Monocular thermal imaging for pedestrian detection and ranging

Eugene Petilli *
Owl Autonomous Imaging, Inc., Fairport, New York, United States

Abstract. To react to the presence of pedestrians, an automated braking system must first find the pedestrian(s) including locations relative to the automobile both in angular position and in distance. In the daytime, cameras and radar can provide the necessary information, but this combination, which requires ambient or active illumination, fails at night. Passive thermal sensors are now being enlisted to dramatically improve imaging at night, whereas substantial effort is underway to assure proper fusing of object information from the thermal sensor with other sensors on the automobile. To simplify the acquisition of information needed to make valid automated braking decisions at night, a camera system with a thermal image sensor was developed that identifies pedestrians and labels each identified pedestrian with location and distance data. The camera utilizes a single uncooled custom microbolometer sensor and a software suite implementing artificial intelligence and machine learning capabilities, running on a combination of convolutional neural networks and fusion processing to provide the data that an automobile host computer needs to implement fast, safe, and accurate automatic braking. We present details of the system construction and operation as well as initial test results showing the potential this technology has to dramatically reduce pedestrian fatalities at night as well as augment safety across all conditions, whether day, night, fog, rain, snow, dust, sun glare, or headlight glare. [DOI: 10.1117/1.OE.62.3.031212]

Keywords: autonomous; thermal imaging; neural network; object fusion; pedestrian; night; pedestrian autonomous emergency braking.

Paper 20220930SS received Aug. 22, 2022; accepted for publication Nov. 28, 2022; published online Dec. 23, 2022.

1 Introduction

In 2019, a European paper on autosafety said, "At the moment, different autonomous emergency braking (AEB) pedestrian systems are already available on the market, based on radar, (stereo) camera, or a combination of these sensors. The performance of these systems varies significantly based on the sensors used. The first generation of camera-based systems typically switch off during low-ambient lighting conditions as classification of pedestrians in darkness is not reliable enough.[1]" In the intervening years, not much has changed. Currently, pedestrian detection systems are tested at night only in scenarios with sufficient artificial lighting to permit the systems to work. Now, the National Highway Traffic Safety Administration (NHTSA) has decided that it is time for pedestrian AEB systems to work at night without artificial lighting.

NHTSA is acting in response to a mandate included in the recently passed Bipartisan Infrastructure Law in the United States to address hazards automobiles pose to pedestrians by initiating work on a new rule that will require every automobile to include an automatic braking system that reacts to the presence of pedestrians in the path of the automobile. Because NHTSA statistics and testing show that at night pedestrians in poorly lighted areas remain at high risk of not being seen, the systems will be required to include sensing that functions in the dark.

2 Nature of the Problem

A recent report[2] published by the Governors Highway Safety Association (GHSA) highlights two important characteristics in pedestrian deaths. First, in 2020, 60.4% of pedestrian fatalities

*Address all correspondence to Eugene Petilli, genep@owlai.us

0091-3286/2022/$28.00

occurred on highway arterials, those roads that connect directly to restricted-access highways but generally have speed limits low enough to allow AEB systems to stop a vehicle before collision, and second, 76.6% of pedestrian fatalities occurred after dark, when current AEB pedestrian detections systems are ineffective. This shortcoming is further highlighted by the GHSA analysis that shows that while daylight pedestrian fatalities have risen only 15.9% (2010 to 2020), at night the increase was 63.3%.

The discrepancy in these increases might be partially explained in a recent study commissioned by the Insurance Institute for Highway Safety, which concluded that "AEB with pedestrian detection was associated with significant reductions of 25% to 27% in pedestrian crash risk and 29% to 30% in pedestrian injury crash risk. However, there was no evidence that the system was effective in dark conditions without street lighting, at speed limits of 50 mph or greater, or while the AEB-equipped vehicle was turning." [3]

As these statistics indicate, a solution to the detection of pedestrians at night, and at a range sufficient to allow protective action, is sorely needed.

3 Current Mitigation Systems

Effective pedestrian detection systems depend on object fusion, the merging of information from two or more imaging devices to find pedestrians in a viewed scene and label each of them with the distance from the vehicle. As Sec. 1 indicates, the imaging devices generally used today are a pair of video cameras arranged to take stereo images that are used to identify pedestrians and provide approximate distance information at short distances, coupled with a radar to extend the distance range and support distance acquisition when visibility is low. In favorable conditions, the data from these two modalities is sufficient to produce object fusion accurate enough for use in AEB systems. However, because of limitations of these modalities, conditions can easily degrade enough to render them ineffective.

Weather is often a factor; rain, snow, and fog obscure the camera images. Lights shining toward the camera can overload the sensors preventing acquisition of necessary information. Congestion prevents fusing radar data with individual pedestrians. Object fusion is further complicated by the differences in the resolution and repetition rate of the data from cameras and radar. Radar can operate very rapidly but with low spatial resolution, whereas cameras operate more slowly but can have high resolution. Further, if the signal from either modality is interrupted, the processor performing object fusion will be unable to operate, not only during the interruption but for a time afterward as it must reidentify all objects of interest.

Other combinations have been applied to this task, such as LiDAR supplementing or replacing the radar, or stereo thermal cameras, but these have both advantages and limitations. LiDAR can have much higher resolution than radar so fusing objects from LiDAR point clouds with camera images can be easier, but LiDAR has the same difficulties working in obscured scenes as cameras; both signals can be lost in fog or other adverse conditions. Further, LiDAR has yet to demonstrate that it is a cost-effective alternative for deployment in production advanced driver assist system (ADAS) vehicles.

