Intelligent Audio Analytics

Technical Whitepaper

Table of contents

1 Introduction

Traditionally, security and detection systems are based on visual monitoring of assets and people. However, video sensors alone are not always the most ideal solution to provide the needed situational awareness to accurately determine the threat and secure the perimeter Environmental conditions such as darkness, weather, or physical obstacles may even disable vision capabilities.

By including sound, an additional layer of awareness is added to detect certain incidents faster and more reliable. For example, a gunshot or a smoke alarm is better heard than seen. That's why Bosch Building Technologies combines both video and audio sensors with Artificial Intelligence (AI) to enhance awareness and facilitate a quick and appropriate response to an alarming event. With the neural network based Intelligent Audio Analytics the camera helps to recognize incidents of interest in the area around the camera, for both indoor and outdoor applications. Bosch' Intelligent Audio Analytics is based on Bosch SoundSee technology developed to improve operations on board of the International Space Station (ISS). SoundSee interprets sounds using audio AI trained with high quality data to detect unusual sound events.

The FLEXIDOME panoramic 5100i IR is the first Bosch camera to support Intelligent Audio Analytics to intelligently recognize a variety of typical sounds based on a trained dataset (machine learning). With its integrated microphone array, this camera not only recognizes sounds, but also can indicate the direction from where the sound originated. There is no need to record audio, as it is possible to examine sound signatures real-time. The first release of Intelligent Audio Analytics is part of Firmware 8.80

2 Technology

Intelligent Audio Analytics by Bosch is a powerful AI-driven audio signal processing algorithm to detect and identify target sounds from the ambient sound. Using Artificial Intelligence, it can differentiate unusual sounds from the background or surrounding noise.

2.1 SoundSee

The analytics algorithm is based on Bosch owned and developed SoundSee technology. This is a deep sound analytics capability that uses machine learning to analyse information contained in emitted noises. SoundSee was developed in a research partnership between Bosch and Astrobotic Technology Inc. started in 2019 to improve the operations on board of the International Space Station (ISS). Its goal was to use auditory analytics to determine whether machines or their components in the ISS needed repair or replacement. The algorithm uses machine learning to analyze subtle acoustic clues in a machine noise and determines whether a machine, or even a single component of a machine, needs to be repaired or replaced. Now, already launched to the ISS for research experiments, SoundSee is scalable for a broad range of commercial uses here on Earth such as predictive maintenance, early warning systems, building technologies and data-driven healthcare.

2.2 Recognizing and identifying sounds

Intelligent Audio Analytics allows the camera to recognize and identify the unique audio signatures of sounds like gunshots or a smoke alarm sound. High level feature extraction is done by Convolutional Neural Networks (CNN) that mimics the human auditory system for sound perception. By creating large datasets to drive the deep neural networks, Bosch is constantly building and improving the algorithms to differentiate similar audio signatures like a car door slam or a truck backfiring, from an actual gunshot. Each one of these events can make a similar sound to human hearing, and yet the response, especially if the event is known, would be very different. By enabling a so called sound detector in the camera, the camera will detect and classify the sound, while ignoring false positives. It will alert the operator instantaneously when the sound matches the audio signature of the selected sound detector, and can also be configured in combination with non-Bosch management systems, such as Milestone Systems.

