement dataset is marginal. This demonstrates that with more diverse non-signal samples, the model is trained to learn the true distribution of signals and non-signals which contributes to the performance improvement. However, seeing more and more non-signal body actions provide diminishing returns. When the dataset is sufficiently large, significant new data collection effort is needed to achieve higher accuracy.

5.3 ViSig Pipeline Internals: Body Signal Classification Accuracy

5.3.1 Experiment setup. Our evaluation of ViSig’s classification capabilities is divided into two stages: single-user evaluation, and cross-user evaluation. Note that the application domain of body signals typically requires a specific designated person to perform the duties of a signaler, who is also trained specifically for the job. Therefore, ViSig on-body sensor apparatus can be thought of as “owned” by a particular person. The model’s training of

Table 2. Datasets in single-user (SU) and cross-user (CU) evaluation.

	Signal #	Finger signals	SU-test user #	CU-test user #	SU-test sample #	CU-test sample #
Cricket umpire signal [69]	11	Yes	1	9	5520	4158
Baseball umpire signal [7]	13	Yes	1	11	4485	3276
Crane signal [65]	20	Yes	1	11	8602	5040
Flag semaphore [81]	30	No	1	11	8280	7560
Football official signal [9]	47	Yes	1	3	6072	3243

signal interpretation can therefore be performed on that particular individual. Of course, in many general human activity recognition applications, a model trained with multiple users and evaluated on a different set of users is more desirable. Therefore, we also evaluate ViSig across different volunteers to test the generality of our approach. In the first stage (single-user evaluation), one of the researchers collects data for all the 5 applications by wearing the sensors. For each application, the researcher performs all the signals multiple times over different days/scenarios. The purpose of evaluation in this stage is to present the performance of ViSig functioning as a user-specific system in signal interpretation. In the second stage (cross-user evaluation), we recruit 11 volunteers[3] for data collection and model evaluation. The heights of volunteers range from $1.65m$ to $1.83m$ so that we can assess the generality of the model.

In both stages, participant(s) wear 6 on-body sensors and perform every application-specific body signal. Then we segment the streaming samples into multiple $2s-$time windows. We arbitrarily extract k ($20 \leq k \leq 30$) time window samples per person per signal, adjusting for starting and ending times of the signal. In the single-user evaluation, more than 10 sets of signals (each application varies) are collected from the same volunteer in each application. In the cross-user evaluation, we collect one full set of signals from each volunteer in baseball, crane, flag semaphore, football, and two full sets of signals from each volunteer in cricket. The samples in every time window will be the raw data fed into the system for training a model. As described in Section 3, we fuse the distance matrix obtained by UWB and acceleration obtained by IMU to train the model, and then perform FSR with on-finger photodiodes. Data are collected at different locations including (i) outdoor open area; (ii) laboratory environment; (iii) indoor atrium; (iv) apartment; and (v) in a corridor, and at different times of the day. We set a hard range on distances for data sanitation. Only data with distances in $[0, 3m]$ are used. Overall over 30 hours of body signal data was collected for this evaluation which we plan to share publicly. To avoid model bias, we manually balance the number of samples for all volunteers in cross-user evaluation. The details of datasets are described in Table. 2.

For the evaluation, in the first stage, we average the classification accuracy of n-fold cross-*dataset* validation, where n is the number of datasets collected at different days/locations. In the second stage, we perform n-fold cross-*user* validation, where data of $n-1$ volunteers forms the training set and test the model on the last volunteer, for all n volunteers.

5.3.2 Single User Performance with Ablation Study. In this experiment, we evaluate the basic classification accuracy of ViSig on five different applications. Each specific body signal is a class in classification. Mathematically, for every class $i \in$ all classes C, ($|C| = c$), the accuracy is computed from the proportion of correctly predicted labels to all predicted labels (P, T are the predicted and true label):

$$\frac{\sum_i^c Number(P == i, T == i)}{Number_{all}} \tag{2}$$

[3] The user study has been approved by the IRB at Georgia Tech. Protocol Number: H21210 and Amendment #1, Title: TeamTrack: Wireless Tracking of a Team's Topology and Individual Poses.

Table 3. ViSig classification accuracy on different applications in the single-user test.

	IMU-only	UWB-only	ViSig
Cricket	93.9%	91.4%	**98.5%**
Baseball	92.4%	85.8%	**98.0%**
Crane	97.4%	82.9%	**98.7%**
Flag semaphore	89.6%	64.2%	**95.7%**
Football	93.9%	78.9%	**94.7%**

Table 4. ViSig classification accuracy on different applications in the cross-user test.

	IMU-only	UWB-only	ViSig
Cricket	88.0%	90.9%	**97.6%**
Baseball	87.9%	83.7%	**97.1%**
Crane	92.2%	80.9%	**97.7%**
Flag semaphore	80.8%	64.5%	**92.2%**
Football	84.7%	61.5%	**90.7%**