On the other hand, thermal cameras acquire images based on capturing the natural passive heat emitted by every object, a capability that allows operation without active illumination and also works well in the presence of obscurants. Thermal imaging is especially advantageous at differentiating warm bodies, such as people or animals, since warm bodies have a unique heat signature as compared to other objects or backgrounds. Thus detection of pedestrians is pretty straightforward with thermal cameras, even at night. However, determination of distance requires a stereo pair or fusion with other technologies, such as radar or LiDAR, which becomes too expensive for many applications, including automotive. Today, even the lowest-cost uncooled thermal cameras are still expensive, especially when used in pairs, and unfortunately, they have an additional problem of needing to close an internal shutter periodically for recalibration. Finally, using two cameras of any type doubles the likelihood that some external event such as mud on a window will render the system unable to determine distance. In the system presented here, all of these difficulties are addressed.

4 Monocular Thermal Ranging

To reiterate the requirements, a system capable of identifying and ranging pedestrians in clear and degraded visual environments, day and night, able to see the pedestrians by their own emitted energy (no active illumination), and determine both their location and range in the scene. To maximize the object fusion success, the location and range data should be as closely matched as possible in resolution, frame rate, and scene perspective with minimum interruptions in scene acquisition for recalibration. Thus a single shutterless thermal camera coupled with image processing that is capable of extracting both location, and range data would meet all of these challenging requirements and provide dramatic advantages over current systems.

Substantial work has been done to develop and demonstrate the methods for extracting range data from monocular RGB images.[4–6] Universally, these methods use convolutional neural networks (CNNs) to identify objects and determine their range. Since camera images are the source data, the locations in the scene are determined by geometric calculations based on the position of the detected objects on the sensor. A variety of demonstrations[7,8] of monocular ranging have been made available online, including two[9,10] using the equipment described here.

5 Implementing Monocular Thermal Ranging™

Practical and cost-effective thermal ranging calls for developing a camera and processor that are optimized for the application: detection and ranging of pedestrians and other vulnerable road users (VRUs) from a moving automobile. The requirements are straightforward as follows.

- Detection of humans (and animals) by their own emitted body heat using thermal imaging in the 8- to 14-μm thermal band with a single low-cost sensor.
- Sufficient resolution to enable detection and ranging curb to curb.
- A camera that does not need to be shuttered periodically for recalibration.
- CNNs trained to extract location and range information from a monocular thermal image.
- Output data formatted for use in automotive AEB systems.
- Optionally, output video for driver use including markers, and alerts identifying detected VRUs with their ranges.

5.1 *Sensor Array*

Currently, no commercial sensor incorporates all of the features needed to supply the operating modes and the data quality demanded by the requirements above, so design of a custom sensor was undertaken. The final design consists of a VOx microbolometer array incorporating features that provide continuous thermal reference signals, allowing continuous shutterless operation, which is imperative for safe application within moving vehicles. Further, the 1280 (H) × 800 (V) resolution of this array is sufficient to support a wide horizontal field of view within a single camera and sets a new benchmark for automotive thermal cameras.

The 1280×800 pixel active array is surrounded on all four sides by dummy pixels to provide the outermost active pixels with an environment that matches that of the interior pixels. Additional pixels are included for calibration processes and for providing reference signals. The input of the sensor array is a set of analog signals from all of the microbolometer elements, which are immediately digitized and fed to an arithmetic unit on a per pixel basis. This unique sensor, with the ability to digitize on a per pixel basis at input and thereby store frame data in digital format versus analog, unlocks a number of benefits not currently available within today's automotive class of VGA-based thermal sensors. Specifically, this sensor simultaneously achieves 3.5 times higher resolution (1 megapixel), higher dynamic range, better sensitivity, a 15 times reduction in power per pixel, support for extended automotive temperature range operation, shutterless and global shutter-based operation, and significantly lower cost per pixel.

Moreover, the array integrates many of the traditional image signal processing (ISP) functions into the array itself, further reducing footprint and cost. The ISP function within current automotive thermal cameras requires a second chip. This results in a multiboard stack that takes

up more space, requires more power, and adds cost. This novel digital implementation enables a single-board thermal camera system for the first time.

5.2 *ROIC Construction*

To minimize size and, therefore, the cost of both the sensor and its associated optics, the high definition (HD) read out integrated circuit (ROIC) die while providing both 3.5 times the number of pixels as a video graphics adapter (VGA) array and individual analog-to-digital conversion under the pixels, is essentially the same size as a typical VGA array with the same pixel pitch as the HD ROIC, as shown in Fig. 1(a). In the predominately analog VGA device, the nonimaging circuitry is adjacent to the active imaging area, whereas in the new HD device as a result of its unique digital architecture, the entire surface area is available for imaging, and the nonimaging circuitry is below the imaging area.

Placing the ADCs under the pixels enables this HD ROIC to leverage the same wafer scale packaging used on low-cost VGA FPAs. A larger HD ROICs would require the FPA wafer to be diced and mount in a package then vacuum sealed, which precludes high-volume manufacturing required for low-cost automotive applications. This distinction is shown in Fig. 1(b); here, the active pixel imaging area of the VGA ROIC is overlaid (red box) on the top of the active pixel imaging area of the HD ROIC.

