Container monitoring with infrared catadioptric imaging and automatic intruder detection

Victor E. Trujillo II¹ · Mark K. Hinders¹

Received: 17 September 2019 / Accepted: 18 November 2019 / Published online: 25 November 2019

© This is a U.S. Government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2019

Abstract

We describe a framework for global shipping container monitoring using machine learning with low-power sensor hubs and infrared catadioptric imaging. A mesh radio satellite tag architecture provides connectivity anywhere in the world, with or without supporting infrastructure. We discuss the design and testing of a low-cost, long-wave infrared catadioptric imaging device and multi-sensor hub combination as an intelligent edge computing system that, when equipped with physics-based machine learning algorithms, can interpret the scene inside a shipping container to make efficient use of expensive communications bandwidth. The histogram of oriented gradients and T-channel (HOG+) feature is introduced for human detection with low-resolution infrared catadioptric images, and is shown to be effective for various mirror shapes designed to give wide volume coverage with controlled distortion.

Keywords Container monitoring · Infrared camera · Catadioptric imaging · Machine learning

1 Introduction

In the 80 years since a trucker named Malcolm McLean first conceived the modern-day shipping container, shipping has exploded into a \$400B a year industry [1]. The shipping container has radically transformed supply chains, fundamentally inter-twined domestic and international economies across the world, and changed societies in the process. In a recent survey by the World Shipping Council, whose members operate as much as 90% of the global liner ship capacity, the international liner shipping industry transported approximately 130 million containers packed with cargo, with an estimated value of more than \$4 trillion dollars [2]. These metal boxes will remain a staple into the future for an industry that relies upon their ease of use for transporting all types of cargo.

Despite their versatility, shipping containers are easy to break into, and difficult to keep track of, as they are transported around the globe. Cargo theft and cargo loss are estimated to cost the industry at least \$50B annually,

according to The National Cargo Security Council [3]. Losses can occur in a myriad of ways, from containers being mislaid, mislabeled, or simply failing to arrive at their destination, to instances involving premeditated criminal intent, such as breaking into ports to steal goods, or pirates attacking crews at gunpoint for their valuable cargo. In the last several decades, security and visibility of the global supply chain has largely been addressed through improved locking mechanisms and radio tracking devices that are attached to shipping containers, and it has been estimated that there are now over one million remote tracking systems on containers worldwide [4]. The latest adaptations to the existing technology include GPS capable devices that use satellite, cellular, or Wi-Fi connectivity. Containers are often modified post manufacturing with better locks or smart technology for improved security and visibility.

These methods have enhanced the industry but have come at a high monetary cost due to the infrastructure needed to track devices that use short-range radio

✉ Mark K. Hinders, hinders@wm.edu | ¹William and Mary Applied Science Department, Williamsburg, VA 23187-8795, USA.

frequency identification (RFID) and in some cases high bandwidth costs to transmit the data wirelessly. Vulnerabilities and shortfalls plague current RFID tracking methods, especially during disaster relief when wireless networks are destroyed, in emerging markets which exist outside wireless coverage areas, and in providing real-time information to stakeholders.

New technology offers lower cost and higher efficiency sensors connected to the internet that can better inform decision-making. Recent developments in this technology have improved sensing capability with lower power consumption. Ubiquitous devices still have bandwidth and power limitations, but the challenge of working in austere environments has been improved via on-board processors with the ability to do edge computing. By processing data near the sensor, bandwidth is dramatically reduced between the sensors and any central datacenter. Thus, it is now possible to equip containers with a device that can talk to stakeholders during its entire voyage in the transportation network. Better yet, it is possible to put sensors on the inside of the shipping container without drilling holes to feed signal wires by transmitting data through the container wall using ultrasonic waves.

2 GRIDSAT tag architecture

Our goal in this research is to develop a real-time tracking device for containerized goods. The system incorporates the existing GRIDSAT Tag architecture (Fig. 1) to provide location and status of goods as they are transported around the globe.

Low power sensor hubs [7] like those found in wearables and smartphones, fuse the inputs of several different types of MEMS sensors such as accelerometers, magnetometers, and gyroscopes without engaging the main processor, thereby reducing power consumption by up to 95%. A device that incorporates these low power sensor hubs can be attached to the inside of a container to detect intrusion or changes in conditions to monitor the status of goods inside. New compact long-wave infrared (LWIR) imaging systems also provide useful features for our application because they readily detect the body heat of an intruder. The Seek Thermal Compact LWIR camera is a small, inexpensive device which attaches to a smartphone (Fig. 2).

First, we use machine learning techniques to maximize status awareness and minimize data transmission with data collected by various microelectromechanical sensors (MEMS). Second, we transmit information through container walls without drilling holes and feeding wires utilizing ultrasonic guided waves via an appropriate combination of frequency and amplitude modulation (Fig. 3).

Fig. 1 GRIDSAT Tag Architecture utilizes interconnected GRID tags configured in a self-healing mesh network which communicate with a GRIDSAT tag. Messages are compiled and transmitted via satellite communications to stakeholders through a GIS Software Application Package [5]. We use the Iridium satellites for full global coverage [6]

Fig. 2 The left image is a false-color thermography image of a person standing inside the open door of a shipping container. Note that the roof is bright because it's being heated by the sun, as is the background seen through the open door. The middle image shows a similar view with the catadioptric mirror above the door and seek thermal imager attached to the roof identified. The right image shows a close-up view of the data collection apparatus with Seek Thermal LWIR camera, catadioptric mirror, and TI Sensor Tag that communicates to the iPhone 5

This allows the device to be attached to the inside of a container and communicate to a device on the outside of the container, which is connected to a GRIDSAT Tag architecture for worldwide tracking.

Fig. 3 Lamb waves can be transmitted in a pitch-catch setup from a piezoelectric transducer on the inside surface of a container to a transducer on the outside surface of the container. We find that these guided waves can be encoded with information via an appropriate combination of frequency and amplitude modulation [8–24]

3 Catadioptric imaging

With a 36-degree diagonal angle of view, the Seek Thermal cannot fully cover the inside of a shipping container, but there are several methods for increasing that field of view. The first is refractive optics such as a fisheye lens, but lens materials for the LWIR can be expensive and infrared fisheye applications often use a system of multiple lenses. The second is reflective optics which combines a standard imaging system with a convex mirror of polished brass or aluminum. A third method is rotating the imaging system, but that method is prone to failure and our goal is to make the device as robust and low-cost as possible. Since we are talking about a semi-disposable system operating in a harsh environment, our assessment is that a simple, low-cost shaped reflector is the right approach.

Chahl and Srinivasan state the best shape for an omnidirectional mirror is hyperbolic [25] because only the hyperbolic mirror guarantees a linear mapping between the angle of elevation and the radial distance from the center of the image plane. Moreover, the resolution in the omnidirectional image captured increases with growing eccentricity, and hence, guarantees a uniform resolution for the panoramic image after unwarping.

Chahl and Srinivasan also note that mis-alignment of the camera with respect to the reflective surface produces distortions in the image from the camera. Similarly, a lateral displacement (orthogonal to the axis of the surface) results in a distortion of the plane of displacement that is offset from the correct angle of reflection. Therefore, polynomial surfaces must be imaged with the focus very close to the surface. The exact location of the focus varies with radial angle, unlike a plane mirror. To achieve high-resolution images, it is necessary to use a camera with a wide depth of field and the ability to focus at close range. This is simply a result of the surface being curved and is also true for spherical mirrors and cones. Intensive ray tracing is necessary to fully map the location of the focus for objects at different ranges and different angles

of elevation. In practice, however, such a procedure is not required because of the large depth of field of typical charge coupled device video cameras, for which these surfaces are often designed.

Kecskes et al. [26] utilize a commercial off-the-shelf *GoPano* reflector, originally designed for visible light applications, to develop a nighttime surveillance concept to detect moving, human-sized heat sources at ranges between 10m and 70m. They replaced the SiO_2 coating on the aluminum surface of the *GoPano* reflector with a Clausing Beral coating that is highly reflective in the thermal infrared and requires no additional overcoating. Commercial IR coating companies also offer low-temperature application processes for metallic coatings, so in addition to 3D printing and then coating the required geometry as we do here, it is possible to mill a single-surface mold, vacuum thermo-form a sheet of plastic, and then coat it with an IR reflective material, resulting in low-cost custom-shaped front-surface reflectors.

3.1 Catadioptric theory and design

A catadioptric sensor uses a combination of lenses and mirrors configured to capture a wide field of view [27]. Many mirrors today seek an omnidirectional field of view (FOV) which give a warped view of the world in a donut shape on the focal plane array. This invariably renders portions of the rectangular focal plane array unusable, reducing the pixel use efficiency and increasing the effective price per pixel of the detector [28]. The very low-cost LWIR imagers we are using do not have pixels to waste, so that becomes a primary design consideration. It is often the case that more complicated designs require careful registration between the camera lens and the mirror. The environment for our application can be jarring and a container's route over land and sea requires multiple on and off loadings, so we seek a design that is robust to the conditions and one that allows proper sampling of the scene for decision making.

Baker and Nayar [27] developed the complete class of conic section mirrors used in catadioptric sensors that satisfy the single viewpoint (SVP) constraint for omnidirectional viewing. The SVP is rigid by design, so Swaminathan et al. [29] developed the geometry and analysis for non-single viewpoint (non-SVP) catadioptric systems which pose no constraints on mirror shape and relative position to the camera.

Yagi and Yachida [30] developed a tiny omnidirectional image sensor for autonomous robot navigation that adopts a technique used in reflecting telescopes to minimize the influence of blurring. Their design overcomes spherical aberration, astigmatism, and coma by using a two-mirror system with a convex primary mirror

to capture rays from the environment and a concave secondary mirror to converge the rays to a single center of projection. This design is subject to the same limitation as the Chahl and Srinivasan family of mirrors, in that it is not suitable for harsh environments and requires careful calibration for assembly. Hicks [31] developed a conquistador helmet shaped mirror using differential methods that yields a cylindrical projection to the viewer without digital unwarping and leads to a resolution that is more uniform. Similarly, Srinivasan [32] designed equatorial view and polar view mirrors which eliminate the need for computational resources to map circular images into a rectangular form for analysis. Hicks [33] also developed a method for catadioptric sensor design for realizing prescribed projections. Realizing implementation of these types of freeform mirrors can be a manufacturing challenge and thus expensive.

Existing thermal panoramic imagers with parabolic or hyperbolic reflectors have optical aberrations which are typically reduced by imaging at low numerical apertures. At small apertures, any lens operates in a near-pinhole mode, but this approach to reduction of aberrations comes with a significant loss of light. In panoramic catadioptric systems, the mirror size affects the effective aperture. Larger catadioptric panoramic cameras have lower aberrations and produce better images, at the expense of the overall larger size of the optics. Small panoramic optics with low aberrations presents a design challenge for thermal optics where diffraction of the longer wavelength dictates large apertures, regardless of the light throughput, in order to match the diffraction limited spot size with the pixel pitch of the sensor.

Gutin et al. [34] developed an infrared panoramic imaging sensor for 360-degree automatic detection, location, and tracking of targets. In testing the effect of an SVP versus a non-SVP optical system arrangement, they determined a noticeable difference at short distances. However, they found that a critical distance for the detection of an object exists when the angular size of the optics as viewed from the object is equal to or larger than the immediate field of view, or the angle subtended by a single pixel in the object space. They also concluded that a non-SVP optical system with axial symmetry only affects the elevation of the object which still allowed determination of azimuth of the object. In a concept developed by Simon Thibault [35], which he calls the IR panamorph lens, we can think of the surveillance problem in terms of a pixel allocation problem by using distortion as a design parameter. This allows us to optimize our design by providing the highest resolution coverage where it is needed by dividing the coverage area into zones. Once the coverage area is divided into specific security zones a relative angular resolution analysis is used to allocate pixels.

