Chapter 1
Definitions and Performance Measures

1.1 What is Automatic Target Recognition (ATR)?

ATR is often used as an umbrella term for the entire field of military image exploitation. ATR Working Group (ATRWG) workshops cover a wide range of topics including image quality measurement, geo-registration, target tracking, similarity measures, and progress in various military programs. In a narrower sense, ATR refers to the automatic (unaided) processing of sensor data to locate and classify targets. ATR can refer to a set of algorithms, as well as software and hardware to implement the algorithms. As a hardware-oriented description, ATR stands for automatic target recognition system or automatic target recognizer. ATR can also refer to an operating mode of a sensor or system such as a radar. Several similar terms follow:

- AiTR: Aided target recognition. This term emphasizes that a human is in the decision-making loop. The function of the machine is to reduce the workload of the human operator. Most ATR systems can be viewed as AiTR systems in the broader context.
- ATC/R: Aided target cueing and recognition.
- ATD/C: Automatic target detection and classification.
- ATT: Automatic target tracking.
- ISR: Intelligence, surveillance, and reconnaissance.
- NCTR: Non-cooperative target recognition.
- PED: Processing, exploitation, and dissemination.
- SDE: Sensor data exploitation.
- STA: Surveillance and target acquisition.

This chapter sets the stage for the rest of the book. It defines the terms and evaluation criteria critical to ATR design and test. However, every ATR project is different. The terms and criteria presented here will need to be modified to meet the unique circumstances of individual programs. Consider a competition to choose an ATR for a particular military platform. Multiple

ATRs can only be evaluated fairly within a consistent framework—consistent in definition of terms, evaluation criteria, and developmental and test data. All parties being tested must have equal knowledge of test conditions and an equal ability to negotiate changes to the conditions. Regrettably, perfect fairness is impossible to achieve. Bias occurs because important factors cannot be controlled. One ATR developer might be the manufacturer of the sensor and know all about it; this developer is able to collect large amounts of data and tune up the ATR and sensor so that they work well together. Another developer might have influence over the test site, target set, test plan, and performance requirements. Another developer might manufacture the host platform (e.g., aircraft), the processor box, and have a long history of working with the end-user on ConOps. Another developer might simply have more time and money to prepare for a competitive test. When ATR components are tested, bias often arises when a developer has an investment in a favorite approach and gives short shrift to tuning up competing approaches. The definitions and performance measures provided here give all stakeholders a common language for discussion and can help to make competitive tests somewhat fairer, but not absolutely fair.

ATR can be used as a generic term to cover a broad range of military data exploitation technologies and tasks. These include image fusion, target tracking, minefield detection, as well as technologies for specific missions such as persistent surveillance and suppression of enemy air defenses. The term can be broadened to cover homeland security tasks such as border monitoring, building protection, and airport security. It can include environmental efforts such as detection of fires, whales, radioactive material, and gas plumes. Commercial applications similar to the military ATR problem are grouped under the name *video analytics*. These include parking lot security, speed cameras, and advanced signage. Internet companies are making huge investments in image-based search engines and face recognition. Industrial automation and medical applications of machine vision and pattern recognition use the same basic technology. This chapter focuses on the narrower military problem, epitomized by the basic ATR architecture depicted in Fig. 1.1. This architecture consists of two main components: a front-end anomaly detector (prescreener) and a back-end classifier. The classifier completes the detection/clutter-rejection process. The classifier can also assign a target category to a detected object. The performances of the two primary ATR components can be measured separately. Alternatively, the ATR can be treated as a single *black box*. In this latter case, the only concerns are the inputs and outputs. The inner workings of the ATR, that is, how it transforms the inputs to the outputs, might not be of interest to a team evaluating an ATR's technology readiness level (TRL).

Figure 1.1 shows the input data as a 2D image plus ancillary information. This will be the case in point used in this book. Other types of ATRs might

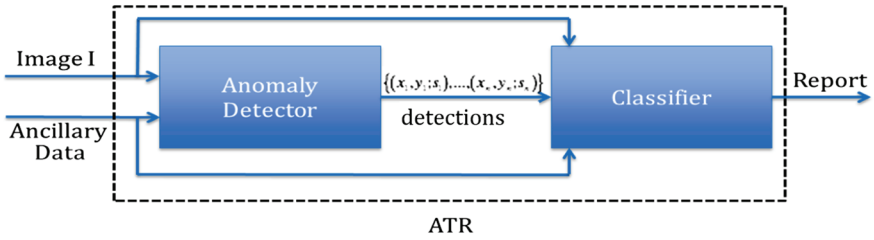


Figure 1.1 Basic traditional ATR architecture. The classifier stage stands for any or all levels of target classification that can take place, as well as supporting processing such as feature extraction and segmentation.

process different types of data, such as 1D or 3D signals, data from multiple sensors, or data in compressed form, just to give a few examples. Some ATRs do not fit the archetype shown in Fig. 1.1.

An ATR might process each input frame of data independently from the next frame, as in a synthetic aperture radar (SAR) or an infrared-step-stare system. A triggered unattended ground system can rarely generate but a single frame of data. Or, the ATR could process video data, using temporal information to help make its decisions.

An ATR often receives ancillary information. The nature of the ancillary data depends on the sensor type and system design. For an electro-optical/infrared (EO/IR) system on a helicopter, this type of metadata includes inertial data, latitude, longitude, altitude, velocity, time, date, digital terrain elevation map, laser range, bad pixel list, and focal plane array nonuniformity. The ATR can also receive target handoff information from another sensor on the same or different platform. The ATR can receive commands requesting that it look for certain targets, switch modes, or render itself useless upon capture. The ATR could as well send commands to the sensor, such as to change integration time, switch modes, or slew in a certain direction.

The borderline between the sensor and ATR is not clear cut. Either the ATR or the EO/IR sensor might perform image correction, frame averaging, stabilization, enhancement, quality measurement, tracking, or image mosaicing. The ATR might implement its own unique version of SAR autofocus and SAR image formation. Some customary ATR functions, such as image compression, could also be handled by other platform components such as a data link or storage system. In the future, the ATR could be just another function within a sensor system, analogous to face detection in handheld color cameras. Or, the future ATR might be given an expanded role to serve as the brains of a robotic platform.

The output of an ATR is a report. The report provides information about targets located and/or tracked, the ATR's health and status, an assessment of the quality of input data, etc. The report could be in the form of graphical overlays for display. The ATR might also output image data for storage or

transmission over a data link, perhaps stitching together frames or compressing target areas to higher fidelity than background areas.

1.1.1 Buyers and sellers

"Academic exercise" is a somewhat pejorative term, meaning something with little or no relevance beyond academe. An ATR study or algorithm is an academic exercise if the authors (1) have no power to implement their approach, and (2) if their recommendations are divorced from such considerations as ConOps, performance requirements, specific sensors and sensor modes, metadata, cost, timelines, logistics, competing technologies, countermeasures, independent test and evaluation (T&E), and the DoD procurement process.

This book treats ATR as a product rather than as an academic exercise. There are buyers and sellers. The buyers and sellers need to use common terminology when discussing a transaction. It is up to a buyer to describe in excruciating detail the specifications and key performance requirements of the product being procured. It is natural for sellers to describe their products in best possible terms. It is up to the buyer to do the requisite independent T&E and due diligence to determine if the seller's product meets all requirements. The following discussion should help.

1.2 Basic Definitions

Image: A 2D array of pixels.