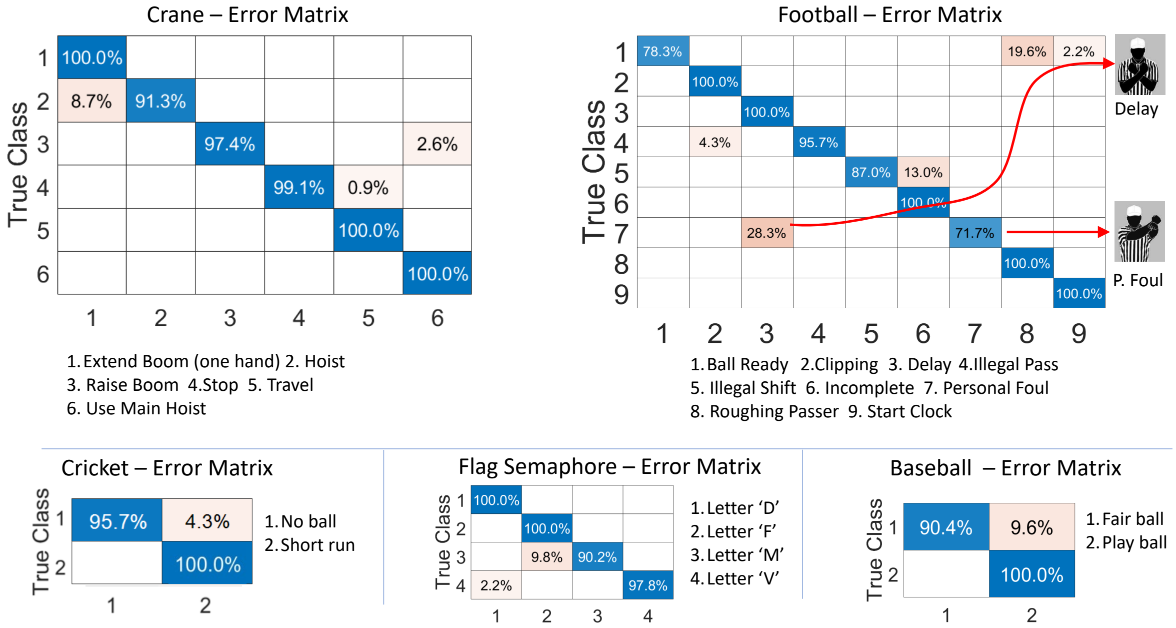

Fig. 16. Confusion matrices for five different applications. Only error confusions plotted for each application. Confusion occurs between very similar actions, as illustrated in the football case.

Since ViSig uses a UWB-IMU sensor fusion approach, we test the value of this fusion by comparing against a UWB-only and an IMU-only system. As shown in Table. 3, in the single-user test, ViSig achieves above 95% accuracy in four of five applications, and 94.7% accuracy in football which contains 47 different signals. Of course, certain body signals are more susceptible to being confused with other signals. Fig. 16 shows some example confusion matrices for different applications in the test. We observe that there are three leading causes of mis-classification: (1) proximity of different signals in the feature space, (2) intra-class data variance, and (3) raw data precision limits. For example, "Delay" and "Personal Foul" in football differ only in how far the crossing arms are from the chest (proximity in feature space). Such difference is blurred further after taking intra-class variance (for the same signal, put the hand at slightly different places but still recognizable), and inherent UWB distance measurement precision (±10 *cm*).

Table. 3 also gives the result of ablation study in the single-user scenario. For cricket, baseball, and flag semaphore signals. ViSig outperforms UWB-only models by 7.1%, 12.2%, 31.5% , and outperforms IMU-only models by 4.6%, 5.6%, 6.1% in the given signal examples. For crane signal and football signals, the improvements

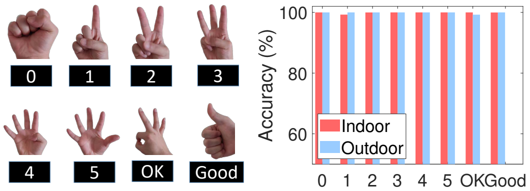

Fig. 17. 8 tested finger signs and the performance of FSR in indoor and outdoor scenarios.

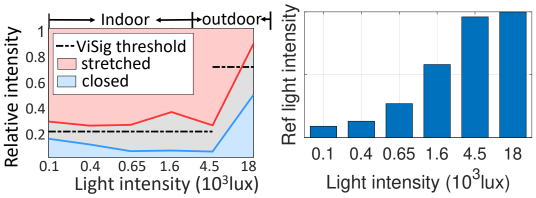

Fig. 18. Optimal threshold range in FSR (left); The measured reference light intensity (right).

of ViSig over IMU-only solutions are marginal, at 1.3% and 0.8% respectively. One of the reasons for the marginal improvement in these two applications is that crane signals and football signals involve many dynamic action where IMU can infer the signal from the motion pattern itself, without necessarily requiring UWB at all. The fusion allows ViSig to achieve 95% accuracy in all cases, which would not be feasible using any one technology alone. The gains come from inherently different properties of the two sensors; even if one modality incurs errors or cannot resolve ambiguity, the other one can still extract key features from raw data to perform signal interpretation.

5.3.3 Cross User Performance with Ablation Study. In the single user study, we computed the accuracy when the classification model was trained on that particular user's data. It is natural to wonder *can a single general model be created that would apply to any individual so that we would not need any per-individual training?* We evaluate ViSig's cross-user performance as specified in the experimental setup. An n-fold cross-user validation is applied to acquire average accuracy numbers.

