# How to benchmark Video Analytics

Evaluating video analytics performance

# Table of contents

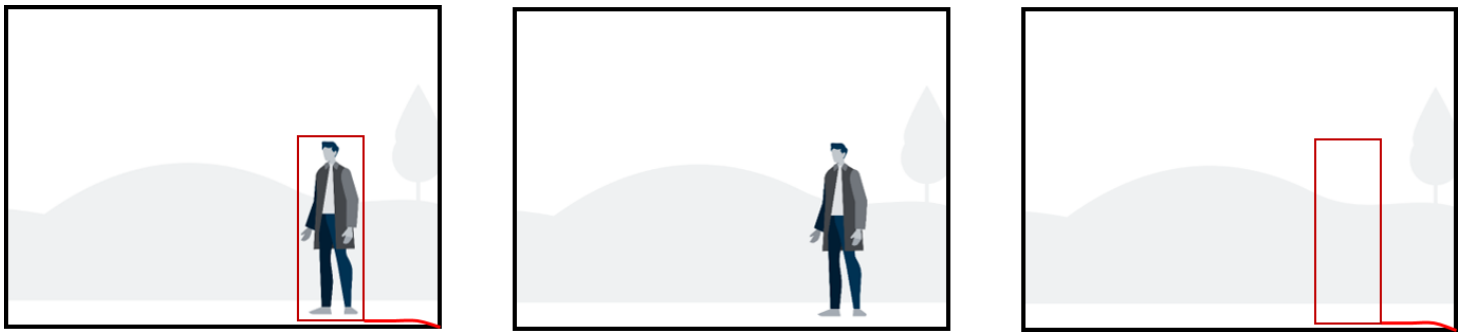
# 1 How to measure video analytics performance?

In this section, the main criteria for video analytics evaluation are presented.

## 1.1 Robustness

Robustness can be determined by counting the following three cases:

- True positive: Object / alarm detected correctly
- False positive: Object / alarm detected though there was none
- False negative: Missed object / alarm



**True, missed and false detection / alarm**
On the left, the object is detected properly. In the middle, while there is an object, the video analytics has missed it. On the right, the video analytics falsely detects an object which is not there.

Both false alarms as well as missed alarms have to be considered in the evaluation of robustness.

In case of intrusion detection, false alerts are very time consuming and annoying and should therefore be minimized as much as possible. If too many false alerts occur, then operators have been known to shut down the video analytics system completely, as they were otherwise no longer able to fulfil their monitoring tasks. Any missed alarms, on the other hand, mean the video analytics did not fulfil their task at all and intruders could enter the premises unhindered. The ratio of true alarms to false alarms is typically very unbalanced. While a single intruder in three month is already much, video analytics can easily generate a multitude of alarms per day.

There is usually a trade-off between the sensitivity of a video analytics algorithm ensuring the detection of all objects / alarms and its false alarm robustness, as a higher sensitivity often means more false alarms, and a higher false alarm robustness often results in less sensitivity. For example, a video analytics that provides large detection distances needs to be more sensitive to be able to detect objects with few pixel only, and thus has more potential to detect false objects than a video analytics that has a reduced detection range and only detects objects covered by many pixels to start with. Exchanging a focus on sensitivity or robustness for the other might make a solution workable for a specific task, but it will not result in a better performance per se. A real progress can only be achieved if both sensitivity and false alarm robustness can be kept and improved. Note that some video analytics focus strongly on reducing false alerts as much as possible, while others focus on ensuring that every intruder will be detected, or on finding a good trade-off between sensitivity and false alarm robustness.

- The sensitivity of a video analytics algorithm is given by the ratio of objects / alarms that have been detected correctly divided by all objects / alarms that should be detected: $(\#\ true\ positives)/(\#\ true\ positives + \#\ false\ negatives)$
- The precision measures false alarms as the ratio of the true detections divided by all detections that have occurred: $(\#\ true\ positives)/(\#\ true\ positives + \#\ false\ positives)$
- An often-used measure of robustness for intrusion detection is the false alarm rate, given by the amount of false alarms over time, e.g. the number of false alarms per hour / day / month.