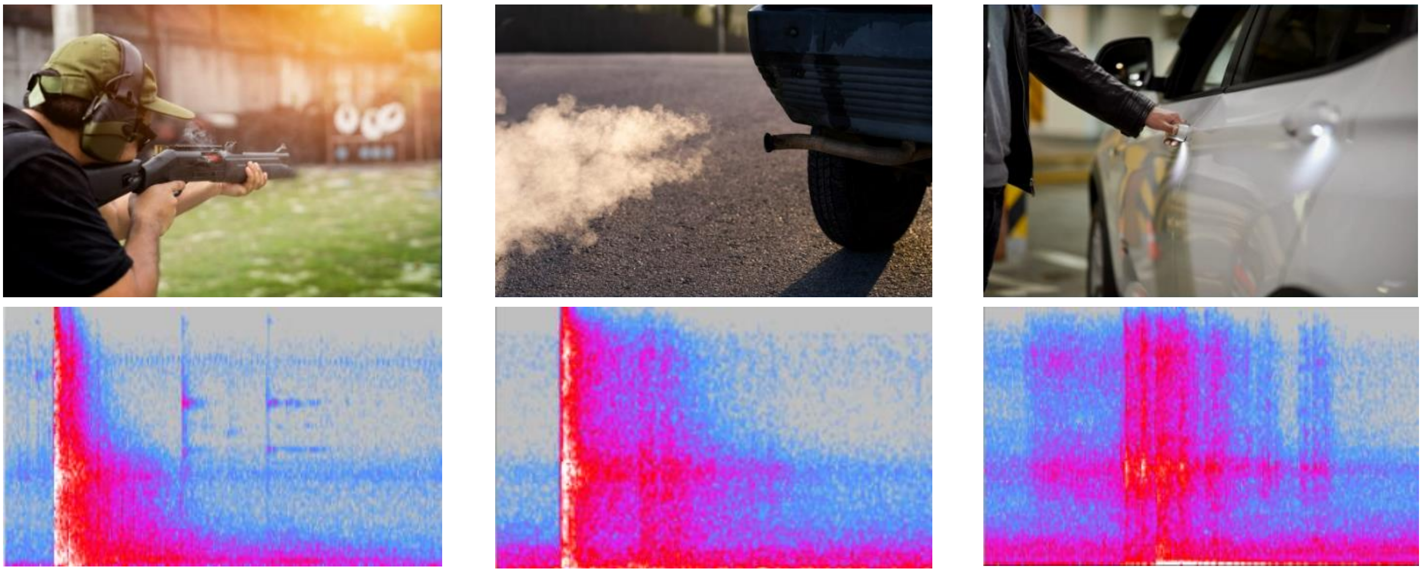


Figure 1: Similar sounds with different audio signatures

The spectrograms above are a visual representation of the audio signatures. The X-axis is time, and the Y-axis is the audio frequency. The intensity is represented by a color. The hotter the color temperature the louder the sound is at that point in frequency and time. The above signatures are clearly different, but most gun types have very similar signatures, yet they are different given the caliber and other specifications of the gun. This information all has to built into the data model.

2.3 Privacy protected

Intelligent Audio Analytics is fully edge-based, utilizing an AI data model and algorithm running directly on the camera. By relying solely on audio signatures, privacy is protected, as no actual audio needs to be recorded or has to leave the camera. Instead, only metadata is streamed together with the video stream. Also, no client side software or external connectivity to the cloud is necessary. For installations with elevated privacy concerns or restrictions, such as public spaces or schools, the audio output can be permanently blocked through a special license. With the license activated, the system operator won't be able to access the actual audio stream to ensure privacy is protected. Even with the audio being blocked, the camera still maintains the capability to run Intelligent Audio Analytics, sending metadata for detected events as needed.

Activation key to disable audio function while maintain analytics metadata for CPP14 cameras is
22-01.63.01-AD386378-E874DC69-F29E8915-8FA63F26-14DE32D7

The permanent block is irreversible by customers and can only be lifted by releasing this license via the Bosch service organization

2.4 Building a reliable audio analytics model

Depending on the application, the deep learning model may consist of more than 30 CNN layers depth to convert input audio to high level features that eventually would be converted to a confidence probability of the target event in the output. A CNN model is mimicking the human brain. The CNN layers are trained through extensive training procedures using trained data. For gunshot detection, the training data includes several thousands of gunshots from various caliber guns, all recorded by Bosch Security Systems. Other training data consists of commercial and public audio recordings for various events including wind, warehouse background noise, crowded court, high noise sorting facility, musical instruments, nature sounds, etc. To make the model more reliable and less prune to false positives, datasets with audio events that are intentionally impulsive and loud (similar to the pattern of a gunshot) such as heavy box dropping, metal materials impact, balloon pop, etc, were also included in the training process.

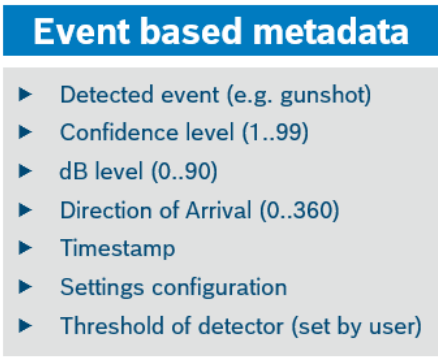
The model listens to the surrounding environment and classifies the input sound from the most recent two seconds as either the targeted event (e.g. gunshot) or “other” event, for the case that the sound was not recognized as target event. And again, no audio will be recorded.