The results are shown in Table. 4. In the cross-user scenario, we observe a slight decrease in the accuracy compared to the single-user scenario, which is caused by variance when different volunteers are performing signals. Overall, ViSig still achieves an over 90% accuracy in all applications. As for the ablation study, ViSig outperforms IMU-only models by 9.6%, 9.2%, 5.5%, 11.4%, 6% for each application, and outperforms UWB-only models by 6.7%, 13.4%, 16.8%, 27.7%, 29.2%. We observe a non-trivial increase in the improvement of ViSig over IMU-only models in the cross-user scenario. This indicates that in the cross-user scenario, there is a significant increase in IMU variance when performing the same signal, while fusing with UWB measurements provides the required robustness, further validating our reasoning of using both IMU and UWB modalities.

5.4 Finger Signal Recognition Accuracy

5.4.1 Overall Accuracy. As finger signs are frequently used in body signals to indicate counts, we conduct an experiment focusing on finger signal recognition accuracy. Fig. 17 shows 8 common finger signs used in this experiment. We collect data in both indoor and outdoor environment. The reference ambient light intensities ($\frac{I_{fin}}{I_{ref}}$) in the two environments are 400 lux and 18000 lux. Whether the finger is stretched or closed is determined by a threshold set to 0.2 in the indoor environment and 0.7 in the outdoor environment (we will empirically determine this threshold in the next section). Fig. 17 presents the FSR accuracy, showing that ViSig achieves higher than 99% accuracy for all signs, demonstrating the effectiveness of ViSig in FSR. The cause of 1% error is that sometimes light can leak to the photodiode on a closed finger. For example, when indicating "OK", light can leak to the photodiode on thumb and index finger at certain hand posture positions, leading to confusion with "four" or "five". Further improvements are possible by exploring better diode placements, but we leave those to future work.

5.4.2 Optimal FSR Threshold. Since the threshold on relative light intensity is important for accurate finger state detection, we now explore how such a threshold can be determined. Recall that we place a reference photodiode on the head which allows us to estimate the ambient light, I_{ref} at any given time. The light intensity measured by

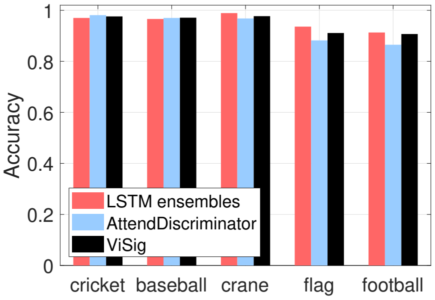

Fig. 19. Classification accuracy of different neural network models.

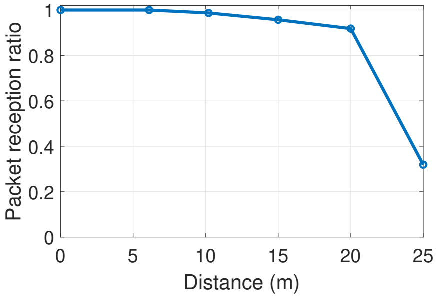

Fig. 20. Packet reception ratio at different distances in the indoor and outdoor environment.

Table 5. ViSig accuracy in different environments.

Location	Outdoor	Hall	Corridor	Room	Lab
Accuracy	100%	100%	100%	100%	100%

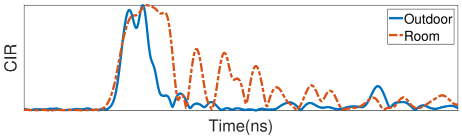

Fig. 21. Example CIR in the outdoor and room environment.

the photodiodes on the hand is typically a fraction of the current reference intensity I_{ref}; a larger fraction when the finger is stretched, and a smaller fraction when the finger is closed. However, in different lighting conditions, the amount of light that leaks between fingers is different. We need to select a threshold between the minimum observed intensity when the finger is stretched and the maximum intensity when the finger is closed.

To understand intensity variations better, we collect data at 6 positions with different light conditions (100, 400, 650, 1600, 4500, 18000lux respectively). Among them, positions with light intensity from 100 to 650 are indoor locations under artificial light. The positions with 1600 and 4500lux light intensity are indoor positions illuminated with sunshine (through glass panels). The position with 18000-lux intensity is an outdoor location. Fig. 18 shows the relative light intensity range of stretched fingers and closed fingers, and the intervening threshold selection margin. We observe that while a single threshold can be used for all indoor locations ($\leq$ 4500lux), a different threshold would be required for outdoor environment (18000 lux). In the indoor environments, under both artificial or through-glass sunshine, a threshold between $[0.1, 0.3]$ is needed. However, in outdoor environments, as light leaking between fingers can still be substantial, the threshold shifts to $[0.48, 0.83]$ range. Thus, our threshold selection algorithm must depend on the absolute measured light intensity level of the reference, selecting between an indoor-threshold or an outdoor-threshold. Note that sunshine intensity can vary significantly based on time of day, day of the year, and latitude. Therefore, a single threshold is unlikely to work; some fine-tuning will always be necessary, and the photodiode on the cap provides a good reference.

5.5 Microbenchmarks

5.5.1 Learning Models. Since the fusion of IMU and UWB features informatively describes a body signal, ViSig employs a simple network architecture to train the model. Furthermore, a model with a simpler network also has the advantage of implementation simplicity and resource efficiency. However, it is still essential to examine if ViSig suffers performance loss as a result of this choice of simplicity, irrespective of computational costs. Recently, a variety of new models have found application in the wearable human action recognition space. We compare our performance with two other promising approaches:

- LSTM ensembles [29]: LSTM ensembles trains multiple LSTM learners to output scores for each signal. These scores are then fed into a meta-classifier for signal interpretation.
- AttendDiscriminator [14]: Self-attention mechanism [77] relates different positions of a single sequence to compute a feature representation. In human action recognition, AttendDiscriminator employs a self-attention layer to encode cross-channel feature interactions which aids accurate recognition of activities.

