Multi-sensor Energy Efficient Obstacle Detection

Anupam Sobti M. Balakrishnan Chetan Arora
Indian Institute of Technology Delhi
{anupamsobti, mbala, chetan}@cse.iitd.ac.in

***Abstract*—With the improvement in technology, both the cost and the power requirement of cameras, as well as other sensors have come down significantly. It has allowed these sensors to be integrated into portable as well as wearable systems. Such systems are usually operated in a hands-free and always-on manner where they need to function continuously in a variety of scenarios. In such situations, relying on a single sensor or a fixed sensor combination can be detrimental to both performance as well as energy requirements. Consider the case of an obstacle detection task. Here using an RGB camera helps in recognizing the obstacle type but takes much more energy than an ultrasonic sensor. Infrared cameras can perform better than RGB camera at night but consume twice the energy. Therefore, an efficient system must use a combination of sensors, with an adaptive control that ensures use of the sensors appropriate to the context. In this adaptation one needs to consider both performance and energy and their trade-off. In this paper, we explore the strengths of different sensors as well their trade-off for developing a deep neural network based wearable device. We choose a specific case study in the context of a mobility assistance device for the visually impaired. The device detects obstacles in the path of a visually impaired person and is required to operate both at day and night with minimal energy to increase the usage time on a single charge. The device employs multiple sensors: ultrasonic sensor, RGB Camera, and NIR Camera along with a deep neural network accelerator for speeding up computation. We show that by adaptively choosing the appropriate sensor for the context, we can achieve up to 90% reduction in energy while maintaining comparable performance to a single sensor system.**

I. Introduction

The design of a portable real-time vision system which continuously monitors the surroundings is complex. Such a system must achieve an acceptable level of performance while making sure that the energy consumption is also within limits. Further, due to their always-on nature, these systems go through a lot more variations in the environment than a typical static system, which makes it imperative to employ a variety of sensors to perform efficiently under different conditions. Therefore, the system designers must classify various contexts, and choose the corresponding best performing and least power consuming sensor in that context, to ensure that the performance levels are met while also saving energy. Examples of such systems include robotic systems working in collaboration with humans, factory floor robots, assistive devices for visually impaired and even battery operated autonomous vehicles.

Technological advances in last few decades have made cameras inexpensive and ubiquitous. At a relatively low power usage, cameras allow us to obtain a large amount of information about the environment thus making a variety of applications possible with vision-based systems. However, extracting this information from the images is still a non-trivial

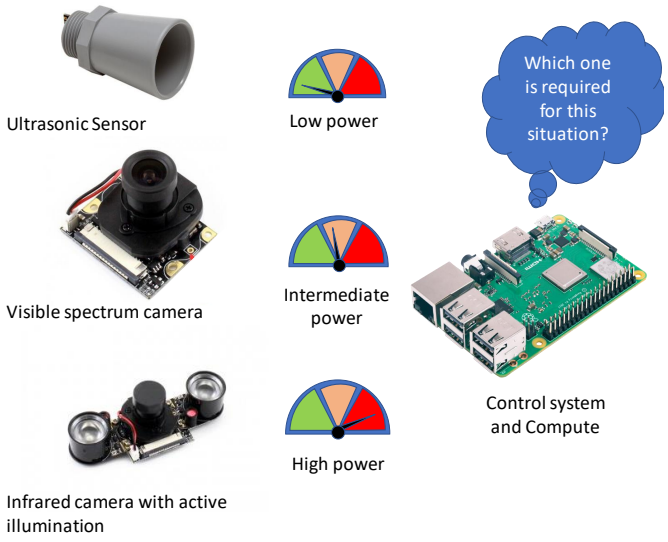


Fig. 1: Common sensors for obstacle detection have different energy consumption pattern along with the associated performance in a certain setting. For example, as shown above, an ultrasonic sensor system is extremely low power, a visible spectrum camera has an intermediate power consumption while an IR camera with active illumination is the most energy consuming. An ideal system should have an effective policy for actual usage of each sensor, to achieve high performance with energy efficiency.

task for computer algorithms. With the advent of deep learning techniques for image classification, object detection etc., some of these inference tasks that were previously considered extremely difficult are now beginning to be realized. However, the computational demand of these tasks is still very high. State-of-the-art object detection techniques use neural network architectures parameterized by a large number of weights ($\sim$5.8 million for the small-sized network we use). Therefore, a single inference involves millions of multiplications and additions and a large number of memory accesses. Several solutions [1], [2], [3] have come up in the form of dedicated accelerators for neural network computations. These solutions help decrease the latency of these computations and intend to reduce the power consumption as well by doing these computations in an energy efficient manner. We take a closer look at the trade-offs involved in the upcoming sections.