When contemplating mirror shapes, we considered field of view, spatial resolution, and distortion for various zones of coverage. There is an inherent interplay between these considerations which invariably leads to tradeoffs. Common security mirror shapes are spherical. The half dome or quarter dome variety are often used to facilitate wide fields of view and to see around corners. However, these mirror shapes do not allocate pixels efficiently, in a catadioptric sensor, when paired with a rectangular focal plane array. We found that an oblate sphere with a somewhat flattened equatorial zone works well to capture the inside of a shipping container. Figure 4 shows an example of our experimental mirror shape inside of a checkboard lined box to illustrate distortions in three zones. The equatorial zone is rounded on the sides to allow viewing into the corners but generally flat offering minimal distortion to the opposite wall. The top and bottom zones are domed and preserve parallel lines, but, very small blind areas exist directly below and above the mirror. In practice, catadioptric design is a process from detector through lens (system) to the reflector such that aberrations are reduced or eliminated. We begin our proof-of-concept using an off-the-shelf, infrared imaging device with a 206×156 pixel focal plane array and estimated its lens parameters [8].

A single effective viewpoint is achieved by positioning the entrance pupil of the lens at one of the focal points of the reflector, making the other focal point the effective viewpoint. However, if the camera is placed between the mirror and the scene, a portion of the scene will be obscured by the camera. Thus, we prefer a non-single viewpoint sensor with no required projection of planes in the scene for two reasons. First, we have limited pixels to capture events, so we can ill afford to obscure any portion of the scene by the camera. Second, the movement of shipping containers in the supply chain will inevitably

Fig. 4 Polished brass mirror shape inside of a checkboard lined box. Head and shoulders region of interest (ROI) is least distorted in the center and distortion steadily increases moving laterally outward as expected

jostle the camera's position relative to the mirror. Thus, we design robustness into the system by selecting a mirror geometry that achieves a full field of view of the inside of a shipping container while allowing for slight shifting during movement over time. The mirror's profile is then generally rectangular, to maximize utilization of the camera's FPA, with the necessary curvature to provide the desired field of view.

Seek Thermal publishes limited data about the technical characteristics of their device, but interaction with their support team has provided sufficient information to make approximations for the design of a catadioptric system. We do not know the true aperture size, thus, we do not know the entrance pupil or f-number of the imaging device. Moreover, we do not know the true detector size as it was estimated with the given pixel pitch of 12 microns [36]. Therefore, we estimate a focal plane array of 2.472×1.872 mm (3.10 mm diagonal). Furthermore, we computed a focal length of 4.77 mm using the manufacturer stated 36° diagonal angle of view. The Seek Thermal lens material is chalcogenide which has a refractive index of approximately 2.60 at $10\mu\text{m}$ assuming BD-2, which is composed of $\text{Ge}_{28}\text{Sb}_{12}\text{Se}_{60}$. It offers low dispersion, broad transmission, and has minimal absorption [37]. Using the Camera Calibration Toolbox for MATLAB [38] we measured tangential and radial distortion of the Seek Thermal. We captured 21 images of a heated $6'' \times 6''$ flat piece of T316 welded stainless steel mesh in different positions by rotating/translating the wire mesh [8]. We took extra caution to detect corners manually by framing the region using the outside edges of the wire and included the wire thickness in the square size computation. Points at the corners of the image were displaced by as much as 45 pixels due to radial distortion. Radial lens distortion is a symmetric distortion caused by the lens due to imperfections in curvature when the lens was ground. By inspection, it is not clear if the measured radial distortion in the Seek Thermal lens produces barrel or pincushion distortions. It is possible that a lateral chromatic aberration exists in the lens which is due to radial and wavelength dependencies [39]. Nonetheless, considering noise effects and errors in finding the exact corners of the non-perfect steel mesh grid pattern we conclude there are no significant non-linear effects of the Seek Thermal lens allowing us to proceed with mirror design.

3.2 Mirror design

The parametric equations of a sphere have parameters s and t and radius r , modified by scalars a , b and c to control the scaling of z as well as to flatten the face of the sphere to reduce distortion in the main region of detection. We use the following parameterization of the mirror surface.

$$\begin{aligned}x &= ar \cos(s) \cos(t) \\y &= ar \sin(s) \cos(t) \\z &= br \sin(ct) \cos(t)\end{aligned}\quad (1)$$

where $0 \leq s \leq \pi$ and $-\pi/2 \leq t \leq \pi/2$ where a , b , c , and r are constants such that $a > b > 0$ and $c > 0$ and r is the radius.

Determining the exact location of the imaging device's entrance pupil to the reflector is important for two reasons. First, we want to capture the entire reflector in the frame of the camera. Second, we need the exact location of the entrance pupil to compute fields of view and spatial resolutions. Our method fits a rectangle to the exact back profile of the mirror then computes the distance of the entrance pupil from the mirror for various pupil heights fixing the lens' diagonal angle of view. The pose (rotation) of the entrance pupil is fixed at the origin of our system, but it can be adjusted through a range of angles as necessary.

The camera's FOV is not the angle of view (36°) used to position the camera. Due to the geometry of the mirror, the FOV is derived in three-dimensions for the non-SVP case [29]. A three-step process begins by computing the caustic [40] which can be thought of as the locus of viewpoints of these rays [41]. We model the incident and reflected rays using surface normals and we find two extreme intersection points along the equator of both the reflector and the caustic surfaces. Finally, we determine the camera's horizontal FOV by computing the angle between two incident rays which are formed from the lens' entrance pupil and extending to the two extreme intersection points on both surfaces.

Referring to Fig. 5, the caustic is formed by letting $S_r(s, t)$ be a point on the three-dimensional reflector, parameterized by (s, t) [29].

$$S_r(s, t) = \langle ar \cos(s) \cos(t), ar \sin(s) \cos(t), br \sin(ct) \rangle \quad (2)$$

Let O denote the position of the entrance pupil of the lens. For any point $S_r(s, t)$ on the reflector, we define the direction of the ray entering the pupil as:

$$V_r(s, t) = O - S_r(s, t) \quad (3)$$

Given a reflector geometry, we define the incoming light rays $V_i(s, t)$ using surface unit normals $N_r(s, t)$, where $\times$ denotes the cross product [29, 42].

$$N_r(s, t) = \frac{S_r(s, t)_t \times S_r(s, t)_s}{|S_r(s, t)_t \times S_r(s, t)_s|} \quad (4)$$

$$V_i(s, t) = V_r(s, t) - 2N_r(s, t)(N_r(s, t) \cdot V_r(s, t)) \quad (5)$$

The caustic is then defined by [9]:

Fig. 5 The model for a lens-based camera placed off-axis with respect to a symmetric reflector parameterized by (s, t) . The view-point locus (caustic surface) of this imaging system is not rotationally symmetric and must be derived in three dimensions. The incident light ray $V_i(s, t)$ reflects off the reflector surface at $S_r(s, t)$ and images after passing through the entrance pupil O along the reflected ray $V_r(s, t)$. The surface normal $N_r(s, t)$ is used in the Law of Reflection, in conjunction with the reflector surface, to define the reflected and incident rays

$$S_c(s, t) = S_r(s, t) + r_c \cdot V_i(s, t) \quad (6)$$

where r_c denotes the distance from the point of reflection at which the caustic lies along $V_i(s, t)$. A point on the caustic is defined as a singularity in the space of scene rays, parameterized by (r_c, s) . Thus, in the limit, traversing infinitesimally along $V_i(s, t)$ (change in r_c) at the caustic is equivalent to traversing from one ray to the next (change in s) [29]. At this point, the determinant of the Jacobian $J(S_r(s, t) + r_c \cdot V_i(s, t))$ must vanish [42]. We find the roots of the quadratic equation that results in solving for r_c after applying the Jacobian method [29, 41] for non-single view-point catadioptric cameras.

$$\det[J(S_r(s, t) + r_c \cdot V_i(s, t))] = 0 \quad (7)$$

The x , y , and z components of the vectors are denoted as $S_r(s, t)_x$, $S_r(s, t)_y$, and $S_r(s, t)_z$, respectively. The roots are substituted into the equation of the caustic $S_c(s, t)$ yielding its analytic form.

Swaminathan et. al proved that the point on the reflector at which a light ray passing through the entrance pupil of the lens grazes the reflector is also its caustic point [29]. Thus, step two of computing the camera's field of view requires finding the two furthest intersection points along the x -axis between the reflector surface and the caustic (Fig. 6).

The x -axis corresponds to the horizontal direction, or equatorial cross section of the mirror, in our coordinate system. Therefore, the extreme x -values correspond to the furthest right-hand and left-hand points on the two surfaces, a two-dimensional computation for the rotationally symmetric case. An Extended Newton Method is used to compute the intersection points between the two surfaces: reflective surface $S_r(s, t)$ and caustic $S_c(u, v)$ [43]. Our system consists of three equations f_1, f_2, f_3 and four variables s, t, u , and v , which are parameterized by w , as defined below.

We proceed by re-parameterizing the caustic surface S_c with parameters u and v , to differentiate the two surfaces, and use differential geometry procedures [43]. Let S_r and S_c be as follows:

$$S_r(s, t) = \begin{pmatrix} S_{r1}(s, t) \\ S_{r2}(s, t) \\ S_{r3}(s, t) \end{pmatrix} \quad S_c(u, v) = \begin{pmatrix} S_{c1}(u, v) \\ S_{c2}(u, v) \\ S_{c3}(u, v) \end{pmatrix} \quad (8)$$

where, $S_r, S_c \in [0, 1]^2$. The problem of intersection is solved by computing the set:

$$M = \{(s, t, u, v) \in [0, 1]^4 \mid F(s, t, u, v) = S_r(s, t) - S_c(u, v) = 0\} \Leftrightarrow \quad (9)$$

$$\begin{aligned} F(s(w), t(w), u(w), v(w)) &= \begin{pmatrix} S_{r1}(s, t) - S_{c1}(u, v) \\ S_{r2}(s, t) - S_{c2}(u, v) \\ S_{r3}(s, t) - S_{c3}(u, v) \end{pmatrix} \\ &= \begin{pmatrix} f_1(s, t, u, v) \\ f_2(s, t, u, v) \\ f_3(s, t, u, v) \end{pmatrix} = 0 \end{aligned} \quad (10)$$

In the non-degenerate case the solution manifold M consists of one or several curves each fulfilling the condition:

$$F(s(w), t(w), u(w), v(w)) = 0 \quad (11)$$

where w is the arc-length parameter. We parameterize by arc-length w which is sometimes used to trace a constant distance along the arc-length of the intersection curve. We interpret the pair of two curves as: $k_1(w) = (s(w), t(w))^T \subseteq [0, 1]^2$ and $k_2(w) = (u(w), v(w))^T \subseteq [0, 1]^2$ which represent the path of intersection on each of the surfaces in the corresponding parameter space.

Following a process described by Alsaidi [43], we implement a routine in MATLAB by first setting $S_r(s, t) = S_c(u, v)$ and solving for the parameters (s, t, u, v) . The process begins by obtaining an appropriate start point lying within the solution manifold. An arbitrary initial guess to start the Extended Newton Method results in sporadic

Fig. 6 A pixel element of area δA in the image plane projects through the entry pupil of the lens onto the reflector as a region of area δS . The pupil is located at $(0, 0, d)$, with respect to the origin. The principal ray from δA reflects off the reflector at

$S_R(x(t, \theta), y(t, \theta), z(t, \theta))$. The corresponding viewpoint on the caustic surface is shown above. The solid angle subtended at this viewpoint is then $\delta\omega = \delta S/r_c^2$, where r_c is the distance of the viewpoint from the reflector. Resolution is then defined as a ratio of $\delta\omega$ to δA

intersection points that are sometimes unwanted intersection points on the back portions of the reflector and caustic. Marching methods, used to find the intersection curves between parametric surfaces, took longer than 72 hours runtime using MATLAB on the William & Mary high-performance computing cluster [44]. This was due to the complicated caustic surface function. We overcame this by estimating surface values over the defined domain with a meshgrid of computed surface values to find the corresponding parameters values for Extended Newton Method initialization.