We implement LSTM ensembles and AttendDiscriminator network and test the performance on the cross-user datasets. For LSTM ensembles, we train 10 LSTM learners. Fig. 19 shows that the accuracy differences on all 5 applications are quite minor (±2%). In crane, flag and football application, LSTM ensembles outperform our current model by $1 \sim 2\%$, but it needs a much larger model (10× memory size) as well as longer time (10.5×) to converge in the training. Preferring resource efficiency, ViSig chooses a simple LSTM followed by fully-connected layers as the neural network architecture.

5.5.2 Effect of Different Environments. ViSig makes a conscious effort to reduce the impact of different environments on performance. We completely shun the magnetometer which is known to be influenced near heavy machinery and even in indoor environments. Our choice of UWB for distance measurements provides robustness due to its ability to separate out wireless multipath caused by nearby objects, and its wireless frequency range ($3.5\,GHz$-$4.5\,GHz$) does not interfere with existing WiFi devices. Furthermore, since the on-body sensors are proximal to each other, we expect minimal influence from multipath. To validate our expectations, we additionally collect two sets of cricket umpire signal data from the same user at a total of 5 locations: outdoor, indoor atrium, corridor, fully-furnished room, and a lab. The observed channel impulse response (CIR) for outdoor space and the fully-furnished room is shown in Fig. 21, to provide a visual guide contrasting two of the extreme environments in this experiment. Table 5 shows that ViSig is able to achieve 100% accuracy in all the tested environments, thus providing substantial confidence in the robustness of our approach.

5.5.3 Overhearing Ranging Test. In our implementation, we have placed the UWB eavesdropper about 2 meters away from the user for data collection. However, an alternative is placing the eavesdropper on body, and then the eavesdropper forwards the received message to an edge device via a different wireless link (like Wi-Fi or LTE) with longer range. While such wireless transport modalities are beyond the scope of this paper, we perform a UWB ranging test to understand the maximum UWB range between the eavesdropper and the user in a pure UWB system. The test is conducted in the outdoor environment. We measure the packet reception ratio (PRR) when the eavesdropper is placed at different distances from the user. Results are shown in Fig. 20. In the outdoor environment, PRR rapidly drops from 91.8% to 31.9% when moving from $20m$ to $25m$, indicating the maximum available range in the outdoor environment is approximately $20m$. Again, this range is not the maximum range of ViSig. For instance, collecting data via Wi-Fi ($2.4/5\ GHz$) can extend the range to $100m$ [2].

5.5.4 Power Consumption. As signal interpretation is done externally on a laptop, the main power consumption of ViSig lies in the UWB ranging, IMU, and photodiode data collection. Each device has a power consumption of about 391mW. On a small 3.7V/1200mAh Lipo battery (4× 3cm^2, 24g), ViSig can work for 11.5 hours before requiring a recharge, satisfying the requirements of most applications.

5.6 Comparison with Computer Vision State-of-the-Art

Prevalent body signal interpretation solutions are mainly computer vision based. The state-of-the-art CV-based approaches propose using graph convolutional network (GCN) for body signal interpretation, showing an improvement over traditional network architecture. However, the performance of CV-based approaches is

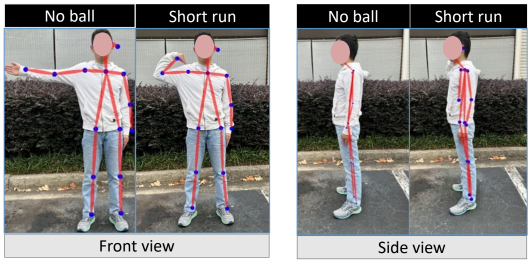

Fig. 22. The front view and side view of “No ball” and “short run” in cricket.

Table 6. Qualitative comparison with ST-GCN.

	ViSig	ST-GCN
System Deployment	On-body	External
Processing Rate	16Hz	1-60Hz
Dataset requirement	Small	Large
Memory/Bandwidth	Small	Large
Environment Influence	Small	Large
Multi-user Tracking	No	Yes

strongly influenced by camera perspective, self occlusion, subject’s distance from camera, etc. In contrast, ViSig presents a different dimension of solution which performs body signal interpretation via wearable computing. In this section, we compare the performance of ViSig with state-of-the-art visual systems both from the perspective of classification accuracy (quantitative), as well as other practical metrics (qualitative).

5.6.1 Classification Accuracy. To compare the classification accuracy, we choose ST-GCN [84], one of the state-of-the-art CV solutions, as our baseline. ST-GCN is a skeleton-based action recognition approach. It first extracts human skeleton joints from the image with OpenPose [21]. Then the joints are fed into a graph convolutional network for signal classification. We use cricket umpire signals as test application. A Google Pixel 4a is used to capture videos as volunteers are performing body signals. To evaluate ST-GCN’s performance, we start from the case when all the samples in the training/test dataset have clear front view (Fig. 22 left). Then we gradually incorporate in the training/testing dataset some self-blocking samples, where key skeleton joints are blocked by other parts of the body (Fig. 22 right). We observe significant performance degradation from **99.9%** to **61.4%** as we increase the proportion of self-blocking samples to 50%. This demonstrates that visual-based systems require high-quality, front-facing pictures for accurate signal interpretation, which is often impractical in many real-world applications. Many signals, because of self-blocking, become ambiguous at certain angles, such as “no ball" and “short run" in Fig. 22 when viewed from the side. In contrast, ViSig does not suffer from such obscuring of the input data, even in the cross-user scenario, and retains high accuracy over 90%.