Discrete image samples (pixels) can be single valued, representing grayscale pictures. Unless otherwise noted in the text, pixels will be regarded as 8- to 20-bit integers. For certain other sensor types, pixels can be vector quantities: dual-band, third-generation IR (2 band); visual color or commercial IR color (CIR) (3 band); multispectral (4–16 band); or hyperspectral (17–1000 band). Image samples can also be complex-valued signals, as in radar or sonar data, or they can be matrix-valued from polarization cameras. A radar can have multiple modes of operation, each producing different kinds of data. Image samples can have embedded information. For example, the most significant bit might be a good/bad pixel indicator. Some ATRs grab and digitize frames or fields of analog video. Ancillary information can be embedded in the first few lines of each frame of data. Alternatively, a file of ancillary information might be associated with each frame of image data or with multiple frames. The temporal synchronization of sensor data and metadata is a critical issue. ATR systems can also operate on 1D signal data or 3D LADAR data. For example, an ATR might process the Automatic Identification System (AIS) data broadcast by commercial ships.

The ATR and a human viewer will generally receive sensor data from different paths. For example, the ATR might receive 14-bit/pixel image data at 120 frames per second, while the human might view 8-bit/pixel video with annotation overlays at 30 frames per second. The ATR might receive complex-valued SAR data, while the human views magnitude SAR data.

Operating conditions (OCs): All factors that might affect how well a given ATR performs.

OCs characterize:

- targets (articulation, damage, operating history, etc.),
- sensor (type, spectral band, operating mode, depression angle, etc.),
- environment (background, clutter level, atmosphere, etc.),
- ATR (settings, *a priori* target probability assumptions, etc.) and
- interactions (tree lines, revetments, etc.).[1]

OCs are the independent conditions of the experimental design. A bin (experimental bin or OC bin) is data, such as a set of test images that meet some pattern of OCs. For example, one bin might include only day images, and another bin only night images. Even simple terms such as day and night should be clearly defined.

Ground truth: Reference data available from a data collection.

This information is generally of two types:

(1) scenario information: climatic zone, weather, time, date, sun angle; target locations, types, conditions, etc.
(2) sensor information: sensor location, pointing angles, operating mode, characteristics, etc.

Ground truth is a term used in various fields to refer to the absolute truth of something. Thus, it can refer to truth about ships and space targets, not just ground targets. Although ground truth might provide target location, velocity, direction, and range [for example, by global positioning system (GPS) transponder on each target], it will not indicate the pixels in the scene that are on target. Determining which pixels are on target is not as easy as it might at first seem, as illustrated in Fig. 1.2.

Target: Any object of military interest.

Traditional targets are strategic and tactical military craft. This will be the case in point used in this text. However, today, the list can also include improvised explosive devices (IEDs), enemy combatants, human activities, muzzle flashes, fixed sites, commercial vehicles, land minefields, tunnels, undersea mines, and technicals (commercial vehicles modified to contain armament).

Figure 1.2 Pixels on target must be labeled according to a set of rules.

Image truth target location or region: A single reference pixel on target or set of pixels on target (target region) as estimated by an image analyst, using ground truth when available.

Bounding box: Rectangle around all of the target or the main body of the target.

For forward-looking imagery, the bounding box is generally rectilinearly oriented (Fig. 1.3). For down-looking imagery, the bounding box will be at an angle with respect to the axes of the image (Fig. 1.4).

For forward-looking imagery, ground truth target location will generally be pinned to the ground surface rather than at the center of the grayscale mass of the vehicle. This is because the range to the point that the target touches the ground is different from the range along the view-ray through the target center to the ground. This truth will have an associated target location error (TLE) in geographical and pixel coordinates. The TLE for database targets can only be specified statistically. The truthing process might indicate the set of pixels on the target, known as the *target region*. These pixels can match the shape of the target, or be more crudely specified as a rectangular (as in Fig. 1.3) or elliptical region. The target region is generally, but not always, contiguous. A target region can even be smaller than a single pixel, as for the case of low-resolution hyperspectral imagery.

Target report: Report output by ATR generally providing, as a minimum: location in the image of detection (by its reference pixel), the equivalent

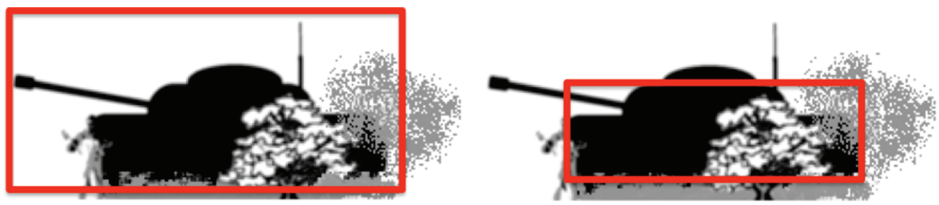


Figure 1.3 Illustration of a bounding box (a) around the entire target or (b) around the main body of the target.

Figure 1.4 Boxes around targets in overhead imagery can be at any angle.

location as latitude and longitude on an earth map, various categories of classification assigned to the target, and associated probability estimates.

The information contained in the target report can be quite extensive, but only parts of it can be disseminated due to mission and bandwidth. A popular protocol is MITRE's Cursor-on-Target (CoT). The CoT event data model defines an XML data schema for exchanging time-sensitive position of moving objects between systems: "what, "when," and "where" information.

Target location and/or region as reported by ATR: Estimated target reference pixel p_{ATR} or region R_{ATR} as provided in an ATR's report.

The ATR will report a target location. This can be the target's geometric center, the center of grayscale mass, the center of the rectangle about the target, the brightest point on the target, or a point where the target touches the ground. The ATR might estimate the pixels on target through a segmentation process. ATR engineers should understand the scoring process and end-user requirements, so as to know how best to report a target.

Target detection: Correct association of target location p_{ATR} or target region R_{ATR}, as reported by the ATR, with the corresponding target location p_t or target region R_t in the truth database.

Detection criterion: The rule used to score whether an ATR's reported target location or region sufficiently matches the location or region given in the truth database.

Note that the truth database can contain mitigating circumstances for which the ATR is given a pass if it doesn't detect particular targets. Such circumstances can be: target out of range, not discernable by eye, mostly obscured, half-off image, covered by camouflage netting, etc. Such objects are referred to as *non-spec targets*.

Once a tracker locks onto a detected target, performance is measured by rules associated with trackers rather than detectors. Tracker evaluation criteria are well established, but are not covered in this book.

Specific kinds of detections include:

- **Multiple detection**: Detections on a target, beyond the first or strongest one reported (for a single frame).
- **Group detection**: A single detection on an assemblage of objects in close proximity, such as a huddle of combatants.
- **Event detection**: Detection of an occurrence, such as: a missile ready to launch, persons unloading a truck, or a person planting an IED.
- **Flash detection**: Detection of the location in image coordinates of a muzzle flash.
- **Muzzle blast detection**: Detection in geographic coordinates of the origin of the auditory (sound) and non-auditory (overpressure wave) components after a muzzle flash.
- **Change detection**: Detection of something in an image that wasn't perceived at that location at a previous point in time.
- **Detection of disturbed earth**: Place where an IED or landmine might have been buried.
- **Standoff detection**: Detection of a dangerous object from a safe distance.
- **Brownout detection**: Detection of the presence of a dust cloud degrading the visual environment.
- **Extended-object detection**: Detection of something very long with no obvious beginning or end, such as power lines, a pipeline, a tunnel, a string of landmines or an underwater cable.
- **Fingerprinting**: Detection not of a *type* of vehicle, but one *particular* vehicle, for example, the car with the bombers in it.