Our first approach to find the intersection between the reflector and caustic surfaces met resistance in multiple ways. First, it is not possible to plot the caustic directly. Not being able to visualize the two surfaces meant it was a guessing game which root to use for the caustic surface. Second, our code resulted in three roots when solving the Jacobian, which led to some trial and error to find the correct root. After multiple attempts trying each root value, we decided to estimate the caustic surface and use a discrete meshgrid to aid the convergence of iterative methods. With good approximate initial values of parameters

s, t, u , and v convergence was fast, about 5 minutes with 25 iterations on average, using a tolerance value of 1×10^{-7} .

3.3 Horizontal and vertical fields of view

The camera's horizontal field of view (C-HFOV) is computed as the angle between a vector from the entrance pupil to extreme point #1 (corresponding to the maximum x -value) and a vector from the entrance pupil to extreme point #2 (corresponding to the minimum x -value). This computed C-HFOV is greater than the nominal 28.35 degree horizontal angle of view (computed from the Seek Thermal's 36 degree diagonal angle of view) used to position the camera. The reason for different values is due to the geometry of the reflector. It is important to go through this process, as merely using the incident rays on the reflector's surface does not account for the locus of viewpoints as described above. The catadioptric system's horizontal field of view (CS-HFOV) follows as the angle between the same incident rays just described. We conclude that this must be the CS-HFOV by visual inspection of the combined plot of caustic and corresponding reflector surfaces. There are no other

incident rays with a wider field of view that reflect off of the mirror and enter the pupil.

We compute the catadioptric system's vertical field of view (CS-VFOV) as the angle between incident rays on the top and bottom of the mirror surface. These incident rays are found by searching along elliptical contours on the top and the bottom of the reflector surface in 5-degree increments until incident rays that reach the pupil are found. This also helps to trim the surface by removing unnecessary curvature that cannot be viewed from each pupil location. Moreover, this procedure does not change any characteristics of the mirror, but instead defines the CS-VFOV from our model definition.

We derive the 2D spatial resolution for a conventional perspective camera, which has a frontal image plane located at a distance f from the camera pupil (modeled as a pinhole) and whose optical axis is aligned with the axis of symmetry of the mirror. See Fig. 6 for an illustration of this scenario. Following the method by Baker and Nayar [45], we consider an infinitesimal area δA on the image plane. If this infinitesimal pixel images an infinitesimal solid angle $\delta\omega$ of the world, the resolution of the sensor (as a function of the point on the image plane at the center of the infinitesimal area δA) is $\delta A/\delta\omega$. If ψ_i is the angle between the optical axis and the line joining the pinhole to the center of the infinitesimal area δA , the solid angle subtended by the infinitesimal area δA at the pinhole is:

$$\delta\theta_i = \frac{\delta A \cos \phi_i}{f^2 / \cos^2 \phi_i} = \frac{\delta A \cos^3 \psi_i}{f^2} \quad (12)$$

Therefore, the resolution of the conventional camera is:

$$\frac{\delta A}{\delta\theta_i} = \frac{f^2}{\cos^3 \psi_i} \quad (13)$$

whose behavior tends to decrease as $\psi_i \rightarrow 0$, so higher resolution areas on the sensor plane continuously increase the farther away they get from the optical center. Then, the area of the mirror imaged by the infinitesimal area δA is:

$$\delta S = \frac{\delta\theta_i (p - S_r(s, t))^2}{\cos \phi_i \cos \psi_i} = \frac{\delta A (p - S_r(s, t))^2 \cos \psi_i}{f^2 \cos \phi_i} \quad (14)$$

where p is the location of the pupil (x_o, y_o, z_o), $S_r(s, t)$ is the surface value of the mirror, and ϕ is the angle between the normal to the mirror and the line joining the pinhole to the mirror. This distance is denoted d in the diagram. Since reflection at the mirror is specular the solid angle of the world imaged by the catadioptric camera is:

$$\delta\omega = \frac{\delta S \cos \phi_i}{r_c(s, t)} = \frac{\delta A (p - S_r(s, t))^2 \cos \phi_i}{f^2 (r_c(s, t))^2} \quad (15)$$

Hence, the spatial resolution of the catadioptric system is:

$$\frac{\delta A}{\delta\omega} = \frac{f^2 (r_c(s, t))^2}{(p - S_r(s, t))^2 \cos \psi_i} \quad (16)$$

From this equation we can examine our system in two ways. First, we want to know how resolution behaves for regions of interest (ROIs) in the scene. Second, we want to know the spot size resolution for a given heat source in the scene. Referring to Fig. 4, we superimpose a grid onto the image of our reflector, placed inside of a black and white checkerboard box. Then, five real-world planes (three walls, the floor, and the roof) are mapped to the reflector. We determine each region's location on the reflector using the superimposed grid and note the distortions for each travelling in lateral directions away from the center of the reflector. The distortion is easily seen in Fig. 4 which depicts a sixth region we call the *Head and Shoulders* ROI. This ROI is the most likely location of a human's torso in the image. As expected, the distortions in the center of the reflector are small and increase moving towards the reflector's edge. The real-world corners, closest to the reflector, represent the areas where we see the most distortion. In general, from this image, we see that the near corners, and the floor area just below the reflector, are areas where we expect scene objects to be the most distorted. We keep this in mind for image processing as it may be possible to use a measurement metric to classify shapes.

For each of the six ROIs, we computed the spatial resolution along lateral lines of the reflector to understand behavior. This modeled behavior uses a scene point that is infinitesimally small and an incident ray from infinity, so the spatial resolution values are extremely small. Then, we computed spatial resolution curves along three radial slices (top, middle, and bottom) for each ROI. Next, we use the derived spatial resolution (16) to compute a spot size of a given area in the scene onto the focal plane array of the Seek Thermal.

3.4 Maximizing seek thermal pixels

The distribution of the Seek Thermal's pixels onto the reflector surface can be computed by mapping each pixel from reflector to image plane according to calibration techniques. By transformation and projection, mapping pixels is shown to work well for rotationally symmetric shapes [46]. Our reflector surface does not directly adapt to the derived models, however, since our basic need is to evaluate the pixel use for various mirror shapes, so we employ a new approach. We project the mirror surface onto the y -plane, according to our coordinate system, and compute the area difference between said projection and the smallest fitting rectangle. This simplistic approach is not quite good enough, however, because it does not allow us to evaluate the effect on

viewed area from different pupil heights. We improve this basic approach by using the computed visible area according to our model, first, and then project the viewable mirror's surface to the image plane.

The four points that allow us to compute CS-HFOV and CS-VFOV are used to compute the visible area of the mirror for each pupil location. Since the computations of our model account for viewing direction, we use this fact to perform the projection and then find the area difference between the projected mirror surface and the smallest fitting rectangle. We do not assume any rectangle according to the projected mirror shape, but instead use computed surface values to draw the rectangle so that we verify it is the correct projection. The simple equations that capture the definition of Area Difference and Pixel Efficiency are then:

Area Difference = $A_r - A_{S_r}$ and

$$\text{Pixel Efficiency} = \frac{A_r}{A_{S_r}} \times 100$$

where A_r is the area of the smallest fitting rectangle and A_{S_r} is the area of the reflector surface as projected onto the y-plane.

3.5 Positioning the catadioptric sensor inside the container

The placement of an infrared imaging system on the inside of a shipping container can be investigated systematically by a geometry of space analysis. First, sampling the scene for infrared emittance from a heat source should be considered by computing the viewing angle for various camera placements. This procedure requires setting up a geometry of space model from which to calculate angles. Due to the symmetry of the geometry of a shipping container, we investigated camera positions at three locations.

First, at the top/center of the door and the corresponding location at the top/center of the back wall. Second, at the top corner of the container and third, at the top/middle along one side of the container. Any other sensor locations would be obscured by container contents and/or would quickly be damaged by the loading and unloading of contents. We did not consider a sensor location in the top/center of the roof of the container, which provides a good vantage point, because containers can be loaded to the roof and this position would lead to sensor damage and/or obscuration. Moreover, a catadioptric sensor in this location would require an omnidirectional mirror, which is not an efficient use of scarce pixels.

3.6 Mirror design results

After a number of initial trials with mirror shapes, we downselected ten variations of mirror shapes which include three different x- and y-scale factors, two different z-scale factors, and two different curvature factors. Using three different pupil heights we show the results of 30 mirror configurations (Table 1). A summary of results follows:

- The average C-HFOV is 33.23°, with standard deviation 2.02 and variance 3.94
- The average CS-HFOV is 212.13°, with standard deviation 1.24 and variance 1.47
- The average CS-VFOV is 144.28°, with standard deviation 17.45 and variance 294.49
- The average Area Difference is 750.02, with standard deviation 204.49 and variance 40,423.90
- The average Pixel Efficiency is 91.09, with standard deviation 0.46 and variance 0.20

The effect of parameter values on CS-HFOV are as follows (increase means better):

- Holding x- and y-scale factor and z-scale factor constant, while increasing the curvature factor, resulted in an increase in CS-HFOV.
- Holding z-scale factor and curvature factors constant, while increasing the x- and y-scale factors, resulted in an increase in CS-HFOV.
- Holding x- and y-scale factor and curvature factors constant, while increasing the z-scale factor, resulted in a decrease in CS-HFOV.
- Pupil height has very little effect on CS-HFOV, which is favorable because that means we have small variance in the equatorial region of our reflector. This is a design trait that translates to assuming lower distortion exists in the main region of detection.

The effect of parameter values on CS-VFOV are as follows (increase means better):

- Holding x- and y-scale factor and z-scale factor constant, while increasing the curvature factor, resulted in an increase in CS-VFOV.
- Holding z-scale factor and curvature factors constant, while increasing the x- and y-scale factors, resulted in a decrease in CS-VFOV.
- Holding x- and y-scale factor and curvature factors constant, while increasing the z-scale factor, resulted in a decrease in CS-VFOV.
- Raising the pupil height resulted in an increase in CS-VFOV.