The problem of object detection becomes extremely chal-

lenging when there is not enough ambient light for capturing good quality images. Loh and Chan [19] show the state of object detection models on dark images. The problem becomes even more difficult when the images are taken from a portable device. This is because reduced shutter speeds are required to capture more light in the images, which causes a blur when camera is moving. Infra-red cameras have been shown to perform well during night conditions with the help of active illumination [16], where some of the techniques for object detection in a day light from visible spectrum images have been shown to work on the images taken by infra-red cameras as well.

Ultrasonic sensors work by emitting sound waves and measuring the object distance using time of flight. These sensors offer a coarse level of information about the presence of an obstacle, i.e., the distance of an obstacle within its beam width. However, reliability of this sensing principle allows robust detection of obstacles [26]. The complementary nature of camera and ultrasonic modalities make it attractive to deploy a fusion of these sensors for a more efficient perception of the environment.

Consider a device to be used for obstacle detection by a visually impaired person. In a typical computer vision based obstacle detection system, the parameters of the object detection model are learnt in a supervised setting. Apart from pre-training the model on a large data set (like Imagenet [25]) for classification, images annotated with bounding boxes and the class of object in the bounding boxes are used to train the object detection model for the intended set of objects. The system is, thereafter, able to recognize the objects that it is trained for. Therefore, the set of objects defines the visual vocabulary of the system. However, this also limits the efficacy of the vision system to only the variety of objects and conditions that it has been trained for. Any unseen object is simply not detected. Even for those in the visual vocabulary, success is limited by the accuracy of the algorithms involved. On the other hand, the detection from ultrasonic sensors is more robust and can be used to detect a large number of obstacles with very few exceptions. This helps to reduce false positives, as well as, helps in detecting obstacles which the object detection system may not have been trained for. An ultrasonic sensor also proves to be a very low power solution for continuous monitoring of obstacles as demonstrated later in the manuscript.

Human beings use variety of sensory organs for effectively perceiving the environment. We adaptively choose the sensing stream for performance and energy efficiency in our cognition. e.g., recognition through audio may often make the visual recognition redundant. The energy consuming visual modality is turned off when we sleep, while we still get alerted with audio cues, which suggests that we use it as a low energy trigger even though it provides a limited amount of information.

The focus of this work is to demonstrate the adaptive use of sensors in the obstacle recognition system for the visually impaired. Figure 1 shows the scenario in the obstacle recognition system under consideration. For reducing the high power requirement of day-night image based object detection, we propose a hybrid system that consists of a low power ultrasonic sensor, a camera with modes for both near-infrared and visible spectrum images as well as active illumination in the form of infrared LEDs. We use the low power ultrasonic sensor as an always-on sensor to trigger a more energy-hungry sensor/computation only when necessary. The need for illumination is also automatically detected and the infrared LEDs are turned on. This adaptive control enables the system to have a reliable performance while significantly saving the energy consumed.

In summary, the contributions of this work are as follows:

1) We design a multi-sensor vision system for obstacle recognition by the visually impaired people.
2) We demonstrate an adaptive sensor selection strategy for power optimization in the proposed system.
3) We explore the trade-offs in latency, energy and accuracy for using accelerators in the proposed system.

II. Related Work

Design of a vision based portable system is a highly complex design problem. When more sensors are added, the problem becomes even more complex and additional tasks for scheduling and determining the correct set of sensors has to be handled. The correct set of design points is a trade-off between energy efficiency, application performance and scheduling strategy. The related work comes from three domains:

a) Object detection in vision systems: After the breakthrough success of Alexnet [15], image understanding started to progress at a different pace. Works like RCNN [12], Faster RCNN [24], RFCN [8] etc. along with data sets like Pascal VOC [11], COCO [17] etc. have brought in a great amount of progress to object detection techniques. The operations carried out by the models are abstracted in terms of layers, i.e., convolutional, pooling etc which are parameterized with a large number of weights. The model, therefore, has millions of parameters. The weights of the first few layers are obtained by pre-training the models on larger data sets to learn generic feature extractors. Thereafter, a smaller data set is used to learn bounding boxes and object appearances for object detection. Techniques like Faster RCNN, RCNN etc. use bounding box proposals which are object independent and then classify them using a classifier. A slightly different class of models like SSD [18], YOLO [23] etc. jointly predict the bounding boxes as well as the classes in a single forward pass making them much faster, although, at the cost of accuracy. Objects are predicted at different scales and aspect ratios along with confidence scores for each object. The drawback of this approach is that the data sets have to be carefully curated and enough variety has to be captured in order to obtain a reasonably accurate working model. Even then, there is no scope of recognition of unseen obstacles/unaccounted variations in the environment (like illumination changes). Therefore, though camera based inferences are becoming increasingly popular,

operation of devices with only a camera as the sensor may not be feasible/reliable in many of the applications.

b) Energy optimization in multi-sensor systems: Adaptively using low/high power sensors to achieve energy efficiency has been demonstrated in some of the earlier works. Dutta et al. [9] demonstrate a system for detecting rare events over a large area where low power continuous sensing followed by a high power confirmation is used to reduce the system power in the sensor network. Tan et al. [29] demonstrate a vehicle counting system where a closed loop calibration mechanism is set up using a camera system and low-power PIR sensors, thus allowing the system to provide accurate predictions over a long period of time without needing manual intervention. Wren et al. [30] use a mixture of high-fidelity camera sensors along with motion detectors which contextually stitch information thus providing a higher context awareness coverage than what is possible with high-fidelity sensors alone due to both deployment and processing costs. The concept of using distributed levels of sensing with different resolution cameras has been demonstrated by Hengstler et al. [13] in an application of distributed intelligence surveillance.