Caveats and ambiguities

Although we will use the common term *image truth*, we note that what is normally referred to as image truth is more realistically *expert opinion*, rather than absolute omnipotent truth. Image truth is often produced by one or more image analysts using ground truth information when obtainable. Image truth can include supporting information such as target aspect angle, image quality near target, and the number of pixels on target. Image truth can also include the truther's opinion of clutter level. Although the image truthing process

currently involves significant manual labor, work is underway to automate parts or all of the process.

Image truth is best produced during a data collection rather than months afterwards. Image truth will contain errors. For example, in IR imagery, some parts of the target will fade into the background. Exhaust can heat up the ground. A puff of smoke or kicked up dust can obscure the target or appear to be part of the target. Before starting image truthing, it must be made clear what should get labeled as part of a target. Consider possible cases of ambiguity (some of which are illustrated in Fig. 1.2): a bush in front of the target, antenna, flag, chain, open space within target's convex hull, gun and backpack carried by dismounted combatant (dismount), target behind another target, vehicle transported on truck bed, camel carrying combatant or weapons, decoy, netting draped over and off of target, hulk of vehicle, fuel supply vehicle adjacent to target vehicle, object towed by target, target shadow, false data produced by turbulence, dust trail behind moving vehicle, aircraft's contrail, ship's wake, etc.

If a scene is synthetically generated, pixels on target are known. Even then, a decision must be made about how to label a pixel that is part target and part background.

Specifying region on target, whether by man or machine, is a nice concept in theory. However, in practice, it might not be possible in some circumstances. Parts of an object in thermal imagery might be darker (colder) than the background, and other parts much hotter, but much of the target can be of similar temperature to the background. This can happen if the object is a vehicle (Fig. 1.5) or dismount. In such cases, it is not obvious which groups of pixels combine to form the region on target. In the visible band, painted camouflage patterns on vehicles and camouflage uniforms can make segmentation of vehicles or soldiers problematic, depending on the background color and texture. In overhead visible band imagery, dark vehicles tend to blend in with

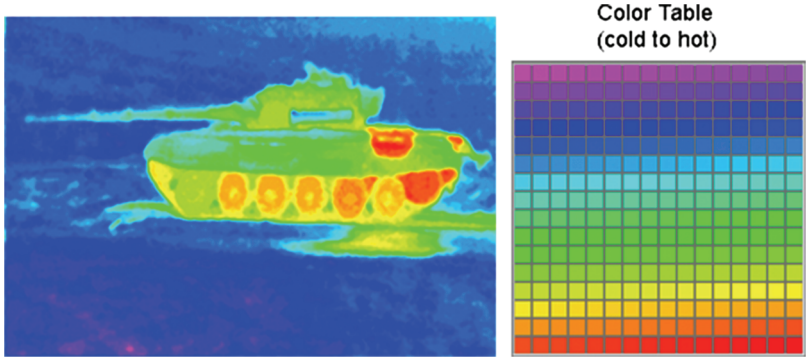


Figure 1.5 Target data may be multimodal in IR imagery. Some pixels can be much hotter than the background, while other pixels match the background temperature. (Shown in false color in electronic book formats.)

shadows. In SAR, it is difficult to determine target region when vehicles are tucked into a tree line.

1.3 Detection Criteria

It is quite challenging to precisely and unambiguously stipulate what is meant by *target detection*. Let us first consider some relevant terms:

$|R|$ = cardinality of R = the number of pixels in region R.
Let R_t = region on target as indicated by truth data.
R_{ATR} = region on target as reported by ATR.
p_t = point (or reference pixel) on target according to the truth data, i.e., the target reference pixel.
p_{ATR} = point (or reference pixel) on target as reported by ATR.
$\|a - b\|$ = distance between points a and b.

First, let us suppose that the ATR outputs a single detection point per object and the truth database contains a single detection point per target. Let

$A = \{p_{ATR}\}$ denote the set of detection points output by the ATR, and
$T = \{p_t\}$ denote the set of detection points in the truth database.

The set C of correct detections output by the ATR is such that each detection in C matches a target in the truth database T according to some match criterion. Here, we will define several common detection criteria, illustrated in Fig. 1.6.

Minimum distance criterion: If the minimum distance between an ATR reported target point and the nearest target point in the truth database is less

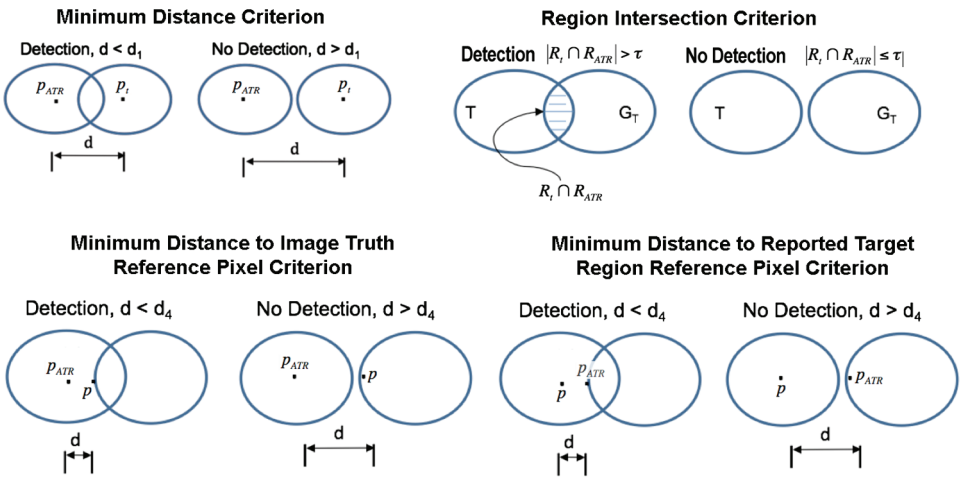


Figure 1.6 Illustration of several detection criteria.

than a preselected value d, then the ATR has detected a valid target, as defined by

$$p_t \in C \text{ iff } \min_{p_t \in T} \|p_{ATR} - p_t\| \leq d.$$

The number of correct detections is given by $|C|$.

This definition allows for, at most, one correct detection on each target in the truth database T. Note that for the purpose of scoring an ATR, if the exact same vehicle appears in more than one image, it is usually considered a different target for each image in which it appears. Detecting the exact same target in two different images results in two correct detections. It is possible, but unlikely, that one detection point reported by the ATR will result in a count of two or more correct detections, for example, if part of a target is in front of another target.

Another way of defining a correct detection is if a target region in a truth database intersects the corresponding target region output by the ATR.

Region intersection criterion: $|R_t \cap R_{ATR}| > \tau$, where τ is a threshold.

Alternatively, a target indicated by a single reference pixel can be said to be detected if the reference pixel falls within a target region. This leads to two additional detection criteria.

Minimum distance to image truth reference pixel criterion: $\min_{p \in R_t} \|p - p_{ATR}\| < \tau$.

Minimum distance to reported target region reference pixel criterion: $\min_{p \in R_{ATR}} \|p - p_t\| < \tau$.

The opposite of a true detection is a *false alarm*. Defining a false alarm involves the concept of clutter.

Clutter object: Non-target object with characteristics similar to those of a target object. A clutter object can be natural or manmade. A clutter object can be ephemeral, such as a hot patch of ground or an opening between two trees.