Table 1 Results for 30 mirror configurations

Mirror	r	a	b	c	Pupil	C-HFOV	CS-HFOV	CS-VFOV	Area Diff.	Pixel Eff.
1	57.15	1.0	0.6	1.1	(0, 192.895, 40)	33.6915	212.8798	134.298	525.718	90.9148
2	57.15	1.2	0.6	1.1	(0, 225.631, 40)	34.7806	213.4985	120.2501	618.886	91.0328
3	57.15	1.4	0.6	1.1	(0, 258.951, 40)	35.5025	213.9418	108.5237	713.378	91.1036
4	57.15	1.0	0.9	1.1	(0, 217.692, 40)	29.9225	209.9731	142.3739	881.842	90.1907
5	57.15	1.2	0.9	1.1	(0, 247.278, 40)	31.7685	211.5746	129.3702	1028.85	90.3641
6	57.15	1.4	0.9	1.1	(0, 278.100, 40)	33.0746	212.6844	118.0754	1175.73	90.4895
7	57.15	1.0	0.6	1.2	(0, 194.496, 40)	33.4248	212.8094	144.0401	538.36	91.2207
8	57.15	1.2	0.6	1.2	(0, 227.003, 40)	34.5792	213.5901	130.182	634.892	91.3187
9	57.15	1.4	0.6	1.2	(0, 260.149, 40)	35.2909	213.3899	118.4657	715.027	91.5858
10	57.15	1.0	0.9	1.2	(0, 220.823, 40)	29.5056	209.6476	154.0401	888.597	90.6297
11	57.15	1.0	0.6	1.1	(0, 189.878, 50)	33.7847	212.5246	151.4483	505.153	91.1873
12	57.15	1.2	0.6	1.1	(0, 223.161, 50)	34.8294	212.9957	137.3196	596.924	91.2776
13	57.15	1.4	0.6	1.1	(0, 256.862, 50)	35.5255	213.2946	125.3549	689.274	91.3378
14	57.15	1.0	0.9	1.1	(0, 214.579, 50)	30.0577	210.1754	153.461	857.272	90.3909
15	57.15	1.2	0.9	1.1	(0, 244.691, 50)	31.8593	211.6083	140.2159	999.864	90.5652
16	57.15	1.4	0.9	1.1	(0, 275.898, 50)	33.1358	212.5916	128.6535	1142.48	90.6905
17	57.15	1.0	0.6	1.2	(0, 191.524, 50)	33.5166	212.6517	159.5219	522.339	91.4233
18	57.15	1.2	0.6	1.2	(0, 224.567, 50)	34.6262	213.195	146.0054	616.669	91.5144
19	57.15	1.4	0.6	1.2	(0, 258.088, 50)	35.3565	213.2353	134.4495	707.961	91.6243
20	57.15	1.0	0.9	1.2	(0, 217.789, 50)	29.6399	209.8357	163.9189	867.052	90.8072
21	57.15	1.0	0.6	1.1	(0, 186.175, 60)	33.8883	211.8818	169.6191	484.734	91.4745
22	57.15	1.2	0.6	1.1	(0, 220.149, 60)	34.8729	212.0712	155.89	574.976	91.543
23	57.15	1.4	0.6	1.1	(0, 254.322, 60)	35.535	212.1906	144.2088	665.4179	1.5908
24	57.15	1.0	0.9	1.1	(0, 210.78, 60)	30.2239	210.3021	164.6718	829.148	90.6338
25	57.15	1.2	0.9	1.1	(0, 241.55, 60)	31.9651	211.4626	151.3013	967.384	90.8063
26	57.15	1.4	0.9	1.1	(0, 273.238, 60)	33.2042	212.4868	139.6005	1111.33	90.8805
27	57.15	1.0	0.6	1.2	(0, 187.868, 60)	33.6215	212.1797	175.3828	504.256	91.6637
28	57.15	1.2	0.6	1.2	(0, 221.59, 60)	34.6736	212.4512	162.5479	597.642	91.7348
29	57.15	1.4	0.6	1.2	(0, 255.576, 60)	35.383	212.6245	151.58	691.299	91.784
30	57.15	1.0	0.9	1.2	(0, 214.076, 60)	29.7912	210.0998	173.7272	848.062	90.9549

The first 10 are for Pupil Height = 40. The second 10 are for Pupil Height = 50. The third 10 are for Pupil Height = 60. Mirror # 21 is the selected prototype design. This mirror has the lowest Area Difference and third highest CS-VFOV. The lowest Area Difference means a more uniform distribution of camera pixels and higher spatial resolution

The effect of parameter values on Area Difference are as follows (increase means worse):

- Holding x- and y-scale factor and z-scale factor constant, while increasing the curvature factor, resulted in an increase in Area Difference.
- Holding z-scale factor and curvature factors constant, while increasing the x- and y-scale factors, resulted in an increase in Area Difference.
- Holding x- and y-scale factor and curvature factors constant, while increasing the z-scale factor, resulted in an increase in Area Difference.
- Raising the pupil height resulted in a decrease in Area Difference.

The effect of parameter values to Pixel Efficiency are negligible for all tested mirror shapes. We expected this because the mirror is matched to the pupil location to enforce the camera's full angle of view.

For the Head and Shoulders ROI, we observed the highest spatial resolution in the center with a steady decrease moving laterally outward until reaching zero resolution at $s = 2\pi/9$ and $s = 7\pi/9$. It makes sense that lower resolutions are in the near corners as previously discussed. The values for when $s = 0$ above, and throughout this section, correspond to the viewable edges of the mirror. We do not do anything with these values as they only confirm the fields of view with no discrepancies noted for falling outside the range of the computed CS-HFOV or CS-VFOV.

The results of spatial resolution behavior for the LHS Wall ROI and the RHS Wall ROI both show high resolution from the center of the reflector decreasing toward the edges as expected. This result means we have higher resolution to detect heat sources in the far corners of the container. Results for the Back Wall ROI are as expected with the highest resolution in the center of the ROI decreasing as we move to the edges. The slope of the curves is relatively flat for this region meaning mostly uniform resolution. It follows that the Floor ROI has similar spatial resolution behavior in the two regions closest to the Back Wall ROI since they are also in the main equatorial region of the mirror. The concern is the bottom floor area, just below the reflector, which has significantly lower resolution than any other ROI. There is nothing unusual about the spatial resolution for the Roof ROI. It is highest in the center and decreases moving outward.

For the purposes of prototyping a mirror we selected the design that has the lowest Area Difference value and the third highest CS-VFOV. There are only 4 degrees of separation between the highest CS-HFOV and the lowest with all mirror shapes above 209 degree horizontal field of view making it a metric not used for selection. It turns out that the Area Difference is the metric that differentiated the mirror shapes the most. In choosing the mirror with the lowest Area Difference, we select the mirror with the highest spatial resolution. This was observed when comparing spatial resolution plots between mirror shapes. The rationale is that a smaller Area Difference directly translates to a more uniform distribution of camera pixels to image the reflector. Moreover, the selected mirror is more compact in size with a shorter pupil to reflector length and had the highest pupil height, meaning less obscuration of the scene by the camera. The prototype mirror parameters are: $r = 57.15$ mm, $a = 1.0$, $b = 0.6$, $c = 1.1$ and Pupil = (0, 186.175 mm, 60 mm).

Prototyping was conducted in two ways that can be used for high-volume production [8]. The first is to model the mirror in a CAD program that then controls the milling of a form in either aluminum or wood. We found it preferable to use aluminum so that we can both test it directly after polishing, and then use it as the block for

vacuum thermo-forming an acrylic sheet which then can be coated with an IR reflective material. The second is simply 3D printing the mirror, filling and polishing imperfections, and the coating it (Fig. 7).

4 Embedded machine learning

Detection based on infrared and visible image fusion is widely used for many real-life applications [47–49]. Our goal is to automatically detect people in images captured by our custom designed infrared catadioptric sensor. We consider a detector to be robust if it can handle a wide variety of articulated poses, complex backgrounds, partial occlusions, and scale differences. Our goal is not facial recognition, which is a very active area of research [50], because the low-cost LWIR imagers do not have enough pixels.

Early human detection and tracking in computer vision relied on modeling people through an analysis of blobs which represented spatial and color mapping. This was done by Gaussian distribution analysis to group pixels as members of a blob. Significant contributions to this type of real-time tracking is attributed to Wren et. al [51] who developed the highly cited system called Pfinder. The central aspects of their approach are the learning of a scene by acquiring a sequence of video frames that do not contain a person, then searching for large deviations from this model by measuring the Mahalanobis distance to identify regions of interest for analysis, and finally identifying head, hands, and feet locations based on 2D contour shapes. We investigated a similar approach, but found that this method is challenged by using very specific parameters that make it difficult to generalize for an autonomous application, especially one in real-time. Moreover, contour shape analysis is noisy and has difficulty generalizing which we found is exacerbated in low-resolution images.

With improved computational power in the early 2000's, state-of-the-art human detection methods included Haar wavelet-based Adaboost cascade [52], histogram of oriented gradient (HOG) features combined with a linear SVM [53], artificial neural network (ANN) using local receptive

Fig. 7 Four 3D-printed and then aluminum coated mirrors. Left to right: XL, L, M, S

fields (LRF) [54], and key point detection methods [55]. These seminal works have withstood the tests of extreme datasets and scrutiny as observed from the amount of citations received and application implementations in the last couple of decades. Also, they have been optimized for many applications outside their first use which shows their adaptability to perform various classification tasks.

Tens of thousands of papers are being published each year on applications of machine learning. Often that machine learning is being performed remotely via cloud computing, but for sophisticated Internet of Things (IoT) applications the images and/or sensor data is intended to be processed locally. The Support Vector Machines (SVM) approach has less classification complexity than Convolutional Neural Networks (CNN) [56] by taking advantage of only a small amount of training data, called support vectors, to construct the optimal hyperplane decision boundary [57] which can make it more amenable to edge-computing implementations.

There is less research on surveillance using fisheye cameras [58] relative to perspective cameras. The relative dearth of omnidirectional images to train CNNs means that deep learning approaches have not been used as much for omnidirectional human detection [59–61] and standard human detectors are not directly applicable to top-view fisheye frames [62]. For these reasons [63] use an SVM human detector with an HOG descriptor.

Despite a heavy focus on human detection in computer vision, little research has been conducted on human detection in catadioptric images for security applications, especially those captured in low resolution. Omnidirectional systems are the most commonly used catadioptrics which capture a warped version of the 360° scene onto a donut-shaped image. This warped image is usually transformed into a panoramic or a sequence of perspective images before applying classification techniques. As part of the proposed algorithm in [64], a direct approach for human detection involves HOG features and SVM classification. The HOG computation undergoes a modification using a Riemannian metric to form an omnidirectional (non-rectangular) sliding window. This method has some interesting aspects that are applicable to our problem as we will see that human objects in our dataset are skewed near the edges of the reflector. Using training and search windows that conform to match the geometry of the reflector may improve detection performance.

One way of separating the task of human detection is by methods that use background subtraction and those that perform direct detection [65]. This is certainly an important consideration in our application. Some background subtraction methods can be versatile in that no background model needs to be learned. In contrast, direct detection methods are attractive because features can be

immediately extracted and fed to a classifier. Ultimately, a decision can be determined after analysis of performance, computational, and battery expenses for each method in a test environment.

In general, we are cautious about methods that require full scans of the image, multiple passes over the same data, or cookbook/dictionary methods which require the most long-term memory. The current state-of-the-art methods unfortunately use full scans at multiple scales. The work-around is to find a sufficient search window overlap and size that minimizes false alarms, has high detection, and reduces computations. Since computations can be separated into parallel cores by treating pixels and cells of pixels as independent of each other, potential gain is sometimes minimal. Additionally, it is possible to take advantage of features with highly discriminative power by reducing feature vector complexity through Adaboost methods [66]. These methods use a combination of weak learners to build a stronger classifier. Finally, for the plain fact that many of these methods are already employed in real-time embedded applications, we consider them in this work.

The research of Tang et al. [67] is similar to ours in that they use an infrared catadioptric sensor to autonomously detect a human. The authors use HOG features, a Support Vector Machine (SVM), and a particle filter to track a human. The HOG features are extracted directly from the image without unwarping. Instead of mapping the target, whose size changes based on its position relative to the center of the reflector, nonlinearity is handled using a particle filter known as a Sequential Monte Carlo (SMC) method. This method uses prior information about an object's state to formulate Bayesian models. The posterior distribution updates are computed as the data becomes available. SMC differs from the well-known methods of Kalman Filter and Hidden Markov Model (HMM) Filter, because Kalman and HMM assume the data can be modelled by a linear Gaussian state-space model and SMC does not [68]. The authors found that a fixed reference template outperformed a filter with an adaptive reference template due to the adaptive method being sensitive to background changes.