In health care applications, a large number of patients require continuous monitoring of certain parameters. Here wireless body sensor networks have been used for monitoring, enabling the patients to remain mobile and active during the data capture. There have been a lot of efforts to reduce energy consumption in such systems. A detailed survey has been done by Rault et al. [22]. Sensor set selection is a technique used where a subset of sensors, which are more critical are selected to be transmitted. We use the technique similar to sensor set selection for our system. However, one major distinction between such systems and a portable system such as ours is the extent of computation done on-board. A large amount of computation for wireless health monitoring systems is offloaded to a central server infrastructure and therefore, the major energy consumption is in the form of communication cost. Sun et al. [28] use a cheaper accelerometer based activity detector for engaging high-energy sensors. Other related works include context detection for communication minimization/energy harvesting techniques. Possas et al. [21] use a reinforcement learning approach to switch between IMU and camera based approach for activity detection.

c) Multi-sensor assistive solutions for visually impaired: Mocanu et al. [20] demonstrated the use of ultrasonic sensors along with a camera where obstacle detection was done using feature based methods and then clustered using ultrasonic sensor readings. Elmannai et al. [10] provide a detailed survey on assistive devices built for visually impaired. However, none of these systems use a hybrid multi-modal approach to sensing and continue to rely on a single modality for perception. Ultrasonic sensors have been shown [26] to be particularly useful for detecting knee-above obstacles by augmenting them with a cane. This has motivated using these sensors along with cameras to provide a more detailed information about the obstacles.

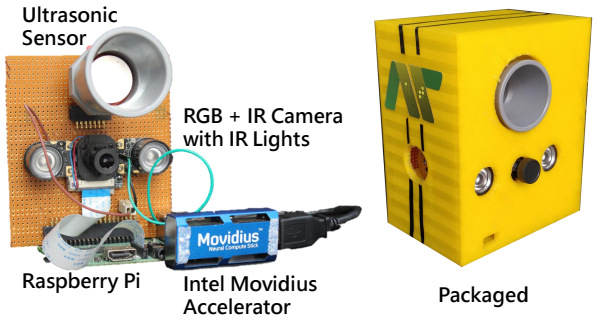


Fig. 2: The ultrasonic sensor is mounted on top of the camera. Both the sensors are connected through the GPIO pins of the Raspberry Pi. The complete system is powered through a power bank.

In the following sections, we demonstrate the design points applicable for a multi-sensor vision-based system. We also explore the pros and cons of including an accelerator in the system.

III. System Description

In this section, we discuss the hardware and software details of the proposed obstacle detection system. In addition, we describe the available design choices and define the design points being explored in further sections.

A. Prototype

We use a Raspberry Pi 3B [4] as the base platform which is connected to the Maxbotix Ultrasonic sensor (MB7383) [5] and a Waveshare RPi IR-CUT Camera [6]. The camera has an electronically controllable filter for filtering the IR light in bright light and another mode for capturing IR light emitted from the LEDs. The typical control of the LEDs is using light dependent resistors. However, we control the LEDs directly through software allowing for finer control on the time for which the LEDs are on. A picture of the prototype is shown in Figure 2. A movidius neural compute stick (v1) [1] is used for accelerating the deep neural network computation for object detection.

B. Software Setup

The Raspberry Pi 3B [4] runs a Raspian Stretch operating system. The ultrasonic sensor outputs the detection distance. Therefore, in the absence of an obstacle, we get the maximum value (9999 mm in this case). We use an ultrasonic threshold (7000 mm) to determine if the user is advancing towards an obstacle. A single thread constantly reads ultrasonic sensor data from the serial interface and compares it to a threshold. If an obstacle is detected, it sets an event flag which captures the image in the selected capture mode and then processes the image in the selected processing engine. We use the SSD Mobilenet [14] model for object detection. Figure 3 explains

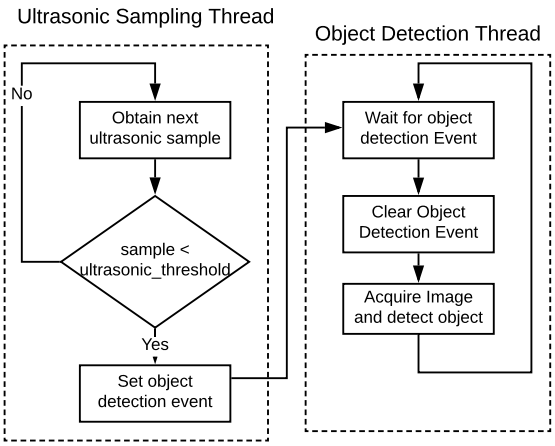


Fig. 3: Execution flow for the software setup. Ultrasonic samples are available in the serial buffer at the rate of 6 samples per second.

the flow. The processing engine in our case can be either a Raspberry Pi CPU or a Raspberry Pi CPU accelerated with a Movidius Stick.