2.5 Directional information

As the first Bosch camera to bring Intelligent Audio Analytics, the FLEXIDOME panoramic 5100i (IR) is also equipped with an integrated microphone array. With three digital MEMS (Micro-electromechanical systems or digital) audio sensors it provides directional information on detected sounds. The circular array in this camera is optimized for 360° audio pick up and capture spatial information from the sound field. Audio signal processing is used to identify the signal from each microphone. With this innovative technology the camera can obtain more information from the acoustic environment. Using beamforming, the direction of the source of a sound can be determined from the 3 audio sensors. They work together simultaneously to analyze the received sound wave and the different time delay of arrival at each microphone. This relative delay, together with the known distances between the microphones, is used to determine the direction of sound. With this information the camera can directly alert the security guard to the direction the sound originated from. These capabilities help security personnel to react quickly and appropriately and take direct and targeted action.

2.6 Metadata

Without the need to record audio, Intelligent Audio Analytics generates metadata that is seamlessly integrated with Bosch’ Intelligent Video Analytics metadata stream. The metadata contains information on the detected event, such as confidence level and direction of arrival.
This information is passed to the client or server where alarm rules can be defined to notify the operator. It provides them sufficient information to take actions based on both video and audio footage or analytics. So, in case of a detected sound event, the camera highlights it as an alarm in the browser or client to enable instantaneous reaction. Also, if supported by the Video Management System (VMS), the recorded metadata can be used for a full forensic search (i.e. search events in retrospect) where the rules can be changed even after the fact. Bosch VMS offers this flexibility in event search. New tasks can be defined to search through both video and audio metadata to evaluate incidents or find anomalous sounds more easily. This allows the user to search through hours of footage in a second to find a specific event like a loud scream or a gunshot.


Event based metadata

- Detected event (e.g. gunshot)
- Confidence level (1..99)
- dB level (0..90)
- Direction of Arrival (0..360)
- Timestamp
- Settings configuration
- Threshold of detector (set by user)

Note: the Bosch forensic search plugin might be required for this, depending the used VMS.

Bosch metadata is ONVIF Profile M conformant, thus enabling the use of metadata and event handling into other systems. With FW 9.0, also the audio metadata will be ONVIF Profile M conformant. Profile M standardizes how metadata and events are communicated between analytics-capable services and devices like cameras, and clients like video management software or server- or cloud-based services. The metadata can be used to trigger automatic responses when a particular sound is detected and recognized by the camera. For example an evacuation alarm message can be triggered when the sound of a gunshot has been detected. Also, it can facilitate location mapping using the sound and geolocation metadata of the camera.

3 Usage and Benefits

Since Intelligent Audio Analytics is in constant development, new Sound Detectors can be added with every camera firmware release. In the first release pack of Intelligent Audio Analytics, two Sound Detectors are supported: Gunshot detection and smoke (T3) / carbon monoxide (T4) alarms. Based on these so-called sound detectors and its metadata, the user can set up an alarm to be able to take immediate action.

Sound Detector	Description

Gunshot	Discharge detection of various types of firearms, guns, rifles and automatic weapons in both indoor and outdoor environments
T3 / T4 Alarm	Detection of alarm sounds of nearby detector alarms. Two types of detector alarms are supported: smoke (T3), and carbon monoxide (T4). A T3 signal is recognized by three intermittent beeps followed by a period of silence, and a T4 signal emits four intermittent beeps followed by a period of silence.

3.1 Expected performance

The environment where the cameras are installed highly influences the performance of Intelligent Audio Analytics. Outdoor environments are more prone to nature unwanted sounds such as wind, rain, lightening, or human made noises such as traffic, impact of objects, car backfire, etc. Indoor environments are challenging in terms of the reverberation of loud sounds inside the room, which depends on the surrounding materials (wall, ceiling, floor), and size of the room. Sounds with signatures similar to a gunshot such as balloon pop, heavy box dropping, etc, can accumulate energy from the reverberation and trigger false positives. Accordingly, the indoor Gunshot Detection system, is trained in a slightly different manner to account for the challenges of the indoor environments.

But, as explained earlier, the Sound Detectors are trained for various environments and background noises (like crowded school environments with lots of talking and playing kids) to improve overall performance. On top of this, the user can influence the performance by changing the threshold. A lower threshold can cause more false positive detections, while a higher value will trigger less false positives but can cause more missed true positives (correct detections). The default value provides the best balanced performance.