4.1 Data collection

We collected a dataset that includes 30 videos inside of two shipping containers over a six-month period (Nov 2018 to Apr 2019). We use two forty-foot containers in an open parking lot located at 37°16'37", 76°43'11" (Fig. 2). The videos were captured at various times of the day and night with different outside ambient temperatures to provide a sample of different backgrounds with variability in detected radiance. The persons in the scenes are

in various poses including side profile, back profile, head on, bent over, and squatting down. In addition to moving around in the container, the humans are partially occluded by container contents which include desks, chairs, wardrobes, and tables. Partial occlusion occurs during the act of picking up and moving chairs and climbing on top of container contents. A montage of image samples can be seen in Figs. 8 and 9.

During data collection, the camera is positioned directly above the door (as seen in Fig. 2) in three positions: center, left corner, and right corner. These positions were selected because they offer the most protection to damage during loading and unloading and provide the best vantage point for intruders' most probable point of entry. The dataset consists of 19,485 images (13,032 positive and 6453 negative examples) as captured by five different mirrors. Table 1 provides a summary of the data collected with a brass orb reflector (seen in Fig. 4) which includes 34 background

sequences and 51 foreground sequences. A sequence is a set of consecutive frames that have no person (background) or a person in the frame (foreground). Table 2 provides a summary of images captured by four 3D-printed then aluminum coated reflectors (Fig. 7) of different geometries. We refer to these four aluminum coated mirrors based on their size (extra-large = XL, large = L, medium = M, and small = S). Table 3 provides size and geometric information about these four reflectors.

Montages of 36 daytime sample images and 21 nighttime sample images (Figs. 8 and 9) show the variability in backgrounds as well as human poses within the dataset. An initial analysis on these 57 images highlights shape and intensity differences of people in the dataset by location in the image. We attribute the variability in the measurements to pose articulation. We defined seven regions where we find the human intruder for sample images and refer to these seven regions as FacePosition. The width,

Fig. 8 Montage of daytime images. Note that there can be quite significant difference in the thermal signature of the container roof depending upon whether it is sunny or cloudy since temperature determines gray scale value in the thermograms. The main feature

that the eye is drawn to in these images is the fixture holding the Seek Thermal camera, which is top-center in each image. The person is the small, bright feature in the lower part of each image

Fig. 9 Montage of nighttime images. Again the IR camera and fixture is in the top center of each image, with the human at a variety of different positions and poses in the lower part of each image

height, and max intensity of the intruder's head for each FacePosition is compared by time of day using the MATLAB *imtool* function.

A comparison of FaceHeight, which is measured as number of pixels, with respect to FacePosition for day and night shows an approximate range of 10 to 60 pixels with tighter groupings according to within FacePosition regions. This is to say, FaceHeights captured in a certain FacePosition, like MiddleHigh for example, are more closely grouped than the full range of measured FaceHeight values. FaceWidth, also measured in pixels, shows an approximate range of 5 to 45 pixels. In general, we observe that the relationship between FaceHeight and FaceWidth correspond in a similar fashion as compared to FacePosition albeit at different scales. MaxIntensity, which is the highest intensity value for the head object, has a range between 150 and 250. Lower variances (see Table 4) for FaceWidth and FaceHeight might be features helpful in determining human shapes from non-human shapes during supervised learning. However, higher variance in intensity values could be an alarm for approaching intensity-based methods with caution. It is not obvious from this sample of images if we can cluster these features according to FacePosition, however, it is promising that FaceWidth and FaceHeight are correlated indicating distortion is minimal and the human form can be assumed consistent with what we would expect.

Next, we applied segmentation methods systematically to the dataset. Since infrared sensors capture emitted radiation from objects in the scene which have variable intensities and detection differences according to atmospheric conditions and viewing angles, we want to understand the performance of different segmentation methods to identify regions of interest to both check its viability as an initial first step to perform simple human detection methods and

Table 2 Summary of aluminum mirror data

Mirror	30 March 2019 (0336 h)	30 March 2018 (1255 h)	31 March 2019 (1659 h)
XL	429 foreground (positive) frames 2 Foreground Sequences	830 foreground (positive) frames 3 Foreground Sequences	383 foreground (positive) frames 1 Foreground Sequence
L	357 foreground (positive) frames 2 Foreground Sequences	392 foreground (positive) frames 1 Foreground Sequence	432 foreground (positive) frames 1 Foreground Sequence
M	230 foreground (positive) frames 1 Foreground Sequence	658 foreground (positive) frames 2 Foreground Sequences	289 foreground (positive) frames 1 Foreground Sequence
S	271 foreground (positive) frames 1 Foreground Sequences	687 foreground (positive) frames 2 Foreground Sequences	350 foreground (positive) frames 1 Foreground Sequence
Outside	Ambient Temp. 44°F Conditions: Clear	Ambient Temp. 76°F Conditions: Overcast	Ambient Temp. 54°F Conditions: Overcast

MirrorXL includes 1642 foreground (positive) frames from 6 foreground sequences. MirrorL includes 1181 foreground (positive) frames from 4 foreground sequences. MirrorM includes 1177 foreground (positive) frames from 4 foreground sequences. MirrorS includes 1308 foreground (positive) frames from 4 foreground sequences. Location is Williamsburg, VA (37°16'37" × 76°43'11") at the dates and times shown

Table 3 Summary of four aluminum-coated mirrors

Mirror	Dimensions (mm)	Parametric Eq. Vals			Camera position values (mm)			
		a	b	c	r	x ₀	y ₀	z ₀
S	114.3 × 67.31 × 60.325	1.0	0.6	1.1	57.15	0.0	186.175	60.0
M	114.3 × 101.6 × 57.15	1.0	0.9	1.2	57.15	0.0	214.076	60.0
L	133.35 × 101.6 × 69.85	1.2	0.9	1.1	57.15	0.0	241.55	60.0
XL	158.75 × 101.6 × 82.55	1.4	0.9	1.1	57.15	0.0	273.238	60.0
Mirror	C-HFOV (deg)	CS-HFOV (deg)	CS-VFOV (deg)	Pixel efficiency				
S	33.8883	211.8818	169.6191	91.4745				
M	29.7912	210.0998	173.7272	90.9549				
L	31.9651	211.4626	151.3013	90.8063				
XL	33.2042	212.4868	139.6005	90.8805				

These four mirrors are 3D-printed using ABS, sanded to remove print lines, imperfections filled with bondo, wet-sanded, and then coated with aluminum by H. L. Clausen Inc. Optical Coatings [8]

to see if it could be used to enhance direct methods. Also, in our application, it is not clear there are no unanticipated adverse effects of using a reflector to capture emitted radiation. We look at separating foreground from background as a way to facilitate feature extraction for scenarios when no contrast exists. Thus, we try edge detection, threshold segmentation, and frame differencing techniques.

We reduce bias or variance in the procedure by first selecting pre-processing steps that allow subsequent routines to be executed sequentially with consistent results. We experimented with histogram equalization methods as a low-level pre-processing step for each method. However, we found that this step is not necessary for all methods.

Initial analysis showed that standard thresholding methods which include global, adaptive mean, adaptive Gaussian, and binary do a poor job segmenting the images in this dataset. Another segmentation method, which uses an active contour model evolves a curve so that it separates foreground from background based on the means of the two regions [69]. The technique is very robust to initialization and returns good results when there is a difference between the foreground and the background means. Chan and Vese [69] developed such an active contour model based on Mumford-Shah segmentation techniques and the level set method. It does not use the classical methods of edge detection to stop the evolving curve on the desired boundary and claims to not need very smooth boundaries defined by a gradient.

Active contour segmentation [70] does a fairly good job for images in our dataset. In images where enough contrast between foreground and background exists the human object is isolated. We note that this method isolated the torso of the human, which corresponds to the brightest part. This method did a particularly excellent job for the images captured when outside conditions were fair or windy and temperatures were in the mid to low 50s.

We found Sobel, Prewitt, and Roberts edge detection work well enough to pursue as a possible enhancement after testing on more data. Our results show sensitivity values between 0.015 and 0.040 using built-in MATLAB functions with both horizontal and vertical edge directions perform equally well for all reflectors and for all ambient conditions. We do run into implementation challenges, just like with thresholding, where we need a way to auto-find appropriate sensitivity values. We found the challenge here to be the amount of noise in the images and the low resolution. Roberts edge detection left the most noise in the image. The amount of noise was not found to be an issue that could not be cleaned up by morphological operations. It was easy to find a person using a sliding bar to select the sensitivity value, but more processing steps may be required to clean up the jittery human form as we saw in the results for March 31, 2019 taken at 1659 hours and

Table 4 Standard deviation and variance of Face Position

	Width SD	Width Var	Height SD	Height Var	Intensity SD	Intensity Var
LeftCenter	2.82	6.84	7.05	42.63	21.91	411.35
LeftLow	4.98	18.59	10.84	88.09	7.35	40.50
MiddleHigh	4.54	19.21	4.87	22.18	29.92	835.33
MiddleCenter	5.15	25.45	6.72	43.30	18.04	311.96
MiddleLow	7.94	50.42	5.38	23.16	25.05	502.16
RightCenter	1.58	1.24	7.38	27.25	0.71	0.25

The values in this table represent standard deviation and variance for head width, height, and max intensity for all 57 test images. The values for day versus night align with these values, thus, we do not depict those here. Intensity shows the highest variability

another on March 30, 2019 at 0336 hours. These images were captured with aluminum coated 3-D printed reflectors. We noticed that human objects captured by these reflectors had a glow to them. While this makes the human object easy to see in the image, it complicated edge detection [8].

4.2 Human detection approach

Our initial approach was to apply two broad methods to conduct human detection inside of a shipping container. First we isolate the foreground and then use features on that object to determine if it is a person or not. We divide this approach into two problems. Since we are not performing continuous observation like watching a street corner or train station, we can process images as short sequences or as single frames. If we exploit the information contained in a two-frame sequence, we can use frame differencing or motion methods to locate the target and then apply feature extraction and binary classification. This assumes that the frames at $t - 1$ (frame at a previous time step) and t (current frame) do not both contain a static target. If we perform detection on a single frame, we need an adaptive background model on standby to use whenever we detect a potential break-in with other sensors.

The background inside a shipping container varies by time of day, ambient conditions, container orientation, and other factors. While the change in background is dynamic, it is usually gradual. Thus, it is possible to capture background information at periodic intervals to be prepared for the task of detection. We look at both methods despite their obvious drawbacks because they have advantages in computational expense.

The more sophisticated approach is to use direct methods that do not require segmentation. For this approach we test the HOG feature with additional T Channel information as in the T_{π} HOG feature proposed by Baek et al. [71]. The T Channel uses intensity information within cells defined as an aggregated version of a thermal image. We choose the HOG feature over Haar features because of its

finer detail to capture edge information. Haar features are limited to edges in four directions: vertical, horizontal, and two 45 degree diagonals [53]. This makes Haar features efficient for tasks such as face recognition because faces have well-defined structure. The finer resolution of HOG features makes it robust to pose variation and built into the algorithm is a block normalization step that makes it robust to illumination changes. While illumination changes usually do not affect LWIR imaging, ambient conditions do, so performing block normalization puts feature extraction for images captured under different conditions on a level playing field.

When we started to build a detector, it occurred to us that simple shape features were not discriminative enough to cluster human objects versus all other objects. It was also clear that building these features was time consuming via trial and error. Therefore, this method would lead to a model that is trained only for a specific dataset and would not generalize well to new observations. Furthermore, we verified the HOG feature does an excellent job capturing the human shape even slightly distorted as captured by our catadioptric sensor. Its two major challenges are detecting people on the extreme edges of the reflector and finding a person under high humidity conditions and hot backgrounds, as expected [72]. Containers often transit tropical conditions en route to Panama and Suez canals.