There are two available capturing modes:

1) PiCam Interface: In this interface, the camera is always on and keeps sending frame data to the Raspberry Pi where the on-board GPU keeps receiving and storing the frames [7]. As soon as an ultrasonic detection is made, the frame is transferred to the main memory via DMA and is accessed by the thread running on the CPU, where it uses the processing engine to find objects. It takes $\sim 0.42s$ to get a frame in this mode. The camera constantly consumes 80 mA of current in this mode.

2) PiCam Lazy Interface: In this interface, the camera object is initialized only when an ultrasonic detection is done. The drawback here is the time taken for the camera to open and stabilize. It takes $\sim 2.3s$ for lazily getting a frame. The benefit of using a lazy capture interface is that the power consumption reduces to almost 0W when the camera is not being used.

C. Experiments

In this section, we describe the experiments conducted. There are four design points for the system:

1) PiCam Interface + Accelerator
   In this mode, the camera is always on, however, frame is only captured and processed once an ultrasonic trigger is received. The frame captured by the CPU from the camera is then sent to the Movidius NCS to process and the detection results are obtained from the same.
2) PiCam Interface + CPU
   In this mode, the camera is always on and a thread waits on the ultrasonic trigger to start a multicore object detection model inference on the CPU. This mode allows us to remove the accelerator completely without much loss of latency. Even though the multi-core implementation is power-hungry, the idle current reduces significantly since the Movidius NCS is now unplugged.
3) PiCamLazy Interface + Accelerator
   In this mode, the camera is uninitialized and powered off. The camera is turned on only when a trigger is provided by the ultrasonic sensor. Once a frame is captured, it is sent to the accelerator. This mode prevents the idle current of the camera. It is helpful in cases where the person walks slowly. Therefore, the object is captured with delay, however, the user is notified immediately.
4) PiCamLazy Interface + CPU
   In this mode, the camera is turned off until an ultrasonic trigger is received. The processing is done on the CPU. If the occurrence of obstacles is expected to be relatively low, choosing this mode is preferable over choosing PiCamLazy + Accelerator since the idle power in case of this mode is much lower. The latency of detection suffers and the person might have to wait for a few seconds to get the classification output.

The *accuracy* of the system is defined as the fraction of objects which were identified correctly in at least one of the frames containing the object, similar to the metric defined by Sobti et al. [27]. The SSD Mobilenet [14] model predicts the bounding boxes containing the objects as well as the class of objects which it contains. If the prediction confidence is greater than 0.5, the correct class is predicted and the intersection over union of the predicted box and the ground truth bounding box is more than 0.5, the object is considered to be detected. The accuracy loss in different modes is due to the following reasons:

1) When an accelerator is not used, the individual inferences on frames are slow. Therefore, the classifier gets fewer frames to predict the object correctly as compared to the case when an accelerator is used.
2) When the camera is accessed via the PiCam Lazy Interface, the time it takes for the camera to capture a frame can be substantial. During this time, the object may move/completely exit the frame. Therefore, the accuracy drops in this case as well.

The aim of the experiments which follow is to address the following questions:

1) What is the loss of accuracy when an ultrasonic based trigger is used to initiate the camera/infrared illumination rather than continuously capturing frames? How much benefit in terms of energy can this methodology provide?
2) When using an accelerator to identify the detected objects, what is the benefit in terms of accuracy and what is the cost paid in terms of power?
3) Is it feasible to completely shutdown the camera when using ultrasonic trigger (PiCam Lazy Interface) with an acceptable loss of accuracy?