5.3 *Readout Electronics*

The microbolometer sensor array is patterned onto a custom readout integrated circuit (ROIC), which provides support for array power and scanning and prepares the video for processing by the CNN. This processing includes digitization, analog corrections for gain and offset, and programmable output data formatting, all controlled through an I2C interface. Figure 2 shows the layout of the functional blocks in the application specific integrated circuit (ASIC).

5.3.1 *Digital image frame*

At the top is the digital image frame (IF) where the microbolometer array is bonded. In this section are the driving circuits for the array and the per pixel analog to digital converters. All image data leaving this block are in digital form. To the right of the IF is a block that provides biasing to the array. This is the only part of the ROIC that has analog interconnections with other blocks.

5.3.2 *Digital control and processing blocks*

To the left of the IF is the row timing generator. This drives the readout of the array row by row and accepts commands that define the top and bottom of a region of interest that the user can specify. Similarly, the digital frame multiplexer below the IF receives and formats the pixel data arriving line-by-line from the IF and defines the left and right borders of the ROIC. Coherent timing of all operations is ensured by clocks and controls sent to both the row and frame blocks by the frame controller located near the bottom center.

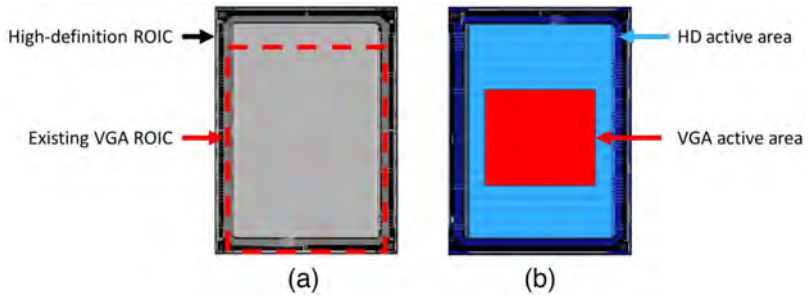

Fig. 1 ROIC size comparison: (a) die area comparison and (b) active area comparison.

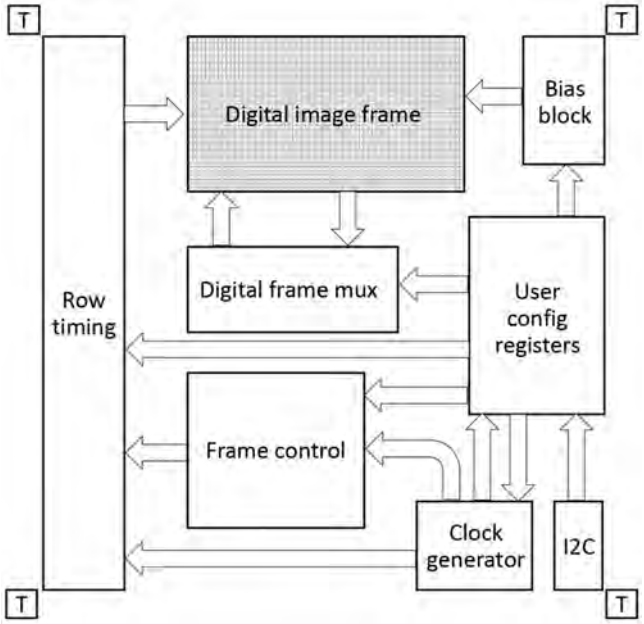

Fig. 2 ROIC block diagram.

To the right of these blocks is the user interface block, which contains an extensive set of addressable registers through which the user can specify the operation of all functions in the ROIC. This is connected to the outside through an I2C interface, shown at the bottom right. To the left of this is the clock input block, which contains a phase-locked loop to assure continuous operation of the sensor even in the presence of noisy external clocks.

Finally, at each corner of the ROIC is a temperature sensor to provide the processor with data necessary to allow compensation for not only changes in the sensor array temperature but also due to temperature variations in the ROIC itself.

5.3.3 *ROIC interfaces*

The ROIC is designed for use in chip-on-board assemblies to facilitate low-cost production. Surrounding the IF is a seal ring to which a lid may be bonded, providing the hermetically sealed vacuum cavity needed for microbolometers to operate properly. This lid supports a window, generally made of silicon, which passes the thermal infrared radiation to be sensed by the microbolometer array.

Electronic connection to the supporting board is provided by a ring of conventional wire bond pads outside the seal ring and along the top and bottom sides of the ROIC die.

5.4 *CNN Inference Platform*

Training of the CNNs is done off-line using a training complex with several NVIDIA A100 GPUs. Auxiliary sensors are used for ground truth reference during the training, but not for real-time processing using the identification and location CNNs. The trained CNNs were used to evaluate using people at known ranges and found to be accurate with a $\pm 10\%$ range error (1 m at 10 m, 6 m at 60 m). This is sufficient for the intended ADAS application since absolute accuracy at long distances does not impact emergency breaking decisions.

After verification, the trained weights are transferred to a runtime inference engine for real-time processing of the thermal video. Runtime CNN object identification and location data are supplied through a Robot Operating System (ROS) Publisher, and visualization images are prepared by an ROS Subscriber currently implemented at the edge on an NVIDIA AGX Orin processor, but the ROS nodes can be readily moved to a centralized processor for other installations.

Figure 3 shows the arrangement of hardware used to capture and process all the images reported in this document. Table 1 lists the equipment in this setup.