We found humidity to be an interesting challenge when processing imagery captured when it had been raining all day. The background seemed almost washed out because any IR waves emitted from scene objects seemed to be attenuated to zero or scattered before reaching the camera's detector. This was not the case for the human in the scene. Nonetheless, while the human object can be seen in the image, the hazy separation between the target and background gave our detector trouble. The fix is foreground detection or background subtraction which can still be done to perform feature extraction and then classification.

These initial results led us to a refined approach of using the HOG feature as our only shape feature for human

detection. We propose to modify this feature by adding intensity information as discussed below. To overcome the challenges of failed detection for human objects captured on the extreme edges of the reflector an approach as suggested by Cinaroglu and Bastanlar [64] to transform the training and search windows to match the geometry of the reflector in the catadioptric system could be a worthwhile pursuit. We do not attempt this method, but we found it encouraging that with more training windows our approach did find people on the edge of the reflector as long as the head and shoulders were captured. If the image only had the head of human object visible there was not enough information to classify the object as a human. We approach the challenge of detecting a human intruder against a hot background or humid conditions by using frame differencing or background modeling and then performing segmentation to give an advantage to the HOG feature extraction. After feature extraction we perform classification using a linear support vector machine (SVM), also discussed below.

To describe a human detector using a direct method with the histogram of oriented gradients (HOG) global feature we use three steps:

1. the method of feature extraction from positive and negative images,
2. the training of a linear support vector machine, and
3. the search method on new images.

Implementation of this detector follows [53] and [73]. We found that the HOG feature on its own does not perform well for our application because it does not capture pixel intensity information. Thus, modifications to McCormick's code [73] include augmenting the HOG feature with a T-Channel descriptor to the feature vector as described in [71]. We call the feature in this detector HOG+ which includes the standard HOG feature plus the T-Channel feature. This adaptation improves performance by exploiting aggregate information from the pixel intensities within cells. We do not include pixel position information as proposed by Baek et al. [71] because we found intensity information was enough to find people in our dataset. Also, the large 3780-value feature vector for HOG combined with the 128-value feature vector for T-Channel creates a 3908-value feature vector which we felt should be kept as lightweight as possible for embedded use.

Feature extraction begins by capturing 66×130 pixel training windows from both positive and negative frames. Positive frames contain a person and negative ones do not. The true size of training windows is 64×128 pixels, however, an extra pixel around the outside is required to compute gradients along the edges. One point to make here is that the dataset has many sequences of a person

or persons in the container. We did not find it necessary to use every pose in a sequence as similarity between frames was enough to identify the person. So, after splitting the data into training and test sets, we selected a variety of poses from the training set to capture training windows and then tested the classifier on the images in the test set. Negative training frames were captured from background only images in the dataset. We felt it would enhance performance to use actual backgrounds from the dataset as will be seen there are hard examples which will need to be addressed by further analysis but using actual backgrounds to form negative training windows seemed to work well. We will also show that there are times when the classifier incorrectly classifies a portion of the background as a person. One would think that the portion that is misclassified would have a human shape, but this was not always the case. This is why we believe future implementations will require some kind of background subtraction step before feature extraction.

Collecting training windows is a tedious process. We wrote MATLAB code that opens an image, allows the user to draw a box over the desired area, then an automatically cropped window of the correct dimensions is saved for training. It is still a manual process but much faster than using the rectangular crop function in MATLAB which uses pixel locations found by opening each image, approximating coordinates to use, then executing the crop function after manual input.

The HOG part of our HOG+ feature is defined as the histogram of the magnitude sum for gradient orientations in a cell. It is widely used as an efficient feature for pedestrian or vehicle detection. It outperforms Haar features because Haar are coarse features and do not have orientation information. It uses four parameters: number of histogram bins, cell size (number of pixels), block size (number of cells) and window size (number of pixels). Dalal and Triggs suggest that the cell size should be about the size of a person's limb and that the block size should be about two to three times the size of a limb. We use a cell size of 8×8 pixels and a block size of 8×16 cells for test purposes. In addition to these parameter values being the most commonly used, recall from data exploration that the head size can be as small as 5×10 pixels in our dataset. A cell size of 8×8 pixels will not capture the fine details of the smallest head, but it performs well enough, and smaller cell sizes take more computations.

The HOG feature is a 3780-value vector formed by putting the magnitudes of 64 gradient vectors for each 8×8 cell into an n -bin histogram. Gradient vectors for every pixel are computed from intensity values using a $[-1 \ 0 \ 1]$ mask in both the vertical and horizontal directions for each pixel in the entire training window. The number of bins n is usually set to 9. Histograms range from 0 to 180 degrees,

so there are 20 degrees per bin, using unsigned gradient values. The choice of $[0, 180]$ or $[0, 360]$ degree orientation bins is based on the object class. For human detection, 0 to 180 has good performance. Since gradient magnitudes are split into bins, the strongest gradients will have a bigger impact on the histogram. Contributions of gradient vectors to bins is done proportionally between the two closest bins. The forming of a histogram from gradient vectors is a way to quantize the 64 vectors with 2 components each into a string of 9 values. The histogram does not encode the locations of each gradient within the cell, it only encodes the distribution of gradients within the cell.

The next step is to normalize the histograms. Normalization is performed locally to increase robustness to intensity changes between frames and variation between foreground-background contrast. Histograms are not normalized individually. Instead, they are grouped into blocks and normalized based on all histograms in a block. This is done by concatenating the histograms of 2×2 cells within a block into a vector with 36 components and dividing by its magnitude. We use a standard L2-Normalization which along with L2-Hysteresis and L1-Square Root were empirically determined to perform the best by Dalal and Triggs.

The T-Channel, originally proposed by Hwang et al. [74], computes the sum of pixel intensities within an 8×8 cell. It was suggested as a feature to classify pedestrians in infrared images taking into account that pedestrians tend to have higher pixel intensity values than backgrounds. The images in our dataset highlight this fact even though they are captured by a catadioptric imaging system. We see especially good contrast for the images captured using the aluminum coated mirrors. To summarize, the 64×128 pixel training window is divided into 7 blocks horizontally by 15 blocks vertically for a total of 105 blocks. Each block contains 4 cells with a 9-bin histogram for a total of 36 values per block. The HOG descriptor vector is then $7 \times 15 \times 4 \times 9 = 3780$ values. The 128 intensity sums per 16×8 cell of the T-Channel feature is concatenated to this vector forming a 3908-value feature vector for HOG+.

The optimization objective of the SVM is to maximize the margin. The margin is defined as the distance between the separating hyperplane (decision boundary) and the training samples that are closest to this hyperplane, which are the so-called support vectors. Given the training set

$$\mathcal{S} = \{(x^{(i)}, y^{(i)})\}_{i=1}^L$$

with L samples, the SVM is trained to classify input data $x^{(i)}$ as positive ($y^{(i)} = 1$) class or negative ($y^{(i)} = -1$) class, where

$$x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \dots, x_N^{(i)}]^T \in \mathbb{R}^N$$

and

$$y^{(i)} \in \{-1, +1\}.$$

If the input data $x^{(i)}$ is mapped to a higher dimensional feature space as $\phi(\cdot)$ then the decision function of the SVM is defined by

$$f(x) = w^T \phi(x) + b$$

where

$$\phi(\cdot) \in \mathbb{R}^D, D \gg N, w \in \mathbb{R}^D$$

is the weight and $b \in \mathbb{R}$ is the bias. The SVM is trained by finding optimal solutions of w and b which maximize the margin between the two classes.

Transforming the training data onto a higher dimensional feature space via the mapping function $\phi(\cdot)$ is known as a kernel trick whereby we train a linear SVM model to classify the data in this new feature space and then use the same mapping function $\phi(\cdot)$ to transform new, unseen data to classify it using the same linear model. Since the computation of new features is expensive, especially when using high-dimensional data, the kernel trick is used to help overcome the problem of solving the quadratic programming task to train the SVM. In practice, the dot product $x^{(i)T} x^{(j)}$ is replaced by $\phi(x^{(i)})^T \phi(x^{(j)})$ leading to the kernel function

$$k(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)}).$$

When dealing with the case of nonlinearly separable classes slack variables ξ_i are introduced. The motivation is that linear constraints need to be relaxed to allow convergence of the optimization in the presence of misclassifications under the appropriate cost penalization. L2-SVMs use the square sum of the slack variables ξ_i in the objective function forming the following optimization problem:

$$\text{minimize } \frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^M \xi_i^2,$$

$$\text{subject to } y_i(w^T x_i + b) \geq 1 - \xi_i,$$

$$\text{for } i = 1, \dots, M.$$

We use the variable C to control the penalty for misclassification. Large values of C correspond to large error penalties and small values indicate that we are less strict about misclassification errors. We can then use the parameter C to control the width of the margin and therefore tune the bias-variance trade-off. We found $C = 1.0$ and the L2-SVM loss function work well on our dataset.

The rationale behind having decision boundaries with large margins is that they tend to have a lower generalization error whereas models with small margins are more prone to overfitting [75]. The high generalization ability has led to the popularity of the SVM over conventional

methods especially when the amount of training data is small. It also offers the advantages of adaptability to various classification problems by changing kernel functions and it has a global optimal solution.

After training windows are captured from training images, the HOG+ feature is extracted from each training window, and an SVM is trained to classify between two classes (person v. no person), the detection phase searches new images at all scales and locations to localize human objects.

First, it is important to note that the HOG+ descriptor window is never resized. Instead, the image is rescaled many times where each image scale is 5% smaller than the previous one until the resulting image is too small to fit a descriptor window. An image scale of 1.00 represents the full image where the idea is to look for persons who are small within the image or farther from the camera. At smaller scales, we search for people who appear larger in the image, meaning closer to the camera. For each scale, the image is downsampled or resized, then HOG+ descriptors are calculated for every possible detection window within the resized image. The HOG+ descriptors are fed to the linear SVM to determine if the detection window contains a person. A person is detected when the value of the computed probability is above a threshold set by the designer. The implementation here does not support variable detection window strides. It is stepped by 1 cell (8×8 pixels) in each direction with no overlap.

4.3 Human detector results

The performance of our human detector on unseen images was analyzed by mirror type, near/far detection of human objects, and day v. night. We trained one SVM classifier with 349 positive and 250 negative training windows (Figs. 10 and 11 respectively). We selected various human poses for training but did not choose all poses in the dataset by design. All training windows came from videos collected with the BrassOrb mirror (Fig. 4).

Test results were recorded for the following number of test images by mirror: BrassOrb (43 images), MirrorXL (44 images), MirrorL (32 images), MirrorM (34 images), and MirrorS (43 images). Test images for each mirror include images for all times of day and weather conditions as per the data collected.

During testing we recorded the threshold, the scale of the detected window, and counted false positives, true positives, and false negatives. The threshold is the value that separates the two classes. Since -1 represents non-persons and 1 represents persons, a value of 0 would normally be selected, however, the threshold can be set higher to say 0.4 to reduce false positives. The actual threshold values for our code range between -25 and

45. This is because we adapted the code to include the T-Channel descriptor which changed the range of threshold values. The range of threshold values where we found people using the adapted code is between 4 and 40. This maps to values of approximately -0.1714 to 0.8571 on the original -1 to 1 scale. Lower thresholds allow more detected windows to be found and higher thresholds reduce the number of detected windows. We tested several threshold values for every test run and recorded only the one that resulted in a low number of detected windows while finding the person. This allowed us to conduct a fair comparison between mirrors.