False alarm: A detection reported by the ATR that does not correspond to any target in the truth database, according to some agreed-on detection criterion.

With this definition, it is possible for the ATR to get penalized for multiple false alarms by reporting multiple detections on the same clutter object within a single image. The ATR could also be penalized for a false alarm by reporting a point between two adjacent targets. This penalty could be relaxed for a particular system if its purpose is to extract (substantially larger than target) regions-of-interest to display to a human operator. For this type of system, two adjacent vehicles would be displayed to the operator for further decision, even if the detection point is between them. The same would be true

for a huddle of combatants. It generally won't be necessary to detect each and every person in the huddle. However, we have seen government tests where detection of each person in a close group was required.

Region-of-interest (ROI): A rectangular image chip about a detected object. It may be scaled as a function of range.

Several examples of ROIs are shown in Fig. 1.7. It is preferable that the detected object be near the center of the ROI.

False alarm rate: A measure of the frequency of occurrence of false alarms in a reference context.

False alarm rate (FAR) is measured differently for forward-looking sensors compared to downward-looking sensors. This is because with a forward-looking sensor it is not possible to determine the ground area covered by the image. Instead, a measure of the solid angle of the optics is used. For example, a sensor might have a horizontal field of view of 2 deg and a vertical field of view of 1.5 deg, or equivalently, 3 square deg. Various measures of FAR can be defined based on different reference contexts:

$$\text{Pixel FAR: } FAR = \frac{N_{FA}}{N_{MP}} = \frac{\text{number of false alarms}}{\text{number of megapixels processed}}.$$

$$\text{Frame FAR: } FAR = \frac{N_{FA}}{N_{FP}} = \frac{\text{number of false alarms}}{\text{number of image frames processed}}.$$

$$\text{Area FAR: } FAR = \frac{N_{FA}}{N_{km}} = \frac{\text{number of false alarms}}{\text{sum of ground area covered by images processed}}.$$

$$\text{Temporal FAR: } FAR = \frac{N_{FA}}{\Delta_{time}} = \frac{\text{number of false alarms}}{\text{time interval}}.$$

$$\text{Angular FAR: } FAR = \frac{N_{FA}}{N_{SD}} = \frac{\text{number of false alarms}}{\text{number of square degrees processed}}.$$

Figure 1.7 Examples of ROIs from several types of sensors: (a) IR, (b) visible, (c) SAR, and (d) sonar.

Ambiguities in measures of false alarm rate

There is some ambiguity in each of these measures. Ambiguity should be eliminated in a particular project based on more precise definitions that take into account the important issues for the project. For example, the ATR might not be able to process and detect targets toward the edges of images, meaning that the total number of pixels in the image set differs from the number of pixels fully processed. The ATR might not process the region above the skyline. Frames might overlap, such that the sum of the ground area processed for the ensemble of frames is greater than the actual ground area covered. Frames or parts of frames might be discarded due to insufficient image quality, or some areas may be outside of range limits. The left side of a non-target might appear in one frame and the right side in the next step-stare frame. With video data, the same rock could produce 30 false alarms per second. Should this be counted as a single false alarm or as multiple false alarms?

There are many other types of ambiguities that must be resolved. If, for example, a military truck or plane is considered a target, should detection of the equivalent civilian truck or plane be treated as a correct detection or as a false alarm? Should the detection of a military vehicle outside the list of targets sought be treated as a correct detection or a false alarm? That is, is the target set open or closed? Some of these issues can be resolved by defining a "don't care" class. Detection of a *don't care object* has no effect on scoring. A calibration target is always a don't care object. A calibration target could be a corner reflector in a radar test, a thermal target board in an IR test, or a color panel in a hyperspectral test.

One problem with scoring a very large geographic area is that objects outside of a military compound will not have associated ground truth. Could there be a military vehicle or similar civilian vehicle on the road to the military base or used as decoration at an armory? If the objective is to verify a rate of 0.001 false alarms per km^2, an area nearly the size of Connecticut would need to be truthed. This is not feasible.

1.4 Performance Measures for Target Detection

Performance measures include truth-normalized measures, report-normalized measures, and various graphical depictions.

1.4.1 Truth-normalized measures

Probability: A way of expressing knowledge or belief that an event will occur *or* has occurred.

There is often some confusion resulting from the dual nature of the definition of *probability* as "has occurred" or "will occur." ATR engineers use

probability rather loosely to mean measured ATR performance over a particular database. Probability is not an unqualified prediction about the future. It is only a measure of performance in a controlled experiment. It only serves as a prediction of future performance to the extent that the future data has characteristics similar to the data processed.

Probability of detection: The probability that the ATR associates a non-redundant detection with a target in the truth database:

$$P_d = \frac{|C|}{|T|} = \frac{\text{number of correct target detections}}{\text{number of ground truth targets}}. \tag{1.1}$$

P_d by itself does not have much utility. It is always possible to declare a detection at each pixel in an image and achieve $P_d = 100\%$. It is the tradeoff between missed detections and false alarms that counts. In textbook terms, this is the normal tradeoff between Type I and Type II errors (Table 1.1).

Probability of a miss: $P_{miss} = 1 - P_{det}$.

Probability of false alarm: The number of false detections normalized by the number of opportunities for false alarm:

$$P_{FA} = \frac{|F|}{|O|} = \frac{\text{number of false alarms}}{\text{number of false alarm opportunities}}. \tag{1.2}$$

The number of false alarm opportunities is an imprecise concept. It is often determined as follows. A polygonal tile is chosen to match the average size of a ground truth target. The size can be a function of range. The number of tiles required to cover the image set is considered to be the number of false alarm opportunities. This doesn't make much sense for forward-looking scenes containing a vanishing point, trees, and sky.

Suppose that we are just testing the back end of the ATR over ROIs. For a fixed database of target and clutter ROIs, the number of false alarm opportunities is then the number of clutter ROIs in the test database. This use of the term P_{FA} makes more sense.

Table 1.1 Tradeoff between Type I and Type II errors.

		Decision	
		Target	**Clutter blob**
Truth	Target	Correct detection (true positive)	Missed detection (Type II error)
	Clutter blob	False alarm (Type I error)	Clutter rejection (true negative)

1.4.1.1 Assigned targets and confusers (AFRL COMPASE Center terminology)

An ATR assigns *cue*s (ID labels) to objects in the test database that sufficiently match signatures in a target data library. The function of this ATR is then solely to assist or "cue the operator." The ATR is referred to as an *automatic target cuer* (ATC). An *assigned target* is a particular target type selected from the target library by the human operator. The ATR can be directed to find that specific target type (or perhaps several assigned target types). This defines the Mission of the Day. A confuser is a target-like object intentionally inserted into an experiment to determine whether or not it confuses the ATR. The ATR correctly rejects a confuser that does not meet its decision criteria for an assigned target. *Bad actors* are confusers that, for a given assigned target, inordinately contribute to the cue error rate.

Cue correct rate (CCR): The ratio of the number of *correct* assigned target cues to the total number of assigned target cues.

Confuser rejection rate (CRR): The percent of confusers that are rejected, i.e., determined not to be an assigned target.