## 1.2 Detection distance

Detection distance measures the area in which an object / alarm can be reliably detected.

Detection distance typically depends on

- Resolution. Thus it also depends on the field of view of the camera. The field of view on the other hand is given by the camera sensor size, the camera lens focal length and possible lens distortions, and the relation of the camera to the surveyed scene. When comparing different video analytics, the same or at least similar fields of view should be used for all of them.
- Target object size. A larger object will automatically have more resolution than a small one, and will thus be detected further away. Therefore the same object (size) should be used to compare detection distances of different video analytics.
- Illumination conditions. At night, artificial illumination is typically needed to detect anything at all.
- Contrast of object to background. A camouflaged object is much harder to detect than one with a good contrast to the background.
- Object motion direction. For an object moving towards the camera, the perceived motion and change in video is much less then for an object crossing the field of view. Therefore the distance at which they are detected can be less.
- Other challenging video analytics environments including shaking camera, grass / tress moving in the wind, snow, waves.

## 1.3 Number of objects

For applications like people counting and traffic applications, the number of objects that can be detected and tracked simultaneously is important. In busy scenes, objects easily occlude each other, and the amount of occlusion an object can have and still be detected is also important.

## 1.4 Features

Typical and useful features for intrusion detection are

- Object within / enter / exit field
- Line crossing
- Loitering
- Classification: Human / Vehicle
- Object size filter
- Object direction filter

Beyond that, other features include

- Counting
- Occupancy / amount of objects in a field
- Tailgating
- License plate recognition

Besides the configuration for live alarming, some video analytics offer to search for recorded alarms, or to change the configuration for a full forensic search. As a full forensic search can also be used to optimize the configuration, it is also valuable for live alarming.

As video analytics live inside the full video surveillance environment, it also needs to be checked whether the target video management system supports the video analytics, and in which depths. For Bosch video analytics, check the integration partner program (IPP) web page (http://ipp.boschsecurity.com) for integration status.

## 1.5 Scenarios

In addition to comparing configuration options, it should also be investigated whether the video analytics can cope with different scenarios like

- Sterile zones
- Water scenes
- Well-populated people-only areas
- Traffic scenes including congestion

Many video analytics have severe troubles when several objects are close to each other, and are not able to separate these objects. Thus, any evaluation of the number of objects (counting, queuing) or paths (loitering) cannot work. Usage of top-down camera views to minimize the occlusions is advisable, but cannot solve the task completely. Some video analytics assume that only people are within an area and are thus able to separate them using knowledge about the perspective in the scene. However, any shopping cart or car will then be detected as a multitude of people as well, so this is only applicable in people-only areas.
Another very challenging and often not supported area of scenarios is given by moving backgrounds like water / waves, elevators, conveyor belts and doors.

## 1.6 Ease of use

Ease of use is related to both robustness and the amount of features. The more robust a system is, the less it needs to be configured to achieve satisfactory results. The more features a system has, the more configuration options are available which might make the system seem more complex at first glance.

Typical measures to determine the ease of use are

- Average installation time
- User satisfaction surveys

One of the most complex and time-consuming tasks during video analytics setup can be the calibration, which is teaching the camera about the perspective in the scene. The perspective describes that an object close to the camera will appear larger than the same object further away from the camera. It also allows to transform the 2D camera image back into 3D measures like metric size, speed and geolocation. Typically, a calibration assumes that the camera is looking at a single, flat, horizontal ground plane only. Any objects walking on stairs, hills or additional levels will not be covered correctly when using the calibration information. The calibration information can be fully described via video sensor size, focal length, the camera angles with respect to the ground plane (tilt & roll angle) and the camera installation height. If instead of the full calibration and exact object location on the ground plane only the size and speed of objects in the video are of interest, then a partial calibration using typically 2-3 human markers can be done.