The mean, median, variance, and standard deviation of threshold values gives us insight into the detection sensitivity of the different mirrors. As expected, from our experimentation of shapes and intuition from their distortion as viewed inside of a checkerboard mock-up (see Figs. 4 and 7), the smallest mirror, MirrorS, has the highest threshold variability of 74.3186 and highest standard deviation of 8.6208. Not surprisingly, the BrassOrb shows the smallest variance of 13.6811 and standard deviation of 3.6811. Since the training windows are formed from BrassOrb data this indicates that further analysis should match training windows to the mirror for best results. The threshold variances and standard deviations for mirrors MirrorM, MirrorL, and MirrorXL are all very similar.

A detected window is simply where a person is identified. For our 480×640 pixel images, we search 24668 windows in about 11 seconds on a 3.2 GHz Intel Core i5. Our code is not optimized to run on parallel cores and is written for us to understand the process and results not for real-time implementation purposes. The detected window is recorded as a row vector which has coordinates and dimensions to be plotted on the image. The values are recorded relative to the original image dimensions.

The scale of the detected window is the size of the downsampled image relative to the original image. The number of detected windows per scale is an output of the code. We only record the range of scales to verify our perception of where the person should be detected according to its distance from the camera. As it turns out, we detect people who are both near and far from the camera at different scales in the same test run. This can be seen by the different sized windows plotted in the results. We also detect the person multiple times per scale. There is not conclusive evidence to say how mirror shapes correlate to scale of detection. All seemed to detect people in the expected scale range. We note that for distances between person and camera in this application we do not expect much variation in the size of the person on the image at distances between 5 to 15 feet. At distances of over 15 feet, the person appears small and was detected less than half the time.

Fig. 10 Montage of 349 positive training windows with humans in various poses and positions within the container

Test results for MirrorS were 65.7% true positive, 34.3% false positive, and 20.8% false negative. For MirrorM, results were 69.7% true positive, 30.3% false positive, and 15.6% false negative. For MirrorL, results were 90% true positive, 9.6% false positive, and 18.4% false negative. Results for MirrorXL were 84.2% true positive, 15.8% false positive, and 26.7% false negative. Finally, for BrassOrb there were 38.5% true positive, 61.5% false positive, and 33.3% false negative.

5 Discussion and future work

The primary contribution of this work is the design and proof-of-concept demonstration of an infrared catadioptric imaging device with physics-based machine learning algorithms that can make effective use of inexpensive infrared imagers which have relatively small detector arrays. The specific goal was to interpret the scene inside

Fig. 11 Montage of 250 negative training windows with humans in various poses and positions within the container

of a shipping container at the point of detection, i.e. in an edge-computing paradigm, and thus make efficient use of expensive communications bandwidth. The histogram of oriented gradients and T-channel (HOG+) feature was implemented for human detection using infrared catadioptric images, and was shown to be effective for various mirror shapes.

While human detection with infrared cameras is not difficult at all if cost is not a consideration, here we are exploiting very inexpensive LWIR imagers that enable real-world applications such as monitoring shipping containers. The primary design challenge is being able to

give coverage of the necessary container volume with an imager tucked up out of the way such that it will survive loading and unloading of the container. Moreover, the catadioptric imaging system described here uses reflective, rather than transmissive, optics for both cost and durability considerations.

Figure 12 shows that it is possible to detect distorted people captured on the edge of the reflector. With less than 600 training windows we correctly found people on the edge in both day and evening hours. Our system correctly classified human objects despite other hot objects in the scene. It correctly detected a person outside the door

Fig. 12 Representative human detection results, with the human automatically identified and marked with a red rectangle in each

with the catadioptric system positioned in the top corner of the shipping container. From this position, we were able to find the person as soon as they opened the door. It is important to note that this position corresponds to the corner above the left door, as looking at the container from the front. The reason this position works best between the two top corner positions is because the door is built such that the left door can only be opened after opening the right door.

Figure 12 also provides positive results of two people near the edge of the reflector and even with one of them just below the camera, and at a distance far from the camera. With no training windows of woman or children the detector correctly found both in the scene purely based on the commonality of human shapes. Finally, this method correctly found two out of three people and a person far from the camera out of focus.

False positive results include people on the edge, finding only one of three people, classifying an object in the background with a similar shape as a person, and 13 false positives of background in one image. Our first videos included Seek Thermal watermarks in the top right-hand corner of the frame, which was misclassified because of intensity miscues, which should not be a factor during field deployment.

In general, we found that the aluminum coated reflectors performed better than the brass orb. The aluminum coated mirrors detected a higher contrast between the human objects and the background. Among the aluminum coated reflectors, the larger ones performed better. Some considerations that should to be explored are smaller cell size, block overlap, number of image scales to search, and background removal before feature extraction. We recommend testing a cell size of 4×4 pixels to

empirically test detection of smaller objects that are far away. It is likely that resolution limits are reached past a certain distance from the camera.

Block overlap is a parameter that can be tested with threshold values and penalty terms using a grid search such as one performed in Python's GridSearchCV. The method is to first build a generic classifier as above and setup a parameter grid to tune parameters over a wide search area and then narrow the search based on score values such as precision, recall, or accuracy. This should be performed after more data collection and training window extraction for each mirror of interest. The search method in this proof of concept uses 28 scales of the original test image which resulted in 24,668 search windows. This was the result of a 5% reduction for each scale until the image scale was not large enough for a single search window. Finding the balance here between true positives and false positives can be another way to both improve detection and reduce computations.

We experimented with several techniques to separate foreground and background. There are many more methods, especially ones that are better suited for real time applications. We found edge detection, active contour segmentation, and frame differencing to be effective methods for this task. Our sense that this enhancement to the algorithm of HOG+ and linear SVM will improve performance.

We have described a framework for global shipping container monitoring using machine learning with multi-sensor hubs and infrared catadioptric imaging. The wireless mesh GRID Tag Architecture provides connectivity anywhere in the world which is a significant improvement to legacy methods. We discussed an infrared catadioptric imaging device and multi-sensor hub combination as an intelligent system that, when equipped with physics-based learning algorithms, can interpret the scene at the point of detection and make efficient use of expensive communications bandwidth. The histogram of oriented gradients and T-channel (HOG+) feature as introduced for human detection on infrared catadioptric images showed to be effective for various mirror shapes.

We left problems such as break-ins by angle grinder, bolt cutter, or sledge hammer to be analyzed in follow-on work since these seem straight forward by already proven signal processing methods. There are already solutions to shock detection, container tip-over, and container tracking [76–78]. The bottom line is that there are a number of use cases for putting a monitoring device on the inside of the container as opposed to the outside.

There is a critical need to know where things are in the supply chain when they are outside of a communications infrastructure. The latest industrial revolution, called Industry 4.0, is spurred by a digital supply chain, smart manufacturing, digital products, and data analytics [79]. In a landscape

of increasing supply chain complexity and manufacturing that chases cheap labor, the need for smart sensors, mobile and tracking solutions, and intelligent algorithms will give the edge to players who can access resources. An interesting race to resources has already started in Africa where Chinese manufacturing has begun to establish itself in places like Nigeria and Ethiopia [80].

Since 2010, China has been the top nation in terms of manufacturing output [81]. However, a generation under the one-child policy has shrunk the country's labor pool, raising labor costs and pushing global manufacturing to low-income countries. Meanwhile, Africa is in the early stages of a population boom creating the largest pool of labor in the world [80]. Sparked by tax breaks, cheap labor, and proximity to lucrative Middle Eastern markets, Africa can provide an appealing location for many business models.

There are also other factors at play in the global trade industry that are already shifting manufacturing out of China. The recent tariffs placed on Chinese exports has further shifted manufacturing to nearby countries Thailand, Vietnam, and Indonesia with India also posting growth of 8% year-to-year [82, 83]. These developments drive the market for companies to find a competitive edge in logistics within the ecosystem of Industry 4.0.

The data collected for this research was for a stationary shipping container, which is the most likely scenario for break-ins. However, we need more data to rule out false positives from container vibrations as it is moved around shipping yards, rail cars, and cargo ship holds. Data from these transit areas are crucial for training machine learning algorithms especially measuring the energy consumption during these activities. We think the next steps in this research are to develop a beta prototype of the inside monitoring device (Figs. 2 and 3) and collect lots of data inside actual shipping containers travelling over sea, land, and possibly air routes [8]. From these observations, we can tune algorithm sensitivity to reduce false positives.

Acknowledgements LTC Trujillo acknowledges full-time support for graduate studies and research by the US Army. The authors thank Midstream Technologies (Williamsburg, VA) for many helpful discussions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Paris C, Di Fonzo T, Llamas L (2018) A brief history of shipping. Wall Str J. <https://www.wsj.com/articles/a-brief-history-of-shipping-1516789500>. Accessed 30 July 2019