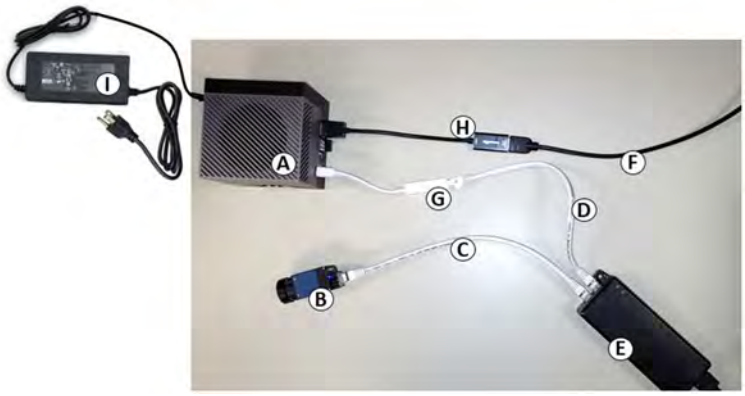

Fig. 3 Inference hardware platform.

Table 1 Inference hardware equipment list.

ID letter	Type	Description
A	Processor	NVIDIA Jetson AGX Orin development kit
B	Camera	Owl AI VGA thermal camera with 14 mm lens
C	Cable	Cat 6 Ethernet cable
D	Cable	Cat 6 Ethernet cable
E	Power	TrendNet TPE-115G11A Power over Ethernet injector
F	Cable	HDMI cable
G	Converter	TP-Link UE300C USB-C to Ethernet converter
H	Converter	Amazon Basics DP-H01 DisplayPort to HDMI active converter
I	Power	NVIDIA AC/DC power module with USB3 output

5.4.1 *Processor*

In the development system, which generated the images shown in this paper, image data from the ROIC are sent to a processor on the supporting board, where all of the data extraction operations are performed.

The configuration of the processor, an inference pipeline, is shown in Fig. 4. An inference pipeline consists of CNN modules separated by fixed processing modules that prepare the data for the next neural network step. For the range detection application, two neural network outputs are needed—one to extract the range data for every pixel in the image and a second to recognize the objects of interest and place them in their proper image locations so they can be tagged with their range.

5.4.2 *Lens distortion correction*

Images recorded from the camera (upper left in Fig. 4) must be corrected to remove the geometric distortion resulting from the use of a wide-angle lens. If this was not done, the objects of interest would have different sizes and curvatures depending on location within image. Although the CNN could be trained to recognize the objects even with such variety, the likelihood of recognizing other objects as falling into the desired class increases. Correcting distortion before the neural processing is one way to normalize the image input so that the recognition process operates more accurately and consistently.

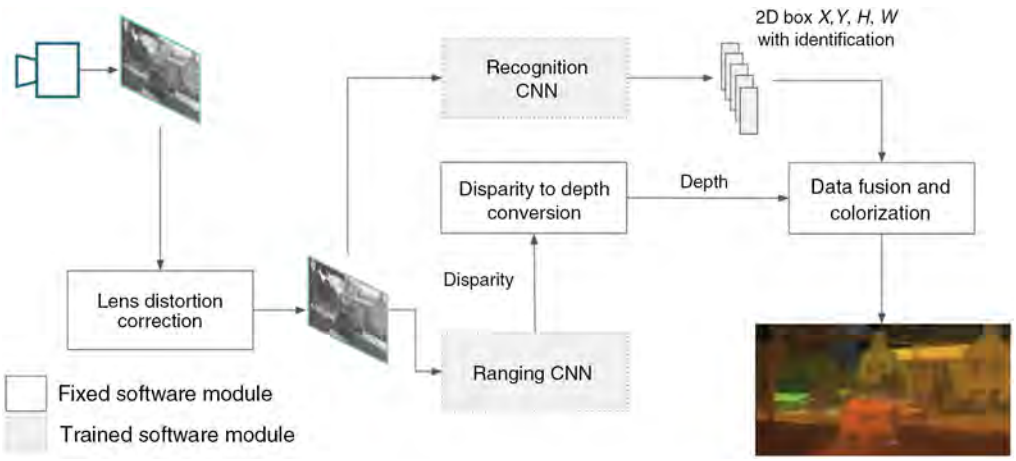

Fig. 4 Inference pipeline.

5.4.3 *Recognition CNN*

The recognition CNN was trained to identify objects of interest, such as pedestrians. This training involves presenting images to the CNN that contain pedestrians (or other objects of interest) and then confirming correct detection and classification when the CNN processes similar images after the training session. The accuracy of CNN processors trained this way improves rapidly at first and then approaches maximum accuracy asymptotically. Training is terminated when the CNN reaches a desired accuracy goal. Appendix B describes the training process developed for determining object location and range.

5.4.4 *Bounding box generator*

Objects identified as pedestrians are passed on to a fixed processor that draws a bounding box around each object using an x, y pixel-based coordinate system and labels it with its type. A confidence value is also generated. If the CNN had been trained to recognize more than one type of object, then each object would be labeled with its recognized type.

5.4.5 *Ranging CNN*

Although the recognition CNN is finding the objects, another CNN is examining every pixel in the image. This CNN, using a proprietary algorithm developed for this system, maps image disparity signatures that can be converted into range plots, also sometimes referred to as point clouds.