The estimated detection distance is different for every Sound Detector and is heavily influenced by the ambient noise level. The sound should be louder than the background noise. This is captured by the Signal-to-noise ratio (SNR) and expressed in dB. SNR is a measure to compare the level of a certain sound to the level of ambient noise. A higher SNR improves the ability to detect the desired sound event from the surrounding noise and classify it. For reliable performance, generally the SNR level of a detector alarm sound or a gunshot is around 30dB or higher in a noisy environment. Below table provides the approximate detection distance in two different environments.

	Normal environment	Noisy or obstructed environment
Gunshot Detector (indoor)	24m / 75ft	16m / 50ft
Gunshot Detector (outdoor)	31m / 100ft	24m / 75ft
T3 / T4 Alarm Detector (domestic application)	11m / 35ft	9m / 30ft
T3 / T4 Alarm Detector (commercial application)	14m / 45ft	12m / 40ft

In environments with loud background noise, the detection distance may reduce due to a lower SNR. Also, for the Gunshot Detector, the detection distance may vary per gun and caliber type. If the user needs the camera to be more sensitive for smaller calibers or longer distances, then the threshold for detection can be decreased in the camera GUI. The camera is likely to capture more calibers and potentially cover a longer distance, however, with a reduced threshold the camera is also likely to trigger on more false positives.

The next sections of this white paper provide performance statistics for a few real use cases tested with several FLEXIDOME panoramic 5100i cameras. Under certain conditions (like windy weather or loud background noises), the detection range and performance are affected.

3.2 Test case performance statistics

Unfortunately, gun violence and school shootings are a reality at American schools and communities. Detecting gunshots in live systems, helps local security staff and police to take instantaneous action. Thorough data collection and testing has been done on multiple sites for creating the Intelligent Audio Analytics Gunshot Detector. The next sections contain results of a series of test cases conducted for both indoor and outdoor applications, encompassing evaluations of both true and false positives. Tests are done at the default detector threshold value of 50%.

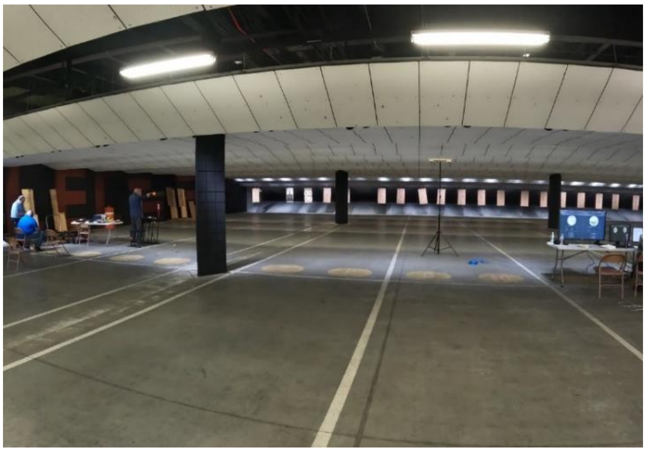
3.2.1 Test case 1 – Elementary school building (indoor)

The first case shows the results from a police training for active shooters in an elementary school in the USA (*n=117*). The test was done with two types of firearms: a 9mm pistol and a .223 rifle. The distance between the camera and the gunshots was between 20 and 100ft (6-30m).
Note: During this test no false positive testing was done.

Location	Handgun (9mm)	Rifle (.223)
Café	100%	100%
Gym	100%	100%
Hallways	100%	100%
Library	100%	100%

3.2.2 Test case 2 – Law enforcement training complex (indoor)

The second indoor test case shows the results at a law enforcement training complex in a community college (*n=1436*). There were four FLEXIDOME panoramic 5100i cameras used for gunshot detection.
Note: During this test no false positive testing was done.