1.4.2 Report-normalized measure

Probability of detection report reliability: The probability that a detection reported by the ATR is a true target:

$$P_{DR} = \frac{|C|}{|A|}. \tag{1.3}$$

1.4.3 Receiver operating characteristic curve

Suppose that the ATR associates a detection strength (score) with each raw detection. Figure 1.8(a) shows sample probability density curves for true targets (true positives) and non-targets (true negatives) versus computed detection strength. This ATR makes soft decisions. It is not declaring objects to be targets. It is only assigning a degree of *targetness* to detected objects. Suppose a threshold τ is set after all of the target reports are generated. If only those detections with strength above the threshold are reported target decisions, then this ATR will have a fixed probability of detection versus probability of false alarm for this test set. If the threshold is adjusted up and down, then a P_d versus P_{fa} curve results [Fig. 1.8(b)]. This is known as the ATR's receiver operating characteristic (ROC) curve. The ROC curve applies to a particular test set and as such is not a prediction of future performance on data of a different nature. Thus, a ROC curve is simply defined as in the definition that follows.

ROC curve: Plot of P_d versus P_{fa}.

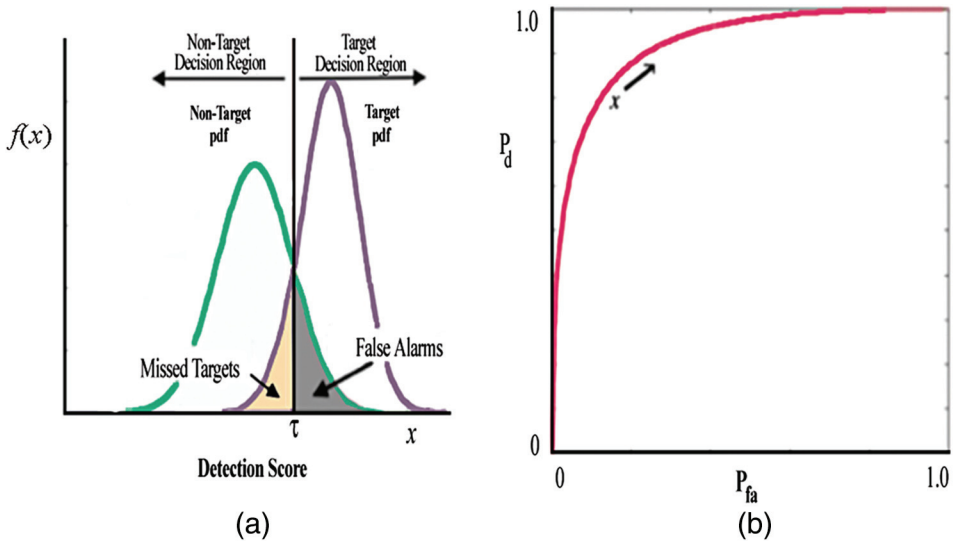


Figure 1.8 (a) Target and non-target probability distribution function (pdf). (b) Probability distribution functions transformed into an operating curve. (See Refs. 2 and 3 for in-depth discussions.)

ROC analysis was developed in the 1950s for evaluating radar systems. The term has since been applied to other types of systems, without regard to its original more specific meaning. Each point on the ROC curve represents a different tradeoff (cost ratio) between false positives and false negatives. The ROC plot thus provides a convenient gestalt of the tradeoff between detection and false alarm performance. If two ROC curves do not intersect (except at their endpoints), then the ATR corresponding to the higher curve performed better than the other. If two ROC curves intersect once, then one ATR performed better at low P_{fa}, and the other performed better at higher P_{fa}.

If a particular ATR can only make a hard decision, then this ATR has no ROC curve, even though it invariably has internal settings that can be adjusted in software. Regardless of the case, there is no assurance that setting a particular threshold on an operational system will produce a pre-specified performance level on new data.

The concept of a ROC curve obtained by adjusting a single threshold doesn't hold up to scrutiny. If the ATR were actually designed to operate at an extremely low false alarm rate, algorithmic changes would be required for best performance. If the ATR were to operate at a very high false alarm rate, changes would also be needed, such as reporting a longer list of raw detections out of the pre-screener. As illustrated in Fig. 1.9, The ROC curve is better suited for comparing different ATR back-end final detectors over a well-chosen set of target and clutter ROIs.

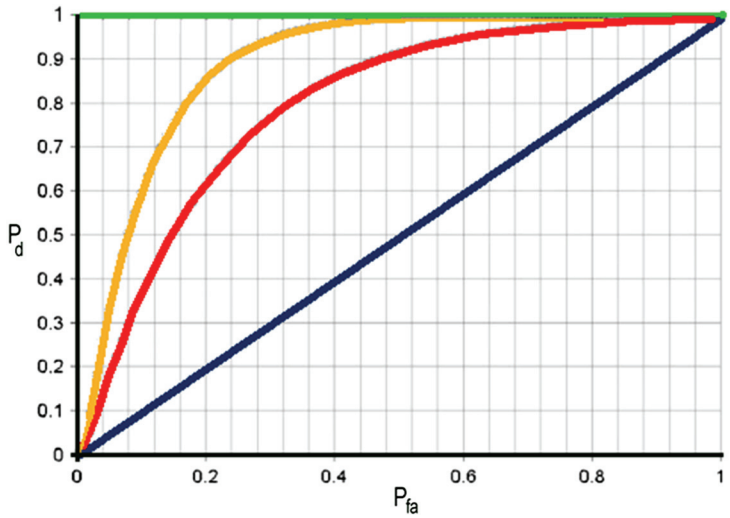


Figure 1.9 ROC curves for four ATRs: The top line represents a perfect ATR. The ATRs with performance indicated by progressively lower curves performed progressively worse. The bottom line indicates an ATR that can't tell the difference between a true target and a non-target.

A good place to find in-depth analyses of ATR performance evaluation methodologies is in Air Force Institute of Technology Ph.D. dissertations, which can be found on-line at the Defense Technical Information Center (DTIC®). ROC curves, the area under ROC curves (AUCs), and alternative ways of comparing ROC curves are provided by Alsing and Bassham.[2,3] The top curve in Fig. 1.9 has a higher AUC than the bottom curve. Two ROC curves can be compared by their AUCs. AUCs make no assumptions concerning target and non-target distributions. Alsing says that if the comparison of classifiers is to be independent of the decision threshold, AUC is a reasonable "metric."[2] However, the AUC measure is not quite a proper metric. Two ROC curves with totally different shapes can have the same AUC value. This violates the *definiteness* property of true metrics. As an alternative, Alsing suggests the use of a multinomial procedure to evaluate competing classifiers.[2] Rather than simultaneously comparing the classifiers over the entire test data set, a multinomial selection procedure compares the performance of each classifier on each data point using some scoring measure.

Bassham analyzes several variants of the ROC curve including localization, frequency, and expected utility, as well as an inverse ROC curve called a response analysis characteristic curve. He also analyzes methods for comparing ROC curves, including:

- average metric distance: the average distance between two ROC curves, using some distance metric,
- area under ROC curve that is above diagonal chance line,

- Kolmogorov method: nonparametric confidence bounds are constructed around ROC curves based on Kolmogorov theory.[3]

Bassham's thesis has a good discussion of the performance measures used during ATR development compared to measures that capture the operational effectiveness of ATRs.

Another variation on the ROC curve is obtained by making whatever internal changes are necessary for the ATR to presumably function best at different rates of false alarm. This generates a collection of $\{P_d, P_{fa}\}$ pairs for a test set. Connecting performance points will produce a P_d versus P_{fa} curve, which is not necessarily well behaved.