For this analysis, we use the prototype discussed in Section III-A and record both video as well as ultrasonic data. This

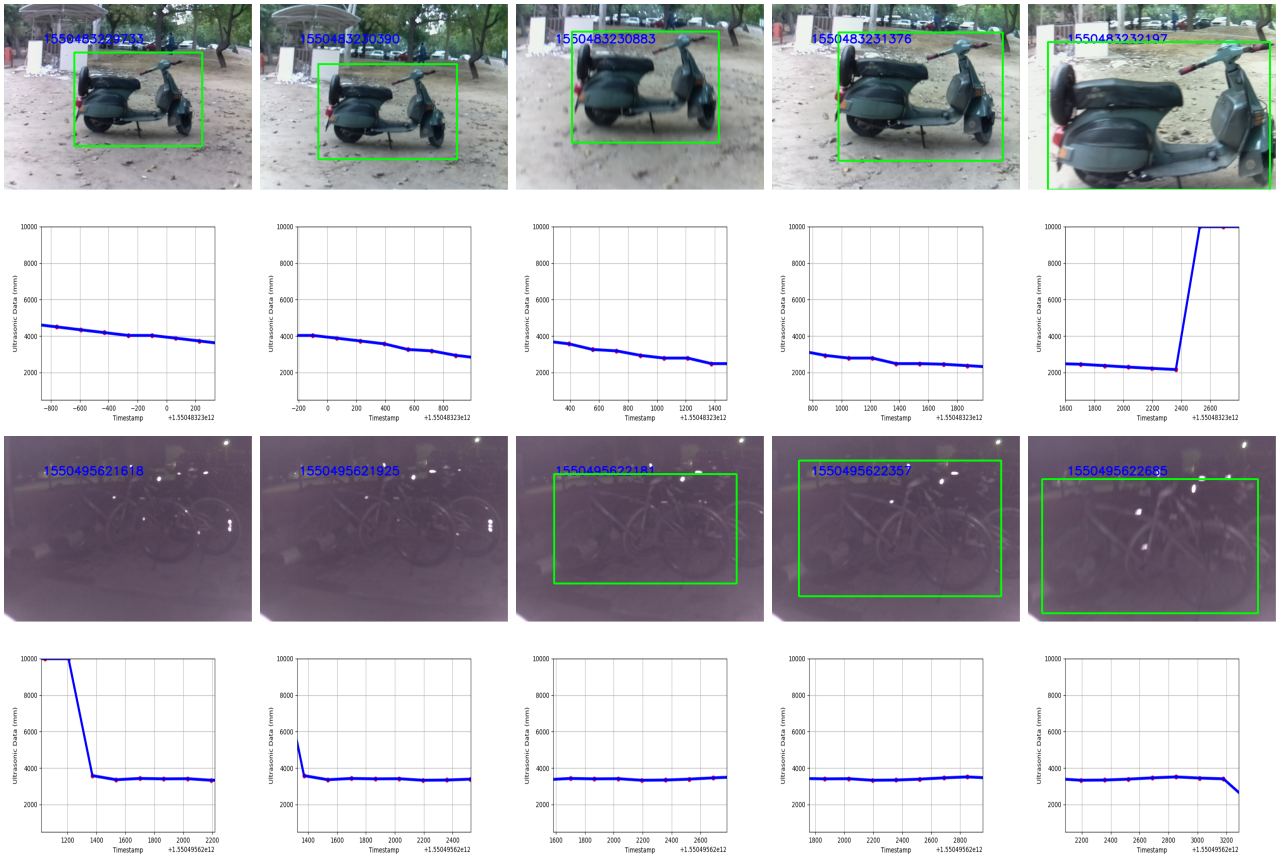


Fig. 4: Samples from the day/night sequence. The data contains time stamped ultrasonic samples and image samples with annotated bounding boxes, object number and class labels. The time series information for the ultrasonic sensor data for 1.2 seconds around the timestamp of the image is shown in the plot. We can see the distance of the object reducing as the user walks towards it.

data is then processed offline to validate the accuracy and power benefit. This was done using the methodology described in Section III-D. We cover essentially three scenarios:

1) No obstacles are present in front of the user
2) Obstacles are present and seen in the visible spectrum in bright conditions
3) Obstacles are present and seen only by the near-infrared active illumination in dark or low-light conditions

We use two recorded sequences – a day time sequence and a night time sequence in order to empirically evaluate the power consumption of the different configurations. Samples of the sequences are shown in Figure 4. We have considered bicycles, cars and motorbikes as obstacles which are recognizable by the vision system but this can be extended to other objects using standard computer vision techniques. Note that we have annotated only the obstacles which are in the path of the user and not all the objects in the image since they are irrelevant for obstacle detection. This is another way in which the hybrid approach outweighs the pure vision based approach since processing has to be done only when the obstacle is in the way. This is still a hard problem to solve with computer vision whereas obstacle recognition is shown to be effective using ultrasonic sensors used in devices like Smartcane [26]. Results are reported in Section IV.

Component	Power
Raspberry Pi	1.020 W
Movidius Stick	0.450 W
Camera	0.470 W
Ultrasonic Sensor	0.015 W

TABLE I: The different components of the static power are shown. The numbers above are from measurements during the idle state.