5.4.6 *Disparity to depth converter*

From the disparity data, the converter computes a range for each pixel in the image, presenting the result as a z-coordinate point cloud measured in meters, suitable for fusion with identification data.

5.4.7 *Optional visualization processor*

Finally, the fusion processor merges the depth data in meters with the x, y pixel-based object thermal data. This process further converts all of the recognized objects into a 3D range map representation where all coordinates are measured in meters (x, y, z) and labels each of them with the range value associated with its location in the image. Optionally, each range is assigned a color and the final display is prepared, showing the scene with color-coding for depth and bounding boxes for the objects of interest, including type labels and ranges. Alternatively, the data may be represented via a 3D bird's eye representation, where the objects of interest are encapsulated within 3D bounding boxes.

5.4.8 *Composite display*

Figure 5 shows the display of the colorized 3D data with all of its computed elements. The pedestrians are delineated by 2D bounding boxes, which are labeled here with the range. The color indicates the range map but is muted outside the bounding boxes for contrast. Note, however, that the left bounding box contains a small green area that indicates the longer-range background behind the pedestrian.

6 Operation Example

The images in Sec. 5 show the results of controlled testing. Figures 6 and 7 illustrate images from a fairly characteristic setting. Both of these figures show still frames from videos, which are part of this paper. Figure 6 shows three steps in the identification and ranging process. In the lower right is the original LWIR thermal image. This is the monocular image data that is provided to the inference processor. In the center is the data extracted from the image presented as a bird's eye view of the data (traditional point cloud representation) showing a blank field for all areas not trained for identification by the CNN. Overlaid are the identified objects of interest (pedestrians and automobiles) surrounded by 3D bounding boxes. At this stage, all coordinates (x, y, z) have been converted from the pixel domain to a 3D domain in meters. Inside the bounding boxes are the 3D images showing the extracted data points that constitute each object. The upper left is the driver view representation of the output image showing the recognized objects with range labels.

Figure 7 shows an expanded view of the driver view output image. In this view, the pedestrian thermal images can be seen with their associated 2D bounding boxes and range labels. Notice on the left, the distinction in range between two pedestrians walking together. In this example, the software labels pedestrian ranges up to 50 m with specific values, whereas pedestrians outside that range are labeled as pedestrians (ped) but do not have range labels. Note that the automobile headlights in the distance are not found to be pedestrians. Video 2 provides a dynamic presentation of this scene.

7 Applying Thermal Ranging

In a real automotive implementation of thermal ranging, only the camera and software would be supplied to the automobile manufacturer or to one of its qualified suppliers (known as tier 1

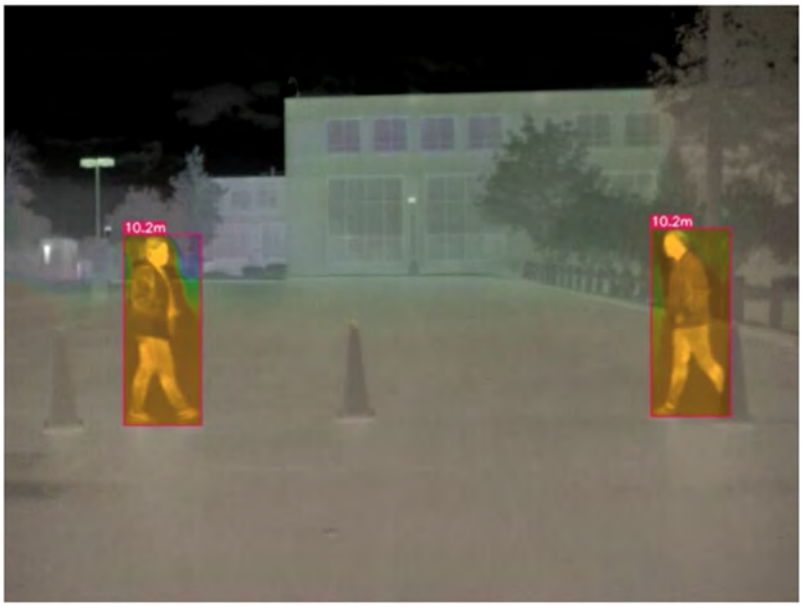

Fig. 5 Pedestrians with labeled bounding boxes.

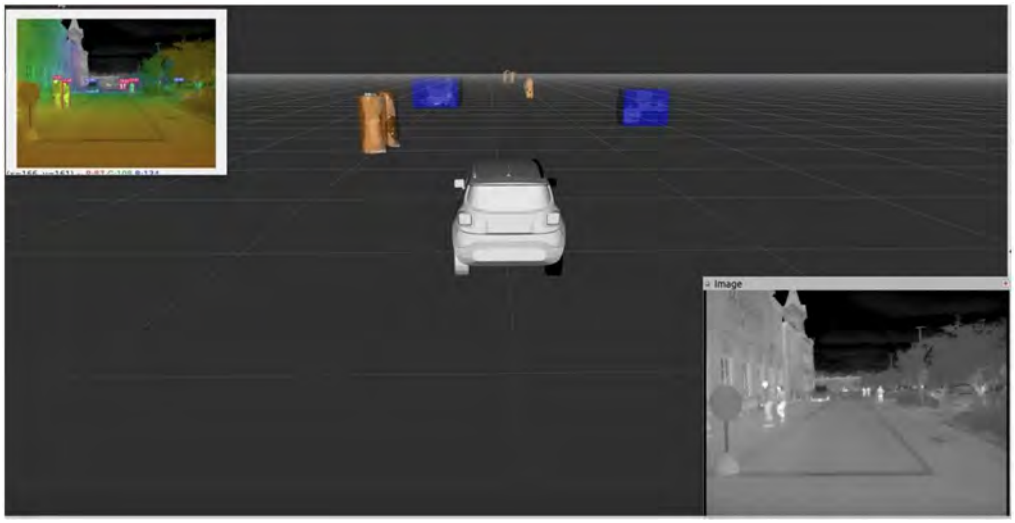

Fig. 6 Stages in identification and labeling (see Video 1 for additional samples) (Video 1, mp4, 12 MB [URL: https://doi.org/10.1117/1.OE.62.3.031212.s1]).