- Some video analytics don’t offer calibration at all, thus not being able to automatically correct for perspective, or to compute object size and speed information.
- Many video analytics assume that the roll angle is zero.
- Some offer fully automatic calibration methods, by observing all detected objects and deriving the calibration from there. This will only work if enough objects move through the scene. Accuracy will depend strongly on the observed objects and their distribution in the image. For sterile zones, this typically does not work as not enough objects move there. For other scenes, this still takes the video analytics minutes to hours to gather the needed object information. Though the installer does not need to supervise the automatic learning, he needs to come back to this camera later on for verification, which may cost additional time.
- Many video analytics offer semi-automatic calibration, by allowing the user to mark vertical and ground lines with their length as well as angles on the ground. Humans can also be used as vertical line markers.
- Some calibration information can be set by the video analytics by reading sensor information. Thus, the video sensor size, the angles, and the focal length can be set, leaving only the configuration of the camera installation height to the customer. The more information gathered that way, the less work needed by the installer.

It is advisable to verify that the calibration is correct independent of the method used. All methods can achieve the same accuracy levels if used expertly.

# 2 Benchmark setup

## 2.1 Representativeness

Video analytics typically run 24/7 in all weather and lighting conditions, in a multitude of different setups concerning perspective, background and complexity of the scene. It is impossible to test all of them in a benchmark setup, but effort should be taken nonetheless to provide as wide a variety as possible and necessary for the video analytics task to be benchmarked.

Typical outdoor challenges for video analytics are

- Low-light situations
- Fast illumination changes
- Snow / hail reducing visibility
- Shaking / vibrating camera
- Moving background like grass / trees in the wind or water / waves
- Low contrast of object to background
- Object moving toward the camera instead of crossing the field of view
- Object rolling / crawling towards the fence
- Fast objects near the camera
- Deep object shadows
- Groups of objects

## 2.2 Camera installation

When planning a real camera installation for a video analytics benchmark, please keep the following in mind:

- Optimal camera installation height is 4-5 meter. When using less height, performance may degrade, especially for video analytics based on ground plane calibration. More height is usually welcome for the video analytics if the mounting is stable enough that the camera does not shake / vibrate in the wind.
- The camera should look down onto the scene as much as possible for best performance. For crowded scenarios, a top-down perspective to reduce occlusions is essential. For intrusion detection, top-down is often not possible as it drastically decreases detection distance, and the angled of the camera is typically given by the camera height.
- Shaking camera. Low-cost tripods in combination with windy conditions will generate heavy camera shaking, which is a hard challenge for video analytics. Use a sturdy mounting instead.
- Use similar field of views for all cameras as far as possible.

## 2.3 Evaluation on stored video footage

An evaluation on stored video footage has the advantage that once the video footage is collected a wide variety of scenarios and environment conditions can be tested. However, nowadays many video analytics are located directly within the camera. The only path to get the video footage to those video analytics is to display it on a monitor and film with the camera from the monitor screen. This has the following disadvantages:

- Compression artefacts. Video needs to be compressed to store it efficiently. This compression leaves visual artefacts, which may degrade performance of video analytics developed to work on uncompressed video directly in the camera.
- Flickering artefacts of the monitor. When filming from a monitor screen, flickering artefacts and line artefacts can occur. This can in part be compensated by the correct camera frequency settings, in part by using a slightly unsharp camera image. In any case, performance of the video analytics may degrade.
- Calibration: As the calibration needs to represent the relation of the camera to the monitor as well as the relation of the camera with which the video footage was captured and the ground plane where the video footage was captured, it is more complex to calibrate and the resulting calibration is typically less accurate. Thus, video analytics performance can degrade. Internal sensor information about the calibration cannot be used either.
- Reflections on the monitor screen and overall illumination levels. Therefore, best practice is to use a dark room for this setup.

**Bosch Sicherheitssysteme GmbH**
Robert-Bosch-Ring 5
85630 Grasbrunn
Germany
www.boschsecurity.com