D. Methodology

a) The base power: The base system contributes significantly to the overall energy consumption of the system. A raspberry pi system doesn't have a power down mode. Therefore, the base power remains at $\sim 1.02W$ even when there is no useful processing except that the device is booted up and running the operating system. In this state, no peripherals are plugged in and the sensing process has not even been started. For a real energy efficient device it is important that the base system has a very low power consumption. A possible alter-

PE	Object	RGB/IR	Interface	Energy (mWh)
Movidius + CPU	Yes	RGB	PiCam	43
			PiCam Lazy	35
		IR	PiCam	63
			PiCam Lazy	42
	No	-	PiCam	36
			PiCam Lazy	29
CPU Multi-core	Yes	RGB	PiCam	38
			PiCam Lazy	34
		IR	PiCam	48
			PiCam Lazy	36
	No	-	PiCam	26
			PiCam Lazy	20

TABLE II: The dynamic power of different scenarios measured over 1 minute of operation. This is used to identify the energy per frame which is further used for the power analysis.

native would have been to use a micro-controller with power down mode(s) or an ARM based SoC where the power-down mode was available. Since the objective is to only illustrate the use of heterogeneous modalities, we use the Raspberry Pi platform and present the results based on incremental power consumption to the base power of the system. There is also a major difference in the compute capability and library support available for micro-controller/ARM based SoCs.

b) Power for components: The energy consumption is modeled in two parts - the static power and the dynamic power.

Static power is the fixed amount of power which gets consumed consistently for a particular configuration. Table I shows the different components of the static power. The power due to Movidius and camera do not contribute in the non-accelerated and the lazy interface respectively since the components are effectively disconnected.

Dynamic power is the power required for operations which consume a variable amount of power during their operation, e.g, capturing and processing a frame. For calculating the dynamic power, we measure the energy dissipated in any particular configuration over one minute of operation and obtain average power measurement from the same. Table II shows the dynamic power measurements for different design points/configurations. As expected, using the PiCam Lazy interface has a lower energy consumption than using the PiCam Interface counterpart. The difference in the energy consumption between the case when an object is present versus the one where it is not is the major contributor to energy saving. Another observation from the table is the difference in Movidius + CPU and CPU Multi-core options. The lack of constant current consumption by the Movidius NCS accelerator makes the system much more energy-efficient. The energy is only consumed when an object is present in front of the system.

Corresponding time of capturing and processing a frame is shown in Table III. We use Algorithm 1 for calculation of the energy and accuracy. This framework enables us to study the effects of different configurations in detail in a convenient and accurate manner.

PE	Interface	Capture and Process time (s)
Movidius + CPU	PiCam	0.789
	PiCam Lazy	2.670
CPU	PiCam	2.114
	PiCam Lazy	3.990

TABLE III: Time taken to capture and process a single frame in one of the configurations

Algorithm 1: Algorithm for energy/accuracy calculation

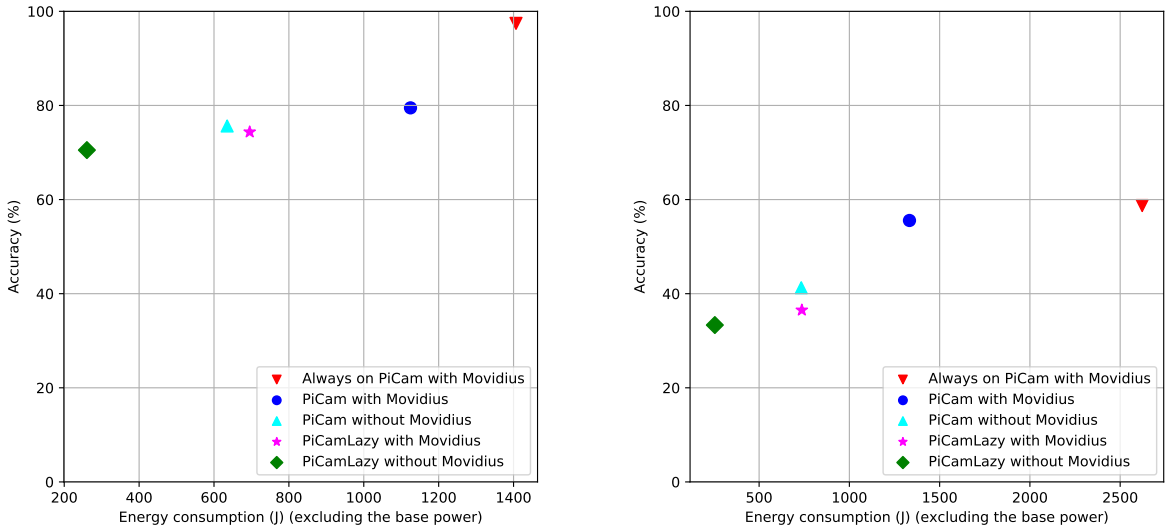
```
object_det_status = {0,0,...};
energy = 0;
while time < end_time do
    if ultrasonic_sample < threshold then
        time += camera_open_time;
        img = image_next_to(time);
        if object_detected(img) then
            object_det_status[obj_id] = 1;
        end
        time += capture_and_process_time;
        energy += energy_per_frame;
    else
        energy += static_power * (time - prev_time);
    end
    prev_time = time;
end
accuracy = sum(object_det_status)/total_objects;
```

IV. Results

The primary objectives behind the experiments were explained in Section 3. Our analysis is based on the sequences described in Section III-C. As mentioned before in Section III-C, accuracy represents the number of correctly recognized obstacles in at least one of the frames where the obstacle was present in the view. The energy calculations are done using the methodology described in Section III-D. Figure 5a shows the energy and accuracy of all the design points analyzed. Note that in any case, the hybrid system based detection is superior since unseen obstacles can be detected by the ultrasonic sensor.