1.4.4 P_d versus FAR curve

Figure 1.10 gives an example of a P_d versus FAR curve. Like the ROC curve, each point on this curve corresponds to a different detection threshold. That is, the ATR is run on a set of data and reports a set of detections, each with an associated strength. In this example, FAR is plotted per square degree of sensor viewing angle, but any measure of FAR could be used. FAR is often plotted on a log scale.

Many such plots would characterize a single system test. For an EO/IR system, these may cover different OCs: fields of view, ranges to target, times of day, clutter level, target types, etc. One would compute separate curves to report performance over dismounts and vehicles. For a SAR system, conditions requiring different P_d versus FAR curves include sensor resolution, clutter level, and target categories.

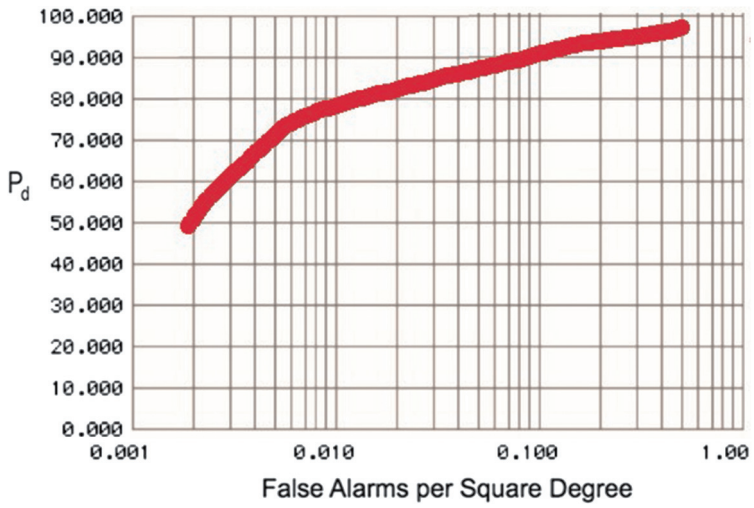


Figure 1.10 Example of a P_d vs. FAR curve.

1.4.5 P_d versus list length

Suppose that the ATR follows the simple model of Fig. 1.1. Suppose that the front-end anomaly detection stage outputs n raw detections per image frame ordered by strength. The processing requirement of the back-end of the ATR is directly proportional to n. The back-end of the ATR has limited processing capacity, which sets an upper bound on n. The longer the list of raw detections out the front-end of the ATR the harder it is for the back-end classifier to reject every non-target. P_d can be plotted against n to determine if there is a point of diminishing returns. Various front-end detectors can be compared in this manner (Fig. 1.11).

1.4.6 Other factors that can enter the detection equation

The equations given so far are for the basic case. A complete equation for a specific project can include other terms relevant to the experimental design and ConOps, such as:

- number of redundant detections on targets,
- number of detections on decoys,
- number of detections on don't care objects,
- number of detections on targets of unknown type,
- number of front-end detections for which the back-end of the ATR makes no decision (this is the opposite of requiring a forced decision),
- number of detections on objects specifically put into the database as confuser objects, and
- number of detections on *spec targets*, i.e., targets meeting specified criteria.

1.4.7 Missile terminology

Determining the effectiveness of a missile or missile defense system involves modeling and simulation (M&S). Verification and validation of the M&S are

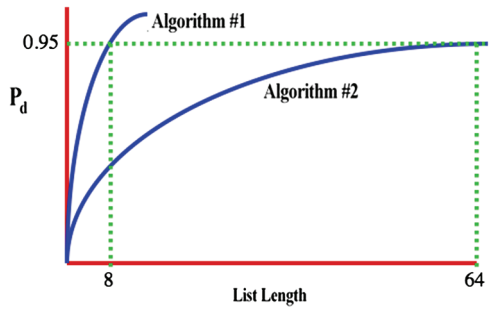


Figure 1.11 Plot of P_d versus list length for two front-end detectors. One detector requires a list of length 8 to detect 95% of targets, while the other requires a list of length 64.

extremely complex. The single missile kill chain is represented by a sequence of events. Each event has its own probability of failure. Each step in the chain decreases the final probability of kill.

Engineers developing missiles and other types of munitions tend to use terminology that differs from that used by the ATR community. A simple example follows:

$$P_{ssk} = P_h \times P_d \times R_m \times R_w,$$

where
P_{ssk} is the probability of a single shot kill,
P_h is the probability of a hit,
P_d is the probability of detecting the target,
R_m is the reliability of the missile, and
R_w is the reliability of the weapon.

1.4.8 Clutter level

Target detection performance should be reported with regard to the clutter level of the database over which performance is measured. There have been many attempts over the years to develop an equation or procedure for characterizing clutter level in images as an alternative to expert opinion. None of these attempts have been thoroughly successful. The problem is one of circular reasoning. If an ATR's P_d versus FAR curve is bad, then the clutter level must have been high as perceived by that particular ATR. A different ATR using different features and algorithms may perform better on the same data. To this ATR, the clutter level is low. Furthermore, if an algorithm can measure clutter level, then it must be able to distinguish targets from clutter, which itself defines an ATR.

Richard Sims did some of the best work in this area.[4] His signal-to-clutter-ratio (SCR) metric is given by[4]

$$SCR = \sum_{i=N_{c+1}}^{N} \frac{\lambda_i}{1-\lambda_i} + \sum_{i=1}^{N_c} \frac{1-\lambda_i}{\lambda_i}. \tag{1.4}$$

This equation derives from eigenvalues of the Karhunen–Love decomposition of a target-sized image region—referred to as the Fukunaga–Koontz transform. The first term encompasses all eigenvalues, denoted by λ_i, where the target dominates. The second term is a measure of the useful information where clutter dominates. The reader is referred to referenced papers for details.[4,5]

1.5 Classification Criteria

Classifier categorization is often represented graphically, while performance is given by a table.

1.5.1 Object taxonomy

In the context of ATR, ontology is a subject of study involving the categories of objects relevant to an experiment, mission, or battlespace. The product of the study, called an *ontology*, is a catalog (a.k.a. library) of the objects assumed to exist in a domain of interest from a military perspective, as well as more precise specification of the basic categories of the objects and their relationships to each other. Such objects can be grouped, related within a hierarchy, and subdivided according to their similarities or differences.

The first stage of a modeling is called *conceptualization*. An ontology is a formal explicit specification of the conceptualization. It provides a shared vocabulary that can be used to model the types of objects in a military domain and their relationships—with sufficient specificity to develop and test an ATR. Ontologies consist of concepts that can be structured hierarchically, thus forming a *taxonomy*.

Taxonomy: Objects arranged in a tree structure according to hyponymy (is a) relations.

For example, a T-72 *is a* tank. A taxonomy places all of the objects into a hierarchy and clarifies the possible labels for an object at various category levels. An example is given in Fig. 1.12.

A taxonomy is a structure for classification. A classifier assigns detected objects to categories based upon the taxonomy. In practice, the classification categories are predetermined, but not exhaustive. That is, the taxonomy will not cover all military vehicle types in the world but should include all vehicles of interest to a military program, mission, or experiment. The categories are

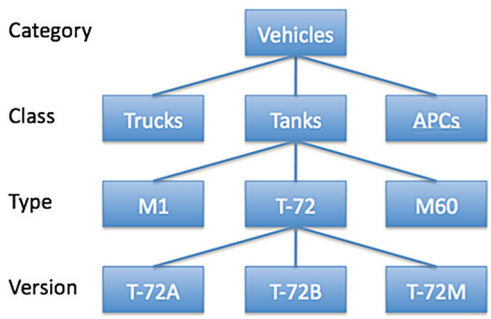


Figure 1.12 Example of a taxonomy.

exclusive. For any level of the taxonomy, an object can be assigned to category A or B, but not both.