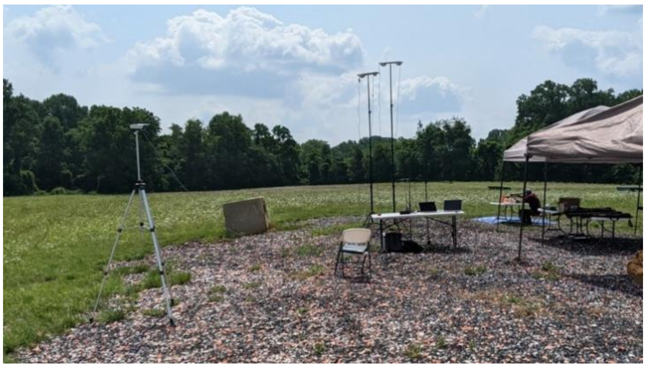
Distance	Handgun	Shotgun	Rifle	Total
25ft / 8m	100%	100%	100%	100%
50ft / 16m	100%	100%	100%	100%
100ft / 31m	83,8%	90%	98,7%	90,8%
150ft / 45m	65,6%	87,5%	100%	83,3%

Gun types per category

- *Handgun (pistols) category includes 9mm, 22lr, 25, 40, 44mag, 45, 357 and 380*
- *Shotgun category includes 12gauge and 20gauge*
- *Rifle (Bolt / Semi-Automatic) category includes 6.5, 223, 242, 243, 30.06, 300blk, 308, 5.45 and 7.62*

3.2.3 Test case 3 – Open field (outdoor)

The third test case shows the results of an outdoor shooting range test conducted in the USA (*n=775*). This test covered 17 different types of firearms at multiple distances between 25 and 150 ft (8-45m) away from the camera.
The tests at 25-50ft (8-16m) and 125-150ft (38-45m) were done in normal weather conditions, and the test at 75-100ft (23-31m) was done under very windy conditions (affecting performance)



Distance	Handgun	Shotgun	Rifle	Total
25-50ft	87,5%	100%	100%	94,3%
75-100ft	89,5%	76,4%	97,9%	92,3%
125-150ft	85,0%	69,5%	97,7%	89,6%

3.2.4 Test case 4 – False positive testing

Key to the success of the Intelligent Audio Analytics Gunshot Detector is to eliminate false alarms. When the system incorrectly identifies a harmless sound as a potential threat, it can lead to unnecessary and costly investigations, wasting resources and causing public alarm. Moreover, repeated false positives may erode trust in the surveillance system, undermining its effectiveness and leading to potential privacy concerns.

Below two tables show the results of deliberate worst case scenarios trying to trigger false positives and deceive the system. Under normal operational conditions, the chances of triggering false positives are significantly lower compared to these test scenarios.
Even in this aggressive test, false positives only occurred within 5ft (1.5m) away from the camera

Forced false positive test on an outdoor parking lot with buildings nearby (*n=496*)

Distance	Pallet drop	Metal shelving	Wood metal hit	Dumpster slam
0ft / 0m	90%	29,1%	16,7%	0%
2.5ft / 0.8m	75%	4,2%	6,7%	0%
5ft / 1.5m	50%	0%	0%	0%
10ft / 3m	0%	0%	0%	0%

Forced false positive test at an indoor mezzanine (*n=472*)

Distance	Metal drop	Pallet drop	Planks hit	Wood metal hit
0ft / 0m	0%	15%	0%	0%
2.5ft / 0.8m	0%	0%	0%	0%
5ft / 1.5m	0%	0%	0%	0%
10ft / 3m	0%	0%	0%	0%

4 Cameras and licenses

4.1 Cameras

Intelligent Audio Analytics is now available on the FLEXIDOME panoramic 5100i (IR). These cameras are also equipped with a microphone array to provide directional information.

In addition to the panoramic cameras mentioned above, there will be future Intelligent Audio Analytics support for other Bosch cameras:

	Sound Detectors	Directional information
FLEXIDOME panoramic 5100i (IR)	FW8.80	FW8.80
FLEXIDOME corner 7100i IR	FW8.81	FW8.81
FLEXIDOME 5100i IR	Future support	-
FLEXIDOME multi 7000i (IR)	Future support	-

4.2 Licences

Intelligent Audio Analytics and its Sound Detectors are a licensed feature that can be purchased per camera. Licenses are available to provide permanent access or for a fixed period of time. Each license can only be activated once on a single camera. All licenses are administered in Bosch Remote Portal. The license process is described in a separate white paper available on the Bosch website.
The T3 / T4 sound detectors are included with the gunshot detector license in FW 8.80 and free of charge with FW 9.0

Material	CTN	Material Description	EAN	UPC	FW
F.01U.412.673	MVC-IAA-GUN	License gunshot detector, perpetual	4060039173782	800549381161	8.80

Bosch Security Systems B.V.
Torenallee 49
5617 BA Eindhoven
Netherlands
www.boschsecurity.com