5.6.2 Qualitative Comparison. Apart from accuracy, we also qualitatively summarize the main differences between ViSig and modern visual systems in Table. 6.

- System deployment: ViSig is a system with 6 wearable devices without the need for any external device (the overhearing device is only used for message communication). Such ego-sensing approach brings in movement flexibility without worrying about losing track. On the contrary, visual systems need multiple cameras to keep track of the user, which increases deployment cost and the difficulty in camera coordinating.
- Processing rate: Both ViSig and visual systems are capable of processing data in real-time. The major time overhead of ViSig lies in the message exchanging of UWB devices. For visual systems, the processing rate mainly depends on the data processing and model architecture, which varies from 1 to 60Hz.
- Training dataset requirement: As ViSig feeds raw physical data (distance and IMU data) which are informative in describing human skeletal poses, ViSig needs only a small dataset to train a model. On the contrary, visual systems use images as input data. Because of the projection and scaling transformation, visual systems need to incorporate images at different angles and distances to find non-biased distribution of different body signals, which significantly increases the demand for training data samples.

- Memory and Bandwidth: In addition to large number of data samples, computer vision systems also significantly increase the memory and bandwidth demands on the system. As an example, in our implementation, one time window sample of ViSig consumes about 3.52KB, while visual systems consumed 1.15MB in the classification comparisons, a 500× increase.
- Environment influence: As evaluated in Section 5.5, ViSig is hardly influenced by multipath in the environment. Moreover, UWB devices, which can work at 3.5-4.5 *GHz*, co-exists well with 2.4/5 *GHz* Wi-Fi due to UWB's low-power transmissions and much larger bandwidth. Given short-distance between the various on-body UWB sensors, the environment has very little influence on ViSig's performance. On the contrary, body signal interpretation with visual systems are influenced by many environment factors, such as ambient light, obstacles, etc.
- Multi-user tracking: As a wearable system, a set of ViSig sensors collect data from a single user for body signal interpretation. In this aspect, visual-based systems could support simultaneous multi-user body signal interpretation as long as human-skeleton data can be inferred from the image. Of course, signal interpretation on multiple users, each of which is wearing a separate ViSig system, is possible in a multi-user scenario, but we leave that for future work.

6 RELATED WORK

6.1 Body Signal Detection

The fundamental task of body signal detection is to identify whether a valid signal is contained in the current data stream. Because of the infinite space of non-signals, lots of research effort focuses on self-supervised solutions: they train with only positive samples (signals) and minimize the signal space [49, 63, 66, 72]. A drawback of such one-class models is that they fail to leverage prior knowledge of the application (e.g., a subset of non-signals). To improve, recent work like [55, 64, 73] propose semi-supervised learning training with both signals and a small portion of non-signals. However, in some applications, as signal space is also very large (football), self-supervised/semi-supervised solutions can incur many false negatives. In contrast, ViSig adopts the supervised learning for body signal detection. To tackle the issue of infinite non-signals, ViSig creates a soft boundary between signals and non-signals in the feature space, and applies a domain-specific threshold for signal detection.

6.2 Body Signal Interpretation

Technological interventions to interpret body signals have been mostly proposed from a viewer's perspective using cameras. Vision-based signal interpretation [18, 57, 60, 87, 88] captures images or videos, extract features and feed them into a machine learning model for interpretation. However, such methods require external infrastructure support, have limited viewing distance and field-view, and are sensitive to lighting conditions. On-body identification is therefore desirable, and initial attempts have been made in limited settings [17, 38], using IMU sensors alone. ViSig provides a robust framework with UWB and IMUs complementing each other. Since body signal identification can be seen as a subset of the general human activity recognition problem (HAR), we will discuss related work in this space as well.

6.3 Motion Capture Systems

Commercially, fine-grained HAR is achieved though a motion capture system such as Vicon [13] or OptiTrack [12]. In a motion capture system, a user mounts reflective markers or active LEDs on body. The position of each marker is captured by multiple synchronized high-speed cameras in a room. It provides precise *mm*−level motion tracking [48]. However, such systems are extremely expensive (starting at $\$10,000$) and only cover a large room.

Our requirements in ViSig which include outdoors and large coverage areas cannot be satisfied by motion capture systems.

6.4 HAR with External Cameras

Vision-based HAR is a well-studied research focus in the field of computer vision. Traditionally, vision-based approaches extract local spatio-temporal features such as HOG [24, 37], HOF [40], SIFT [67], etc., and feed them into classical machine learning classifiers. In recent years, driven by the advances in deep learning and the open source availability of large amount of video data, vision-based HAR approaches have shifted to complicated deep architectures to extract hidden features for recognition [19, 26, 27, 44, 68, 75, 84]. We have dedicated a part of our evaluation to comparison with computer vision based systems. However, vision-based approaches need the deployment of external cameras. Moreover, it fails in camera-denied scenarios, e.g., umpire tracking in crowded sport fields with frequent blocking, or performing flag semaphore operations in foggy marine situations. ViSig, in comparison, does not depend on such external deployments, or environment light condition, and through a careful selection of sensors alleviates the need for complex deep neural network models.