The key observations obtained from the analysis are the following:

a) What is the loss of accuracy when an ultrasonic based trigger is used to initiate the camera/infrared illumination rather than continuously capturing frames? How much benefit in terms of energy can this methodology provide?: In the day sequence, the inference based on the neural network works reasonably well. Therefore, the accuracy is much higher as compared to the accuracy in the night sequence. When using the ultrasonic sensor based trigger, the accuracy suffers a hit of $\sim$17%. The energy reduction is 20% simply by processing the frame only when an obstacle is detected. The latency in



(a) Design points for the day sequence

(b) Design points for the night sequence

Fig. 5: Exploring the energy-accuracy tradeoff for different design points in day and night environments

this case is 0.789 sec, which is similar to the case a pure vision based method. Note that the accuracy numbers do not have objects which are not recognizable by the neural network however, in practice a hybrid system would be better for obstacle detection. In the night sequence, using an ultrasonic trigger alone reduces the energy consumed by half (∼49%). This shows the promise of using an ultrasonic trigger in such systems.

b) When using an accelerator to identify the detected objects, what is the benefit in terms of accuracy and what is the cost paid in terms of power?: When the accelerator is dropped from the configuration, an energy reduction of 55% is achieved at a loss of just ∼4% accuracy. The primary reason is the reduction in the static power of the configuration. A similar configuration in the night sequence provides an energy reduction of 72%.

c) Is it feasible to completely shutdown the camera when using ultrasonic trigger (PiCam Lazy Interface) with an acceptable loss of accuracy?: The PiCam Lazy Interface performs very well in the day, with energy reduction of **81.5%** with <10% decrease in accuracy. However, as the detector becomes less confident (night sequence), the delay incurred by the lazy capture causes much more reduction in accuracy. The energy reduction in the night time sequence is **90%**.

A. Occupancy and Persistence

For further explaining the fall in accuracy and the expected increase in lifetime, we introduce the concept of occupancy and persistence. *Occupancy* is defined as the proportion of time of operation during which an obstacle is present in front of the user. *Persistence* is defined as the number of frames for which a certain object stays in the field of view. A histogram of the persistence of objects in both sequences is shown in Figure 6.

In the recorded sequences, since the focus was on capturing

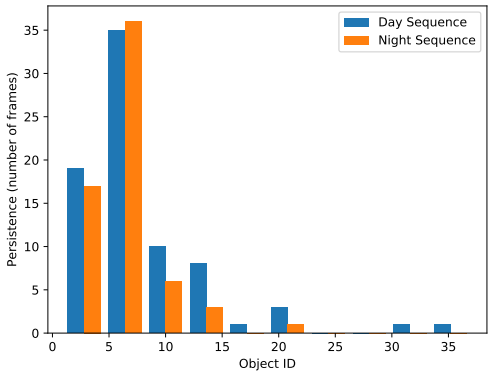


Fig. 6: Histogram of persistence of objects in the two sequences

a significant number of obstacles or analysis, the occupancy is ∼31% and ∼22% for the day and night sequences. In a practical scenario, the occupancy is expected to be much lower since the person is unlikely to come across obstacles that frequently. In such a scenario, the energy saving increases linearly with the occupancy. The slower configurations of Lazy Interface/without accelerator lose accuracy due to lower persistence of objects in the recorded sequence (median of 6.0). Thus, in a particular configuration, if the capture and process takes more time than the capturing of 6 frames, the object would no longer be ”captured” and the object would be missed. Therefore, a person walking at a slower pace may be able to get better energy performance with a slower configuration since the persistence of objects would be higher.

B. Online Validation

Finally, we measure the energy reduction by using the device in the vision only mode and the promising triggered modes. The sequences are recorded by walking through the same environment keeping the device in the respective modes.

We have shown the energy consumption and reduction in average power for these sequences in Table IV. Note that the reduction is not directly comparable to the sequences shown earlier due to different levels of occupancy, however the sequences are recorded to estimate the typical usage scenario.