Fig. 7 An enlarged view of labeled pedestrians (see Video 2 for a dynamic example) (Video 2, mp4, 3.93 MB [URL: https://doi.org/10.1117/1.OE.62.3.031212.s2]).

suppliers). The computing hardware in which the inference engine resides is part of a larger set of computing hardware that provides all of the sensor processing and actuation decisions needed to operate the vehicle safely. Figure 8 shows a representation of the object fusion processing architecture to go from a 2D thermal video stream to a complete 3D world space-based dataset. Full compliance with such an arrangement is essential if a new technology is to be integrated into automobile perception stack designs.

On the left is the thermal image from the sensor and associated imaging camera. The digital video information from the camera is in the form of a 2D pixel matrix, where each pixel intensity value is represented by a digital number. The image data are transmitted in frames accompanied by timing information and certain labels necessary to identify each frame and every pixel location in the frame. Everything to the right of the camera is part of a common electronics set in the vehicle.

Residing in the computer is a processor of either the graphics processing unit (GPU) or the tensor processing unit type. These are both matrix processors capable of running the CNN algorithms in real time. The CNN algorithms and drivers to operate the camera are loaded into the matrix processor along with the current set of CNN weights generated by the training sessions conducted by the camera manufacturer. No training takes place using images acquired in the vehicle itself during run time operation.

As in the development platform, the image data are received by the CNN blocks, which extract position and range data. Both of these remain in pixel space: each pixel is assigned a range and the objects are provided with 2D bounding boxes located in the field of view by pixel counts from a reference on the sensor. At this point in the processing, the data are ready to be mapped to world space.

Using object fusion routines customized for the thermal data the processor determines the range value to assign to the label on each of the bounding boxes as well as the dimensions of the 3D bounding box. Then using know geometric calibration factors, it converts the pixel numbers into dimensions in meters. The bounding box then resides in World Space.

Finally, the real-world $x - y - z$ dimensions of the object fusion dataset are formatted and transmitted to the action part of the computer system, using the ROS schema common to many of the vehicle functions. Once in real-world space, the object and its 3D location can be easily fused with objects generated by other modalities, such as RGB images combined with radar or LIDAR ranging. As Fig. 8 shows, both the coordinate data (represented in the lower image) and the picture data (similar to the image in Fig. 7) are produced for appropriate use.

8 Real-World Testing

Although a production vehicle would include a common or central computer to process the thermal images, a stand-alone demonstration of this technology requires a self-contained system that includes the camera, compute power, and software that generates output using the ROS schema. An ADAS development platform (Fig. 9) has been built to accomplish this.

The platform includes the thermal camera with suitable optics, a custom processor box running Linux that includes a supervisory controller, a GPU to run the CNNs, an object fusion processor, and a variety of external interfaces needed to support connection to automotive electronics. The output is $x - y - z$ data defining real-world coordinates of bounding boxes, label information for those bounding boxes, and a displayable image of the type described above. Provisions are included to allow simple updating of the drivers, algorithms, and CNN weights. It should be noted that the drivers and all other firmware were developed to be readily adaptable for use on current and planned automotive computer systems.

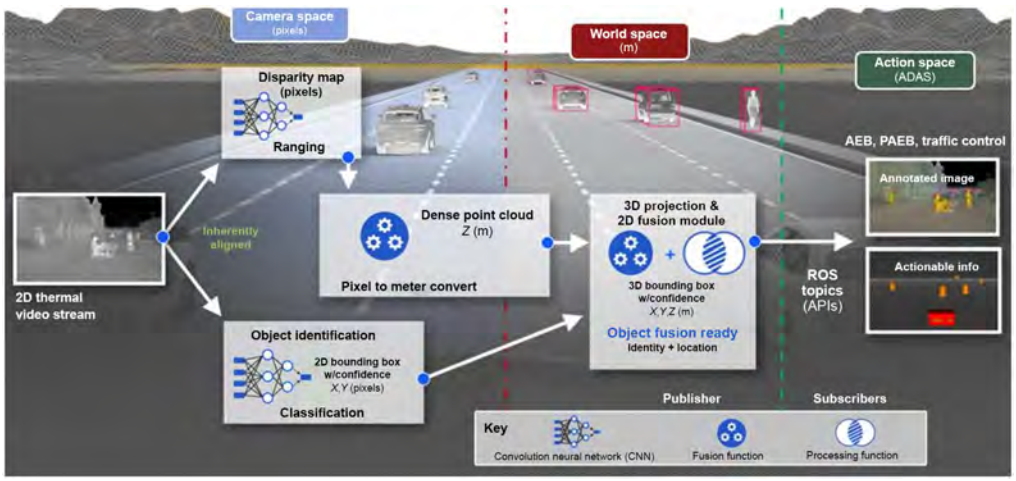

Fig. 8 Data flow and usage.