6.5 HAR with Wearable Sensors

Our approach in ViSig can be classified as a wearable-based HAR, which are also called egocentric tracking since all the sensors are on one's body. Inertial sensors are popular in this space [79]. Features are extracted from obtained IMU sensor data which are then classified using classical machine learning approaches in [16, 20, 41, 46, 59, 70, 74, 90]. However, hand-crafted features require expert domain knowledge and can be blind to non-intuitive hidden features.

Recently, the adoption of deep learning has improved the performance of action recognition significantly through automatic feature learning [35, 51, 76, 89]. CNN-based approaches [85, 89] build deep architectures with multiple convolutional and feed-forward layers to extract deep features automatically. Recurrent neural networks (RNNs) with long-short-term-memory layers (LSTM) encode time-series information effectively [51]. These techniques outperform CNNs in challenging datasets like OPPORTUNITY [62] and Skoda [86]. [29] improves model robustness by combining multiple LSTM learners. [50] employs an attention model to rescale the importance of historical samples, while [14] proposes a framework which extracts inter-modality relation as features by an attentional GRU encoder.

However, the success of these rather complex model architectures has slowed the search for new modalities. Electrocardiogram (ECG) [58, 91], electroencephalogram [22] (EEG), and electromyography (EMG) [43] are explored in healthcare and accessibility oriented human action recognition. However, ECG, EEG, and EMG signals do not directly reflect the human action, which makes it difficult to perform a general human action recognition task. [71] is perhaps the first to explore distance based metrics for HAR. However, they generate distance data through simulation, instead of from real-world systems. [15] uses UWB to measure step size for gait analysis, which indicates the practicability of leveraging distance information for HAR applications. Compared to existing works, ViSig is capable of extracting descriptive features, and has created an end-to-end solution to measure and embed inter-joint distance fused with information from IMU and photodiodes, though for a specific set of body signals. Our choice of sensors is unique and hopefully will pique interest of other researchers as well.

7 CONCLUDING REMARKS

ViSig makes unique contributions in exploring a solution to the problem of automatic interpretation of body signals using on-body sensors. We have shown that by fusing UWB based distance measurements with IMU based orientation, and light sensor based finger configuration detection, a rich and relatable feature set can be created, which provides intuitive understanding of body signals. Doing so improves accuracy of body signal identification in a variety of applications. ViSig also demonstrates that new modalities at the data end matter along with well-built data processing models, broadening the road towards a wider range of human activity recognition (HAR) applications. We expect that ViSig will infuse new energy into the general HAR problem by introducing inter-appendage distances as a new modality for future exploration.

ACKNOWLEDGMENTS

This material is based upon work supported by the NSF under Grant No. 2145278.

REFERENCES

[1] 2002. Taxiing Accident involving Arrow Air APWP6L. https://www.mot.gov.sg/docs/default-source/about-mot/investigation-report/28-feb-2002.pdf.

[2] 2013. IEEE Standard for Information technology– Telecommunications and information exchange between systemsLocal and metropolitan area networks– Specific requirements–Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications–Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz. *IEEE Std 802.11ac-2013 (Amendment to IEEE Std 802.11-2012, as amended by IEEE Std 802.11ae-2012, IEEE Std 802.11aa-2012, and IEEE Std 802.11ad-2012)* (2013), 1–425.

[3] 2016. IEEE Standard for Low-Rate Wireless Networks. *IEEE Std 802.15.4-2015 (Revision of IEEE Std 802.15.4-2011)* (2016), 1–709. https://doi.org/10.1109/IEEESTD.2016.7460875

[4] 2017. Decawave User Manual. https://www.decawave.com/sites/default/files/resources/dw1000_user_manual_2.11.pdf.

[5] 2018. Antenna Delay Calibration of DW1000-based Products and Systems (APS014). https://www.qorvo.com/innovation/ultra-wideband/resources/application-notes.

[6] 2021. 5DT Data Glove Ultra - 5DT. https://5dt.com/5dt-data-glove-ultra/.

[7] 2021. Baseball umpire signal. https://www.nfhs.org/media/1017816/baseball_umpires_signals_2021-1.pdf.

[8] 2021. CyberGlove Systems LLC. http://www.cyberglovesystems.com/.

[9] 2021. Football official signal. https://www.nfhs.org/media/4016213/2021-nfhs-official-football-signals.pdf.

[10] 2021. Industry leading VR techology - Manus VR. https://www.manus-vr.com/.

[11] 2021. Labor Force Statistics from the Current Population Survey. https://www.bls.gov/cps/cpsaat11.htm.

[12] 2021. optiTrack. https://optitrack.com/.

[13] 2021. Vicon motion capture system. https://www.vicon.com/.

[14] Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C Ranasinghe. 2020. Attend And Discriminate: Beyond the State-of-the-Art for Human Activity Recognition using Wearable Sensors. *arXiv preprint arXiv:2007.07172* (2020).

[15] Boyd Anderson, Mingqian Shi, Vincent YF Tan, and Ye Wang. 2019. Mobile gait analysis using foot-mounted UWB sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–22.

[16] Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*. Springer, 1–17.

[17] Sedney R Bedico, Edrhiza Mae L Lope, Erdwin John L Lope, Edward B Lunjas, Andrea Paola D Lustre, and Roselito E Tolentino. 2020. Gesture recognition of basketball referee violation signal by applying dynamic time warping algorithm using a wearable device. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 249–254.