A label of *other* can be included at any level of the taxonomy. For example, at the Type level, *other* can mean any other type of main battle tank not explicitly listed.

Decision tree: A visualization of the complex decision making taking place within an ATR, illustrating possible decisions and possible outcomes, and modeled on the hierarchy of the taxonomy.

The decision tree illustrates all possible classification outcomes and the paths by which they can be reached. While the taxonomy answers the question, "What is a T-72?" the decision tree answers the question, "What kinds of tanks are of interest?"

A decision tree is used as both a visual aid and an analytical tool. It uses its tree-like graph to model the flow of decisions. The decision tree emanates from a starting point (usually a root node at the top of the diagram) and continues through a series of branches and nodes until a final result is reached at the bottom end (leaf of upside-down tree). It illustrates how an ATR's classifier could go about making its pronouncements, step-by-step, coarse-to-fine. (However, we will give some examples later as to why the decision tree structure might not properly model the operation of a particular ATR's classifier.)

At any level of the decision tree, a classifier can make a *declaration*, which is a decision to provide a label. The label corresponds to a node name within the taxonomy. A classifier might, for example, declare a detected object to be a T-72 but may not be able to specify the version of T-72. In this case, all of the levels of the decision tree above the T-72 node would be declared, but not those below the T-72 node.

It is common practice to label the levels of the decision tree by names corresponding to the specificity of the decisions. Thus, in order of increasing specificity, names such as detection, classification, recognition, and identification are commonly used. Such terms must be clearly defined in the context of a particular program. (In this usage, "classification" refers to a specific level of the decision tree rather than the overall operation of the ATR's classifier stage.) Decision trees can be quite different for different programs. In the Dismount Identification Friend or Foe program, dismounts were said to be *identified* if it was determined that they were carrying large weapons rather than confuser objects such as 2×4 lumber or farm tools. For a classifier used to screen people entering a military base, identification could mean naming the person. Several examples of decision trees are given in Fig. 1.13.

Taxonomies and decision trees: ambiguities and exceptions

A node at each level of a taxonomy has only one parent node. It may not be clear how to arrange the taxonomy when this condition is violated, that is,

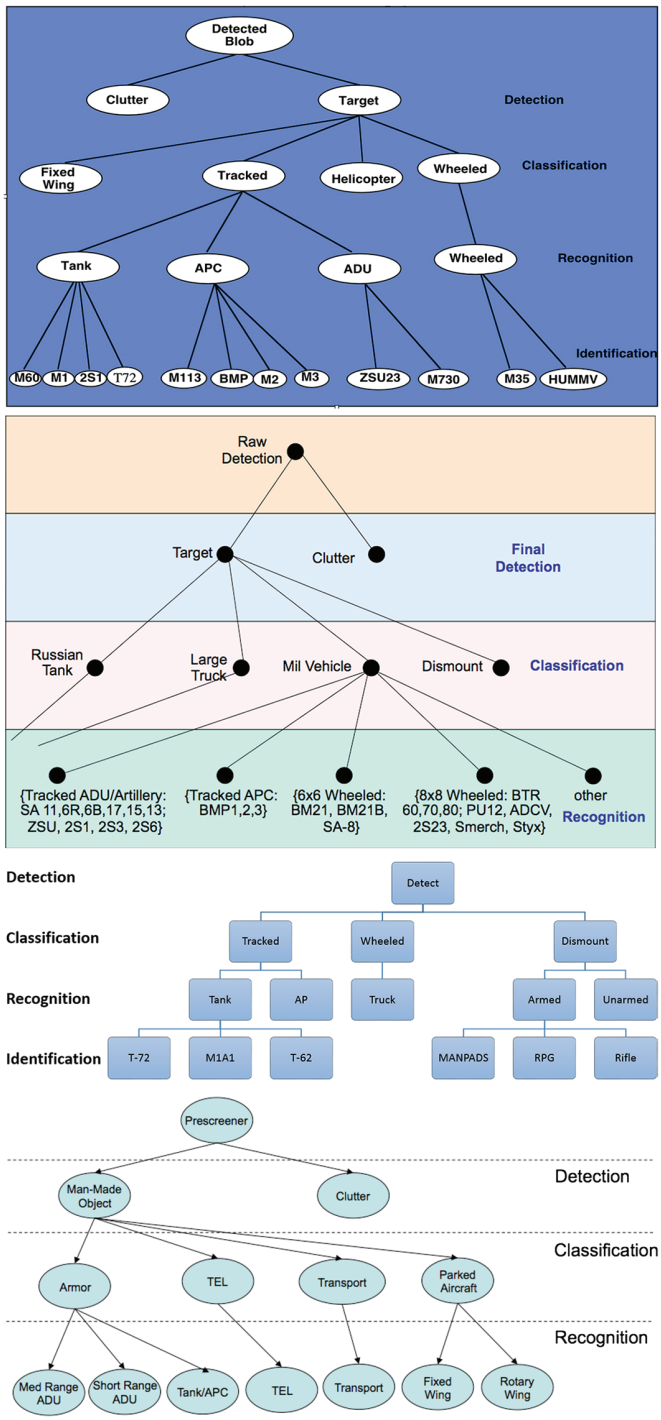


Figure 1.13 Four examples of decision trees. The third example is adapted from Ref 6.

when some vehicles fit into more than one category per Venn diagram (see Fig. 1.14). Several ambiguous cases (from an ATR perspective) follow:

- friend or foe (e.g., enemy T-72 versus NATO T-72),
- tracked or wheeled (e.g., SA-19 on tracks versus SA-19 on wheels), or
- scout car or Air Defense Unit (ADU) (e.g., BRDM scout car with SA-9 ADU weapon).

It is generally the responsibility of those funding a program to specify the targets of interest and their taxonomy. However, the funding organization might not have a good grasp of this issue. An often heard comment is "just try your ATR on the data and see what happens," not understanding that the ATR must be trained to operate with a specified taxonomy. Another common problem occurs when the funding organization supplies test data with targets not fitting the stipulated taxonomy and then complains that obvious targets are not being reported.

A taxonomy is easily transformed into a decision tree. However, a particular ATR's decision process might not fit that of the decision tree corresponding to the taxonomy. For example, a template matcher might only operate at the Type level. Class is obtained by generalizing from Type. Another ATR might utilize separate neural network classifiers for each level of the taxonomy. There is no guarantee that a decision made at the Class level will correspond to a generalization of the decision made at the Type level. An ATR might receive data from multiple sensors, some of which support decisions at one level of the tree and others that support decisions at other levels. For this case, the decision making process might be much more

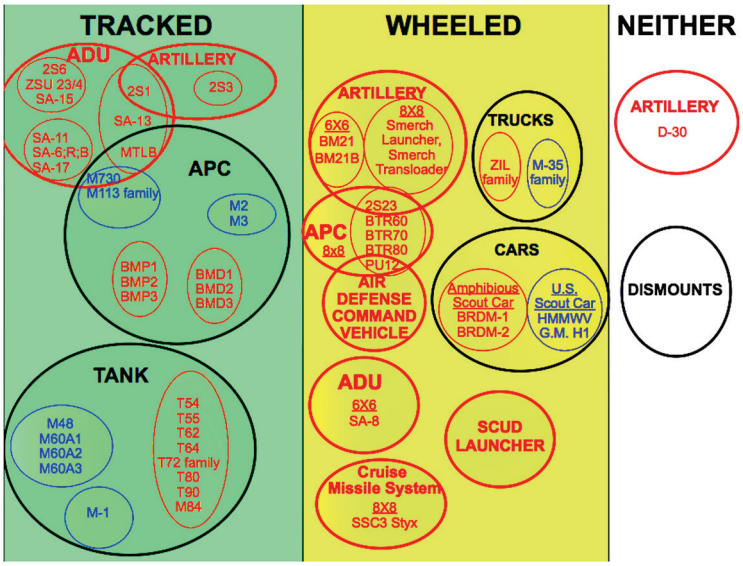


Figure 1.14 Simplified Venn diagram for some common targets types.

complex than indicated by a simple decision tree. Decisions can also change over time as a target is tracked, the sensor and ATR switch modes, or off-board information is received.