Mode	Energy (mWh)	Time (s)	Energy w/o Base (mWh)	Avg. Power Reduction (%)
Vision Mode	215	338	119	0
PiCam (Movidius)	183	318	93	16.9
PiCamLazy (Movidius)	157	326	65	43.3
PiCamLazy (CPU)	122	308	35	67.7

TABLE IV: Online validation results for sequences in the same environment captured in different modes

V. Conclusion and Future Work

Heterogeneity in modalities and their adaptive control enables a system to be energy-efficient while still delivering high performance. In case of an obstacle avoidance system, a multi-sensor system is able to detect a much broader range of objects while being able to recognize objects in the visual vocabulary. Since the modalities are complementary, the device now covers an extended range of environmental contexts like ambient light and the types of objects learnt by the vision system. An efficient and accurate obstacle detection system has been demonstrated with the help of ultrasonic and vision sensors. The use of an accelerator reduces the latency of operation however the energy cost paid is significant. In our experiments, an adaptive control of camera and infra-red illumination along with removal of the accelerator reduces the energy consumption by up to 90% with roughly 25% decrease in accuracy. In future, it would interesting to see the state of the art that can be achieved with low base power micro-controller/microprocessor based systems.

VI. Acknowledgments

This work was supported by a grant from Department of Science and Technology, Government of India. Anupam Sobti has been supported by Visvesvaraya Fellowship, Government of India.

References

[1] software.intel.com/en-us/movidius-ncs. Intel Movidius Neural Compute Stick.

[2] www.gyrfalcontech.ai/solutions/plai/. Gyrfalcon Technology PLAI Plug.

[3] www.orangepi.org/Orange%20Pi%20AI%20Stick%202801. Orange Pi AI Stick.

[4] www.raspberrypi.org/products/raspberry-pi-3-model-b/. Raspberry Pi 3B.

[5] www.maxbotix.com/Ultrasonic_Sensors/MB7383.htm. Maxbotix MB7383 Ultrasonic Sensors.

[6] www.waveshare.com/wiki/RPi_IR-CUT_Camera. Waveshare RPi IR-CUT Camera.

[7] picamera.readthedocs.io/en/release-1.13. PiCamera Documentation.

[8] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[9] P. Dutta, M. Grimmer, A. Arora, S. Bibyk, and D. Culler. Design of a wireless sensor network platform for detecting rare, random, and ephemeral events. In *Proceedings of the 4th international symposium on Information processing in sensor networks*, page 70. IEEE Press, 2005.

[10] W. Elmannai and K. Elleithy. Sensor-based assistive devices for visually-impaired people: current status, challenges, and future directions. *Sensors*, 17(3):565, 2017.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[13] S. Hengstler, D. Prashanth, S. Fong, and H. Aghajan. Mesheye: a hybrid-resolution smart camera mote for applications in distributed intelligent surveillance. In *Proceedings of the 6th international conference on Information processing in sensor networks*, pages 360–369. ACM, 2007.

[14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] J. Li, F. Zhang, L. Wei, T. Yang, and Z. Lu. Nighttime foreground pedestrian detection based on three-dimensional voxel surface model. *Sensors*, 17(10):2354, 2017.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[19] Y. P. Loh and C. S. Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019.

[20] B. Mocanu, R. Tapu, and T. Zaharia. When ultrasonic sensors and computer vision join forces for efficient obstacle detection and recognition. *Sensors*, 16(11):1807, 2016.

[21] R. Possas, S. Pinto Caceres, and F. Ramos. Egocentric activity recognition on a budget. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, 2018.

[22] T. Rault, A. Bouabdallah, Y. Challal, and F. Marin. A survey of energy-efficient context recognition systems using wearable sensors for healthcare applications. *Pervasive and Mobile Computing*, 37:23–44, 2017.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[26] V. Singh, R. Paul, D. Mehra, A. Gupta, V. D. Sharma, S. Jain, C. Agarwal, A. Garg, S. S. Gujral, M. Balakrishnan, et al. 'Smart'Cane for the Visually Impaired: Design and Controlled Field Testing of an Affordable Obstacle Detection System. In *TRANSED 2010: 12th International Conference on Mobility and Transport for Elderly and Disabled PersonsHong Kong Society for RehabilitationS K Yee Medical FoundationTransportation Research Board*, 2010.

[27] A. Sobti, C. Arora, and M. Balakrishnan. Object detection in real-time systems: Going beyond precision. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1020–1028. IEEE, 2018.

[28] F.-T. Sun, C. Kuo, and M. Griss. Pear: Power efficiency through activity recognition (for ecg-based sensing). In *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 115–122. IEEE, 2011.

[29] R. Tan, G. Xing, X. Liu, J. Yao, and Z. Yuan. Adaptive calibration for fusion-based cyber-physical systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 11(4):80, 2012.

[30] C. R. Wren, U. M. Erdem, and A. J. Azarbayejani. Functional calibration for pan-tilt-zoom cameras in hybrid sensor networks. *Multimedia Systems*, 12(3):255–268, 2006.