[18] Sambit Bhattacharya, Bogdan Czejdo, and Nicolas Perez. 2012. Gesture classification with machine learning using kinect sensor data. In *2012 Third International Conference on Emerging Applications of Information Technology*. IEEE, 348–351.

[19] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. 2016. Dynamic Image Networks for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Andreas Bulling, Jamie A Ward, and Hans Gellersen. 2012. Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Transactions on Applied Perception (TAP)* 9, 1 (2012), 1–21.

[21] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[22] Hubert Cecotti and Axel Graser. 2010. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE transactions on pattern analysis and machine intelligence* 33, 3 (2010), 433–445.

[23] Americrane & Hoist Corporation. 2021. CRANE OPERATOR HAND SIGNALS AND THEIR IMPORTANCE. https://www.amchoist.com/news/crane-operator-hand-signals-and-their-importance-46177.

[24] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.

[25] Wilfrid Taylor Dempster. 1955. The anthropometry of body action. (1955).

[26] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[27] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1933–1941.

[28] Marilynn P Green. 2005. N-way time transfer ('NWTT') method for cooperative ranging. *Contribution 802.15-05-0499-00-004a to the IEEE 802.15. 4a Ranging Subcommittee* (2005).

[29] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28.

[30] R Hari and M Wilscy. 2014. Event detection in cricket videos using intensity projection profile of Umpire gestures. In *2014 Annual IEEE India Conference (INDICON)*. IEEE, 1–6.

[31] Jogi Hofmueller, Aaron Bachmann, and IOhannes zmoelnig. 2007. The Transmission of IP Datagrams over the Semaphore Flag Signaling System (SFSS). (2007). https://datatracker.ietf.org/doc/html/rfc4824.

[32] HM Sajjad Hossain, MD Abdullah Al Haiz Khan, and Nirmalya Roy. 2018. DeActive: scaling activity recognition with active deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–23.

[33] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–15.

[34] IEEE802.15.4z. 2020. IEEE Standard for Low-Rate Wireless Networks–Amendment 1: Enhanced Ultra Wideband (UWB) Physical Layers (PHYs) and Associated Ranging Techniques. *IEEE Std 802.15.4z-2020 (Amendment to IEEE Std 802.15.4-2020)* (2020), 1–174. https://doi.org/10.1109/IEEESTD.2020.9179124

[35] Jeya Vikranth Jeyakumar, Liangzhen Lai, Naveen Suda, and Mani Srivastava. 2019. SenseHAR: a robust virtual activity sensor for smartphones and wearables. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 15–28.

[36] Antonio Ramón Jiménez and Fernando Seco. 2016. Comparing Decawave and Bespoon UWB location systems: Indoor/outdoor performance analysis. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 1–8.

[37] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. 2008. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 275–1.

[38] Ming Hsiao Ko, Geoff West, Svetha Venkatesh, and Mohan Kumar. 2005. Online context recognition in multisensor systems using dynamic time warping. In *2005 International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. IEEE, 283–288.

[39] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.

[40] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[41] Oscar D Lara, Alfredo J Pérez, Miguel A Labrador, and José D Posada. 2012. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and mobile computing* 8, 5 (2012), 717–729.

[42] Selena Larson. 2017. Google Home now recognizes your individual voice. https://money.cnn.com/2017/04/20/technology/google-home-voice-recognition/index.html.

[43] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D Hand Pose Tracking using EMG Wearables. In *Proceedings of the Web Conference 2021*. 1471–1482.

[44] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[45] Marine Construction Magazine. 2020. CRANE OPERATION AND ROUTINE SAFETY PROCEDURES. https://marineconstructionmagazine.com/safety/crane-operation-and-routine-safety-procedures/.

[46] Alan Mazankiewicz, Klemens Böhm, and Mario Bergés. 2020. Incremental real-time personalization in human activity recognition using domain adaptive batch normalization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–20.

[47] Michael McLaughlin and Billy Verso. 2016. Asymmetric Double-sided Two-way ranging in an UWB Communication System.

[48] Pierre Merriaux, Yohan Dupuis, Rémi Boutteau, Pascal Vasseur, and Xavier Savatier. 2017. A study of vicon system positioning performance. *Sensors* 17, 7 (2017), 1591.

[49] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. 2020. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5216–5223.

[50] Vishvak S Murahari and Thomas Plötz. 2018. On attention models for human activity recognition. In *Proceedings of the 2018 ACM international symposium on wearable computers*. 100–103.

[51] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[52] International Civil Aviation Organization. 2005. Rules of the Air - Annex 2. https://www.icao.int/Meetings/anconf12/Document%20Archive/an02_cons%5B1%5D.pdf.

[53] Timothy Otim, Alfonso Bahillo, Luis Enrique Díez, Peio Lopez-Iturri, and Francisco Falcone. 2019. FDTD and empirical exploration of human body and UWB radiation interaction on TOF ranging. *IEEE Antennas and Wireless Propagation Letters* 18, 6 (2019), 1119–1123.

[54] Timothy Otim, Alfonso Bahillo, Luis Enrique Díez, Peio Lopez-Iturri, and Francisco Falcone. 2019. Impact of body wearable sensor positions on UWB ranging. *IEEE Sensors Journal* 19, 23 (2019), 11449–11457.

[55] Guansong Pang, Chunhua Shen, and Anton van den Hengel. 2019. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 353–362.