2. Grey E (2017) Cargo theft: a billion dollar problem. Ship Technol. <https://www.ship-technology.com/features/featurecargo-theft-a-billion-dollar-problem-5882653/>. Accessed 1 Jan 2019
3. Team M (2017) Cargo Theft: a billion-dollar problem. MFAME. <http://mfame.guru/cargo-theft-billion-dollar-problem/>. Accessed 26 Oct 2017
4. Edmonds J (2016) The history of the shipping container. Freightos. <https://www.freightos.com/the-history-of-the-shipping-container/>. Accessed 1 Jan 2019
5. Schreib B, Bostic A, Mitchell J, Nomura A, Sanchez M, Lin J, McClintock S, Hale T, Yazdi N, Rein D, Meng G, Wallace C, Skinner E, Hinders M (2015) Geo-referencing identification (GRID) Tag Final Report. Study funded by Bureau of Safety and Environmental Enforcement. Contract # E14PC00027
6. Schreib B, Bostic A, Mitchell J, Nomura A, Sanchez M, Lin J, McClintock S, Hale T, Yazdi N, Rein D, Meng G, Wallace C, Skinner E, Hinders M (2019) Tagging of oil under ice for future recovery final report. Study funded by Bureau of Safety and Environmental Enforcement. Contract # E14PC00028
7. Johnson RC (2015) DSP-based hubs target low-power sensor fusion apps. Embedded. <http://www.embedded.com/electronic-s-products/electronic-product-reviews/mcus-processors-and-socs/4438876/DSP-based-hubs-target-low-power-sensor-fusion-apps>. Accessed 20 Dec 2016
8. Trujillo II V (2019) Global shipping container monitoring using machine learning with multi-sensor hubs and infrared catadioptric imaging, William and Mary, Department of Applied Science Doctoral Dissertation
9. McKeon J, Hinders M (1999) Parallel projection and crosshole Lamb wave contact scanning tomography. *J Acoust Soc Am* 106(5):2568–2577
10. Malyarenko E, Hinders M (2000) Fan beam and double cross-hole Lamb wave tomography for mapping flaws in aging aircraft structures. *J Acoust Soc Am* 108(10):1631–1639
11. Malyarenko E, Hinders M (2001) Ultrasonic Lamb wave diffraction tomography. *Ultrasonics* 39(4):269–281
12. Leonard K, Malyarenko E, Hinders M (2002) Ultrasonic Lamb wave tomography. *Inverse Probl Spec NDE Issue* 18(6):1795–1808
13. Leonard K, Hinders M (2003) Guided wave Helical ultrasonic tomography of pipes. *J Acoust Soc Am* 114(2):767–774
14. Hou J, Leonard Kevin R, Hinders M (2004) Automatic multi-mode Lamb wave arrival time extraction for improved tomographic reconstruction. *Inverse Problems* 20:1873–1888
15. Leonard K, Hinders M (2005) Multi-mode Lamb wave tomography with arrival time sorting. *J Acoust Soc Am* 117(4):2028–2038
16. Leonard K, Hinders M (2005) Lamb wave tomography of pipe-like structures. *Ultrasonics* 44(7):574–583
17. Rudd K, Bingham J, Leonard K, Hinders M (2007) Simulation of guided waves in complex piping geometries using the elastodynamic finite integration technique. *JASA* 121(3):1449–1458
18. Bingham J, Hinders M (2009) Lamb wave detection of delaminations in large diameter pipe coatings. *Open Acoust J* 2:75–86
19. Bingham J, Hinders M (2009) Lamb wave characterization of corrosion-thinning in aircraft stringers: experiment and 3D simulation. *J Acoust Soc Am* 126(1):103–113
20. Bingham J, Hinders M, Friedman A (2009) Lamb wave detection of limpet mines on ship hulls. *Ultrasonics* 49:706–722
21. Bingham J, Hinders M (2010) 3D elastodynamic finite integration technique simulation of guided waves in extended built-up structures containing flaws. *Computat Acoust* 18(2):165–192
22. Leckey C, Rogge M, Miller C, Hinders M (2012) Multiple-mode Lamb wave scattering simulations using 3D elastodynamic finite integration technique. *Ultrasonics* 52(2):193–344. <https://doi.org/10.1016/j.ultras.2011.08.003>
23. Miller C, Hinders M (2014) Classification of flaw severity using pattern recognition for guided wave-based structural health monitoring. *Ultrasonics* 54:247–258. <https://doi.org/10.1016/j.ultras.2013.04.020>
24. Miller C, Hinders M (2014) Multiclass feature selection using computational homology for Lamb wave-based damage characterization. *J Intell Mater Syst Struct* 25:1511. <https://doi.org/10.1177/1045389X13508335>
25. Chahl JS, Srinivasan MV (1997) Reflective surfaces for panoramic imaging. *Appl Opt* 36:31
26. Ian K, Ezra E, Christopher W, George T (2016) Low-cost panoramic infrared surveillance system. *Proc SPIE* 10178:1017801
27. Simon B, Nayar Shree K (1999) A theory of single-viewpoint catadioptric image formation. *Int J Comput Vis* 35:175–196
28. Orges F, Driggers Ronald G, Gerald H, Keith K (2014) Performance analysis of panoramic infrared systems. *Infrared Imaging Syst Des Anal Model Test* 9071:907112
29. Rahul S, Grossberg Michael D, Nayar Shree K (2006) Non-single viewpoint catadioptric cameras: geometry and analysis. *Int J Comput Vis* 66:211–229
30. Yasushi Y, Masahiko Y (2004) Real-time omnidirectional image sensors. *Int J Comput Vis* 58:173–207
31. Hicks RA (2018) Differential methods in catadioptric sensor design with applications to panoramic imaging. Technical Report
32. Srinivasan MV (2003) A new class of mirrors for wide-angle imaging. In: Proceedings of the 2003 conference on computer vision and pattern recognition workshop
33. Hicks RA (2001) Geometric distributions for catadioptric sensor design. In: IEEE proceedings
34. Gutin M, Tsui EK, Gutin O, Wang XM, Gutin A (2006) Thermal infrared panoramic imaging sensor. *Infrared Technol Appl* 6206:62062E
35. Thibault S (2008) IR panomorph lens imager and applications. *Infrared Technol Appl XXXIV* 6940:69401A
36. Seek Thermal Compact Manufacturer's Specifications (2018). <https://www.thermal.com/uploads/1/0/1/3/101388544/compact-sellsheet-usa-web.pdf>. Accessed 12 July 2018
37. SCHOTT North America (2018). <https://www.us.schott.com/d/advanced-optics/866702a2-39e9-41a5-ac2b-c28e10f0fbff/1.1/schott-infrared-chalcogenide-glasses-irg-26-english-us-10042017.pdf>. Accessed 10 Aug 2018
38. Bouguet JY (2015) Camera calibration toolbox for MATLAB. <http://www.vision.caltech.edu/bouguetj/calib-doc/>. Accessed 10 Sep 2018
39. Heikkila J, Silven O (1997) A four-step camera calibration procedure with implicit image correction. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition
40. Born M, Wolf E (1965) Principles of optics. Pergamon Press, Oxford
41. Burkhard DG, Shealy DL (1973) Flux density for ray propagation in geometrical optics. *J Opt Soc Am* 63(3):299–304
42. Shealy DL, Burkhard DG (1973) Caustic surface and irradiance for reflection and refraction from an ellipsoid, elliptic paraboloid, and elliptic cone. *Appl Opt* 12(12):2955–2959
43. Alsaidi R, Musleh A (2014) Two methods for surface/surface intersection problem comparative study. *Int J Comput Appl* 92(5):09758887
44. W&M High Performance Computing Cluster <https://www.wm.edu/offices/it/services/hpc/atwm/index.php>
45. Baker S, Nayar S (1998) A theory of catadioptric image formation. In: Proceedings of the 6th international conference on computer vision

46. Swaminathan R, Grossberg MD, Nayar SK (2004) Designing mirrors for catadioptric systems that minimize errors. In: Fifth workshop on omnidirectional vision
47. Xin J, Qian J, Shaowen Y, Dongming Z, Rencan N, Jinjin H, Kangjian H (2017) A survey of infrared and visual image fusion methods. *Infrared Phys Technol* 85:478–501. <https://doi.org/10.1016/j.infrared.2017.07.010>
48. Dogra A, Goyal B, Agrawal S (2017) From multi-scale decomposition to non-multi-scale decomposition methods: a comprehensive survey of image fusion techniques and its applications. *IEEE Access* 5:16040–16067. <https://doi.org/10.1109/ACCESS.2017.2735865>
49. Jiayi M, Yong M, Chang L (2019) Infrared and visible image fusion methods and applications: a survey. *Inf Fusion* 45:153–178. <https://doi.org/10.1016/j.inffus.2018.02.004>
50. Loy CC, Liu X, Kim TK et al (2019) Special issue on deep learning for face analysis. *Int J Comput Vis* 127:533. <https://doi.org/10.1007/s11263-019-01179-z>
51. Wren CR, Azarbayejani A, Darrell T, Pentland AP (1997) Pfnder: real-time tracking of the human body. *IEEE Trans Pattern Anal Mach Intell* 19(7):780–785
52. Paul V, Jones Michael J, Daniel S (2005) Detecting pedestrians using patterns of motion and appearance. *IJCV* 63(2):153–161
53. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. *Pattern Recognit* 35:886–893
54. Shimizu H, Poggio T (2004) Direction estimation of pedestrian from multiple still images. In: IEEE intelligent vehicles symposium, pp 596–600
55. Krystian M, Cordelia S (2004) Scale and affine invariant interest point detectors. *IJCV* 60(1):63–86
56. Samira P, Saad S, Yilin Y, Haiman T, Yudong T, Presa RM, Mei-Ling S, Shu-Ching C, Iyengar SS (2018) A survey on deep learning: algorithms, techniques, and applications. *ACM Comput Surv* 51:36. <https://doi.org/10.1145/3234150>
57. Xu R et al. (2018) Real-time human objects tracking for smart surveillance at the edge. In: 2018 IEEE international conference on communications (ICC), Kansas City, MO, pp 1–6. <https://doi.org/10.1109/ICC.2018.8422970>
58. Galan M, Strojnik M, Wang Y (2019) Design method for compact, achromatic, high-performance, solid catadioptric system (SoCatS), from visible to IR. *Opt Express* 27:142–149
59. Ababsa F, Hadj-Abdelkader H, Boui M (2019) 3D human tracking with catadioptric omnidirectional camera. In: Proceedings of the 2019 on international conference on multimedia retrieval (ICMR '19). ACM, New York, NY, USA, pp 73–77. <https://doi.org/10.1145/3323873.3325027>
60. Masato T et al. (2019) Omnidirectional pedestrian detection by rotation invariant training. In: 2019 IEEE winter conference on applications of computer vision (WACV), pp 1989–1998
61. Demiröz BE, Salah AA, Bastanlar Y et al (2019) Affordable person detection in omnidirectional cameras using radial integral channel features. *Mach Vis Appl* 30:645. <https://doi.org/10.1007/s00138-019-01016-w>
62. Oded K, Kiryati N (2017) People detection in top-view fish-eye imaging. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–6
63. Srisamosorn V, Kuwahara N, Yamashita A et al (2019) Human position and head direction tracking in fisheye camera using randomized ferns and fisheye histograms of oriented gradients. *Vis Comput*. <https://doi.org/10.1007/s00371-019-01749-9>
64. Cinaroglu I, Bastanlar Y (2014) A direct approach for human detection with catadioptric omnidirectional cameras. In: IEEE 22nd signal processing and communications applications conference (SIU 2014), pp 2275–2279
65. Ogale NA (2006) A survey of techniques for human detection from video. *Survey*. 125:133
66. Wang ZR, Jia YL, Huang H, Tang SM (2008) Pedestrian detection using boosted HOG features. In: Proceedings of the 11th international IEEE. Conference on intelligence transportation systems. Beijing, China, pp 1155–1160
67. Yazhe T, Youfu L, Tianxiang B, Xiaolong Z, Zhongwei L (2011) Human tracking in thermal catadioptric omnidirectional vision
68. Arnaud D, de Freitas N, Gordon N (2001) An introduction to sequential monte carlo methods. In: Sequential Monte Carlo methods in practice. Springer, New York
69. Chan Tony F, Vese Luminita A (2001) Active contours without edges. *IEEE Trans Image Process* 10(2):266–277
70. Lankton S (2016) Active contour segmentation. MathWorks File Exchange. Version 1.0.0.0. Updated March 31, 2016. Retrieved March 15, 2019
71. Baek J, Hong S, Kim J, Kim E (2017) Efficient pedestrian detection at nighttime using a thermal camera. *Sensors* 17:1850
72. FLIR Technical Note (20019). Seeing through fog and rain with a thermal imaging camera: metrological effects of fog and rain upon IR camera performance. <https://www.flirmedia.com/MMC/CSV/Tech-Notes/TN-0001-EN.pdf>. Accessed 10 April 2019
73. McCormick C (2013) HOG Descriptor in MATLAB. <http://mccormickml.com/2013/05/09/hog-descriptor-in-matlab/>. Accessed 28 Mar 2019
74. Hwang S, Park J, Kim N, Choi Y, Kweon I (2015) Multispectral pedestrian detection: benchmark dataset and baseline. In: Proceedings IEEE CVPR., Boston MA, pp 1037–1045
75. Raschka S (2016) Python machine learning. Packt Publishing, Birmingham
76. Arviem (2019) Supply chain visibility via real-time cargo tracking and monitoring. <https://arviem.com/>. Accessed 28 May 2019
77. LinkLabs (2019) Container tracking systems: everything you need to know. <https://www.link-labs.com/blog/container-tracking>. Accessed 28 May 2019
78. AT&T Business (2019) Shipping containers and trailers. Online: <https://www.business.att.com/products/shipping-containers-and-trailers.html>. Accessed 28 May 2019
79. Schrauf S, Bertram P (2015) Industry 4.0—How digitization makes the supply chain more efficient, agile, and customer-focused. PWC. <https://www.strategyand.pwc.com/media/file/Industry4.0.pdf>. Accessed 6 June 2019
80. Sun IY (2017) The World's next great manufacturing center. Harvard Business Review, Harvard
81. West DM, Lansang C (2018) Global manufacturing scorecard: How the US compares to 18 other nations. Brookings. <https://www.brookings.edu/research/global-manufacturing-scorecard-how-the-us-compares-to-18-other-nations/>. Accessed 6 June 2019
82. Eli E (2019) What Country is the world's next manufacturing hub? nectar product development. <https://nectarpd.com/country-worlds-next-manufacturing-hub/>. Accessed 6 June
83. Rapoza K (2018) Trade war casualties: factories shifting out of China. Forbes. <https://www.forbes.com/sites/kenrapoza/2018/07/30/trade-war-casualties-factories-shifting-out-of-china/33f61643103e>. Accessed 6 June 2019

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