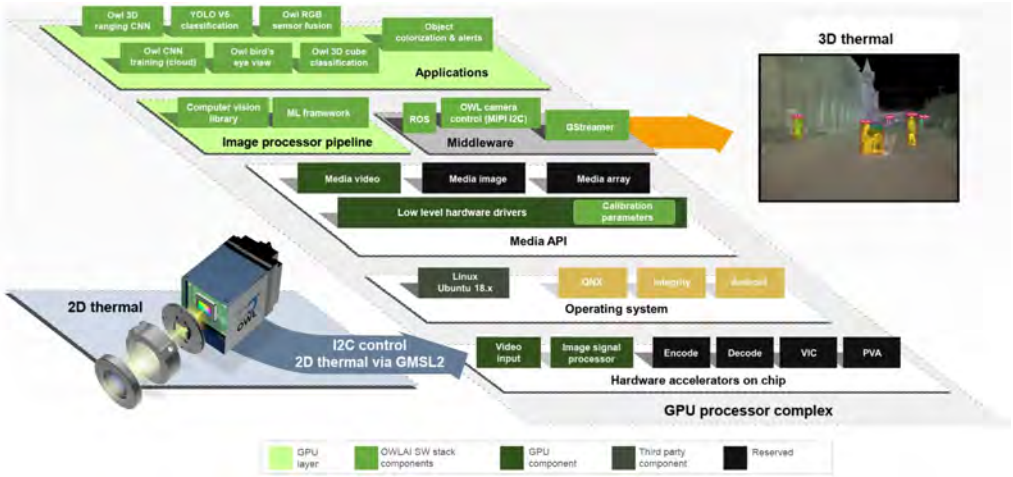

Fig. 9 The thermal ranging development platform.

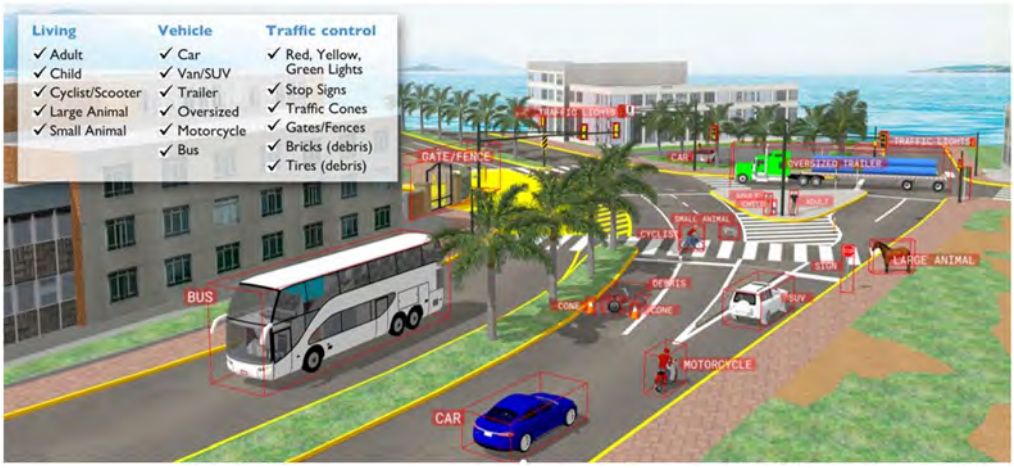

Fig. 10 A traffic scenario.

9 Requirements for Full Deployment

Comprehensive identification of all of the object types in Fig. 10 is essential for the safe deployment of ADAS or fully autonomous vehicles. Although many of these objects can be classified based on their visible images only, thermal data can provide the additional information needed to sort activated objects (human or machine) from the static background. For instance, is a truck at the curb preparing to pull out? Thermal imaging can help determine that the engine is running.

Whether ADAS or autonomous vehicles, continuous operation is essential. The system described here is always on, never blacked out while recalibration is underway. Not only does this avoid interruptions in the thermal data, it also avoids delays resulting from re-establishing frame sync with other sensors during fusion.

10 Conclusion

Initial development and testing of a monocular thermal ranging system has demonstrated that this camera configuration can provide essential location and range information day and night. With a custom thermal detector array and a supporting ROIC that eliminate the need for calibration shuttering, this thermal ranging system is ready for consideration by teams designing for

both driver assistance and fully autonomous applications. A development platform is now available for use by these teams.

11 Appendix A: General Neural Network Training

Neural networks operate by configuring themselves to produce a positive response when they detect a match between a new set of incoming data and a series of datasets that they have received in training. This is similar to the way radiologists are trained to read images. As one radiologist put it, "show them 50,000 images and tell them what they mean and, after seeing these, they will be able to tell you what any new images mean." That is, if the new images were similar to some of those in the training set and the training set was broad enough to cover everything, a radiologist might need to evaluate.

The training curves for a typical CNN show what this means. In training, the CNN is presented with datasets which are identified according to what the CNN should recognize. Initially, the CNN knows nothing, so its accuracy is very low and its failure rate (loss) is very high. With additional training, the CNN improves in accuracy and the loss drops. Typically, the training dataset is presented to the CNN many times in sessions known as epochs to reinforce the selectivity of the CNN.