[56] Sarah Perez. 2019. Alexa developers can now personalize their skills by recognizing the user's voice. https://techcrunch.com/2019/09/26/alexa-developers-can-now-personalize-their-skills-by-recognizing-the-users-voice/.

[57] AJ Piergiovanni and Michael S Ryoo. 2018. Fine-grained activity recognition in baseball videos. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops*. 1740–1748.

[58] Bahareh Pourbabaee, Mehrsan Javan Roshtkhari, and Khashayar Khorasani. 2018. Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48, 12 (2018), 2095–2104.

[59] Nikhil Raveendranathan, Stefano Galzarano, Vitali Loseu, Raffaele Gravina, Roberta Giannantonio, Marco Sgroi, Roozbeh Jafari, and Giancarlo Fortino. 2011. From modeling to implementation of virtual sensors in body sensor networks. *IEEE Sensors Journal* 12, 3 (2011), 583–593.

[60] Aravind Ravi, Harshwin Venugopal, Sruthy Paul, and Hamid R Tizhoosh. 2018. A dataset and preliminary results for umpire pose detection using SVM classification of deep features. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1396–1402.

[61] Daniel Roetenberg, Henk Luinge, and Per Slycke. 2007. Moven: Full 6dof human motion tracking using miniature inertial sensors. *Xsen Technologies, December* 2, 3 (2007), 8.

[62] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*. IEEE, 233–240.

[63] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *International conference on machine learning*. 4393–4402.

[64] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. 2019. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694* (2019).

[65] Occupational Safety and Health Administration (OSHA). 2010. HAND SIGNALS FOR CRANE OPERATION. https://www.osha.gov/sites/default/files/laws-regs/federalregister/2010-08-09.pdf.

[66] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation* (2001), 1443–1471.

[67] Paul Scovanner, Saad Ali, and Mubarak Shah. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*. 357–360.

[68] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199* (2014).

[69] BBC Sport. 2021. The umpire's signals. http://news.bbc.co.uk/sportacademy/hi/sa/cricket/rules/umpire_signals/newsid_3809000/3809867.stm.

[70] Jie Su, Zhenyu Wen, Tao Lin, and Yu Guan. 2022. Learning Disentangled Behaviour Patterns for Wearable-based Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–19.

[71] Luke Sy, Nigel H Lovell, and Stephen J Redmond. 2020. Estimating lower limb kinematics using distance measurements with a reduced wearable inertial sensor count. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 4858–4862.

[72] David MJ Tax and Robert PW Duin. 2004. Support vector data description. *Machine learning* 54, 1 (2004), 45–66.

[73] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. 2021. Weakly-Supervised Video Anomaly Detection With Robust Temporal Feature Magnitude Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4975–4986.

[74] Roberto Luis Shinmoto Torres, Qinfeng Shi, Anton van den Hengel, and Damith C Ranasinghe. 2017. A hierarchical model for recognizing alarming states in a batteryless sensor alarm intervention for preventing falls in older people. *Pervasive and Mobile Computing* 40 (2017), 1–16.

[75] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features With 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[76] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. 2021. FedDL: Federated Learning via Dynamic Layer Sharing for Human Activity Recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 15–28.

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[78] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 349–360.

[79] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119 (2019), 3–11.

[80] Xin Wang and Zhenhua Zhu. 2021. Vision-based hand signal recognition in construction: A feasibility study. *Automation in Construction* 125 (2021), 103625.

[81] Wikipedia. 2021. Flag semaphore. https://en.wikipedia.org/wiki/Flag_semaphore.

[82] Wikipedia. 2021. List of International Cricket Council members. https://en.wikipedia.org/wiki/List_of_International_Cricket_Council_members.

[83] Wikipedia. 2021. Underway replenishment. https://en.wikipedia.org/wiki/Underway_replenishment.

[84] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

[85] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-fourth international joint conference on artificial intelligence*.

[86] Piero Zappi, Clemens Lombriser, Thomas Stiefmeier, Elisabetta Farella, Daniel Roggen, Luca Benini, and Gerhard Tröster. 2008. Activity recognition from on-body sensors: accuracy-power trade-off by dynamic sensor selection. In *European Conference on Wireless Sensor Networks*. Springer, 17–33.

[87] Julius Žemgulys, Vidas Raudonis, Rytis Maskeliūnas, and Robertas Damaševičius. 2018. Recognition of basketball referee signals from videos using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM). *Procedia computer science* 130 (2018), 953–960.

[88] Julius Žemgulys, Vidas Raudonis, Rytis Maskeliūnas, and Robertas Damaševičius. 2020. Recognition of basketball referee signals from real-time videos. *Journal of Ambient Intelligence and Humanized Computing* 11, 3 (2020), 979–991.

[89] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. 2014. Convolutional neural networks for human activity recognition using mobile sensors. In *6th international conference on mobile computing, applications and services*. IEEE, 197–205.

[90] Mi Zhang and Alexander A Sawchuk. 2013. Human daily activity recognition with sparse representation using wearable sensors. *IEEE journal of Biomedical and Health Informatics* 17, 3 (2013), 553–560.

[91] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. 2014. Time series classification using multi-channels deep convolutional neural networks. In *International conference on web-age information management*. Springer, 298–310.

[92] Hao Zhou, Taiting Lu, Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2022. Learning on the Rings: Self-Supervised 3D Finger Motion Tracking Using Wearable Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–31.