However, it is possible that the repeated training data presentations can cause the CNN to ultimately recognize only those datasets that are very close to the training set. This does not show up in the training, but when datasets are presented that were not used for training to validate the function of the CNN, too many of them may be rejected. This phenomenon, called overfitting, can be seen in Fig. 11.[11] Those demonstrates that attempts to make the CNN perform perfectly almost always lead to additional failures. Because the onset of overfitting depends on the configuration of the CNN and on the training datasets used, care must be taken to stop training when the CNN operation is near its optimum.

12 Appendix B: The Specific Training Process

The processor uses neural networks of the type known as CNNs, which, in general, are applied to tasks requiring the extraction of data from images. In operation, the CNN accepts images and outputs a map of features, often the location and orientation of edges indicating transitions in gray level or color. This edge map can then be passed to a second CNN to recognize, in the arrangement of edges, specified types of objects.

As shown in Fig. 2 in Sec. 5.3, the inference pipeline includes two CNN sections, one for object recognition and the other for generation of a range map. The object recognition CNN section is trained by showing it images of the types of objects to be recognized. This is a relatively common RGB image application of the CNN and uses well-known training methods.

Using a CNN to generate a range map from thermal images, however, is not common. Thermal images have very different spatial frequency contents and required special, proprietary training, and calibration techniques to be developed. The goal of building a range map in the system is to label each of the objects identified as humans by the recognition CNN section with a range that matches the range measured by conventional means, data known as ground truth.

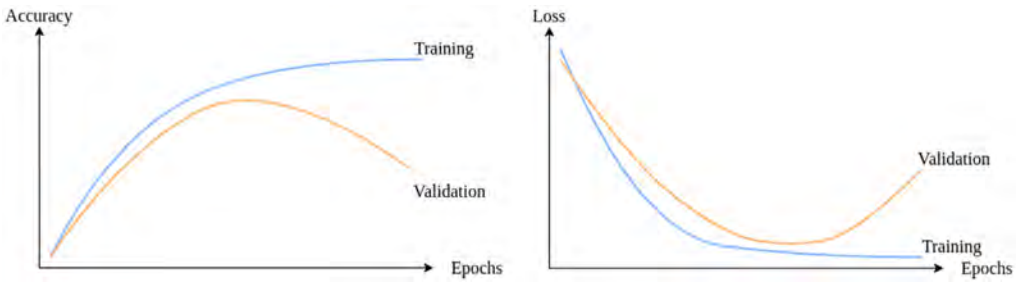

Fig. 11 CNN training curves.

Because of the differences in resolution and wavelength spatial frequency response with longer wavelength radiation, it is difficult to use a ground truth data from RGB cameras as the ground truth set used to supervise training of a thermal CNN. The preliminary results presented here demonstrate the potential of cross modality training.

The first step in this process is to train the CNN to take in images of a specific type that contain disparity information at each pixel that is indicative of range. In the current version of the CNN, the disparity data are proportional to the inverse of depth (1/distance). The output from the CNN is in arbitrary units and is not linear, with an offset from zero.

Acknowledgments

This work was partially funded by an SBIR contract from the National Science Foundation (Contract No. 2014933) and a Prototype Agreement (No. W50RAJ-21-9-0026) from the United States Army Rapid Capabilities and Critical Technologies Office.

References

1. R. Schram et al., "Euro NCAP's first step to assess autonomous emergency braking (AEB) for vulnerable road users," ESV 2019, 15-0277 (2019).
2. K. Macek, *Pedestrian Fatalities by State: Preliminary 2021 Data*, Governors Highway Safety Association (2022).
3. J. B. Cicchino, "Effects of automatic emergency braking systems on pedestrian crash risk," *Accid. Anal. Prevent.* **172**, 106686 (2022).
4. X. Zhang et al., "Object fusion tracking based on visible and infrared images: a comprehensive review," *Inf. Fusion*, **63**, 166–187 (2020).
5. J. Fang et al., "Self-supervised camera self-calibration from video," in *ICRA* (2022).
6. A. Gordon et al., "Depth from videos in the wild: unsupervised monocular depth learning from unknown cameras," in *ICCV* (2019).
7. Y. Aksoy, "Boosting monocular depth estimation to high resolution," CVPR, 2021, https://youtu.be/lDeI17pHlqo (accessed 1 Aug 2022).
8. H. Zhang et al., "Monocular 3D localization of vehicles in road scenes," ICCV Workshop, 2019, https://youtu.be/SY5HGWoViH4 (accessed 1 Aug 2022).
9. Owl Autonomous Imaging, Inc., "Owl lAI thermal range with classification demo," https://youtu.be/BMGLgnxNI6M (accessed 1 Aug 2022).
10. Owl Autonomous Imaging, Inc., "CES 2022 RGB vs. thermal side by side video with voice over," https://youtu.be/TmfzYcGRH_Y (accessed 1 Aug 2022).
11. "Epoch in neural networks," Baeldung, https://www.baeldung.com/cs/epoch-neural-networks (accessed 27 June 2022).

Eugene Petilli received his BEEE degree in electrical engineering and computer science and his MSEE degree in electrical engineering/microelectronics. He is a co-founder and the CTO of Owl Autonomous Imaging. He has more than 35 years of experience in imaging systems and mixed-signal ASIC design. Under his leadership, the Mixed-Signal Center of Excellence at Intrinsix Corp. (a CEVA company) executed more than 25 ASIC projects including imaging focal plane arrays. He holds 24 patents in imaging and data converters.