1.5.2 Confusion matrix

An *error matrix* quantifies the discrepancy between fact and the ATR's opinion. Each column of the matrix represents instances of a reported category, while each row represents the actual category. With a strict interpretation, measures of error are realizable only when truth is known absolutely.[7] In an ATR test, this would occur with synthetically generated data.

In ATR tests with field-collected data, *truth* is more fittingly called *expert opinion*. Human *truthers* provide their expert opinion of a reference point or pixels on a target, using available ground truth (instrumented data) to help determine target type and location. Target labels can be error prone for targets that are imaged but not intentionally put into the test site, such as vehicles at a military base outside of the planned test site. ATR test results are reported in a *confusion matrix* rather than an *error matrix*. The confusion matrix measures the ability of an ATR to generalize from its training data to the test data, within the accuracy of the image truth. (In this context, training includes the development of templates for a template matcher or storing of vectors for a nearest neighbors classifier.) The ATR's classification performance is evaluated from the data in the confusion matrix.

A confusion matrix characterizes the ATR classifier stage's ability to assign categories to detected objects. If the ATR has an adjustable detection threshold or other adjustable internal parameters, then the confusion matrix addresses performance at those particular settings. For example, the front-end detection stage of the ATR might be set up to operate so that only very strong targets are detected. The classifier stage of the ATR will then have an easier time assigning these strong detections to categories than if the front-end detection stage were less restrictive, i.e., operating "full throttle."

Probability of (correct) classification: Number of objects correctly classified divided by total number of objects classified.

For Table 1.2, $P_c = \dfrac{a+e+i}{a+b+c+d+e+f+g+h+i}$.

The form of the confusion matrix in Table 1.2 indicates that this test is only over objects known to be targets. Otherwise, there would be an additional row and column labeled "non-target."

Ambiguities and caveats associated with confusion matrices

An ATR can be properly designed and trained through use of assumptions of the *a priori* probabilities of target categories (as well as those of target versus clutter blobs). It also must make assumptions about operating conditions (OCs). This is not to say that the designers of the ATR, or testing

Table 1.2 Example of a confusion matrix (APC is armored personnel carrier).

		Reported by ATR		
		Tank	**Truck**	**APC**
Truth (actually, expert opinion)	Tank	*a*	*b*	*c*
	Truck	*d*	*e*	*f*
	APC	*g*	*h*	*i*

organization, actually think through these assumptions. Best performance is achieved when these assumptions match the actualities of the test data. A fair test is one in which all parties being tested have equal knowledge of the proportion of targets of each category in the test set, as well as OCs covered by the test data. In an operational setting, there will be intelligence information on enemy forces, so it is not unreasonable to provide such information for pre-production ATR testing. An operational IR ATR will, for example, know if it is day or night. A typical unfair test is one in which the provided training data is for night and the test data is for day. The equation given for P_c is implicitly making the assumption that the sum of the entries in the rows of the confusion matrix are the priors for the populations of interest. Assumptions about the priors can only be avoided by not aggregating the elements of the confusion matrix; however, this still doesn't solve the problem of assumptions about priors used in training the ATR. Prior assumptions used in training and scoring the ATR should be reported along with test results.

1.5.2.1 Compound confusion matrix

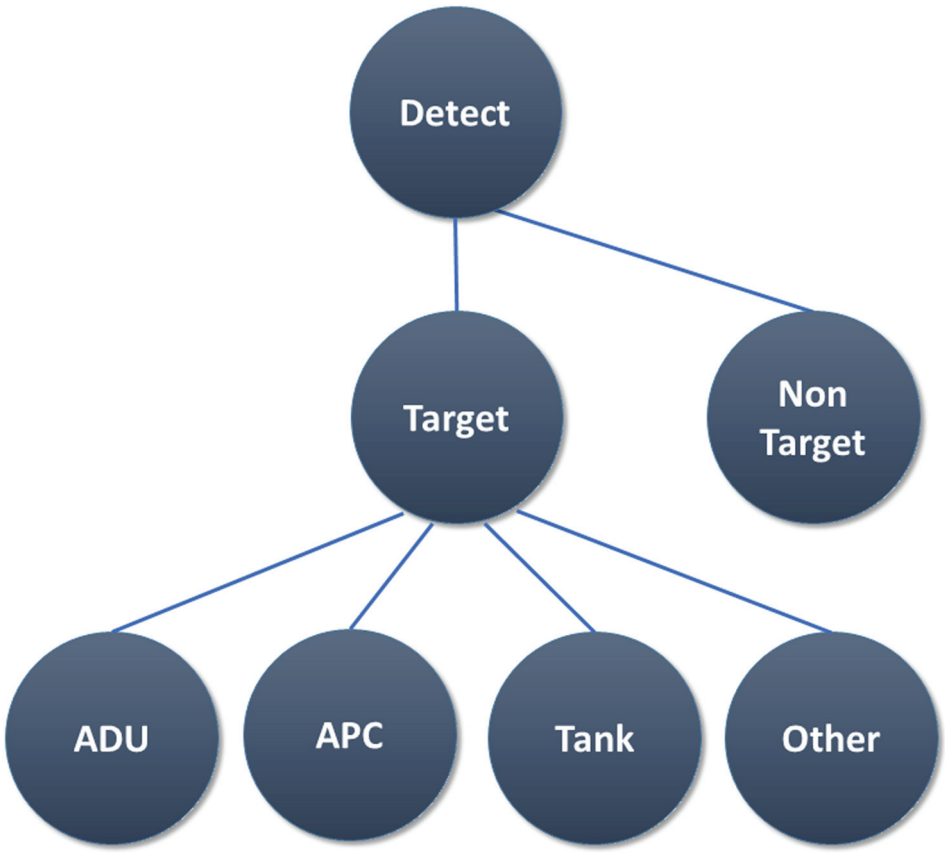
A compound confusion matrix reports results for more than one level of a decision tree. Consider the example shown in Fig. 1.15. The form of this confusion matrix indicates that the ATR's back-end classifier stage is to be tested on target and clutter ROIs. Several performance results are given based on the cells of the confusion matrix.

Decision trees and confusion matrices can be quite complex, as illustrated in the multilevel example given in Fig. 1.16.

1.5.3 Some commonly used terms from probability and statistics

Let us review some terms and introduce a few others. Suppose that the ATR declares a detected target to be a tank with a score of 0.8. This score can be considered a probability estimate with certain restrictions. The sum of all possible outcomes at a given level of specificity, called the sample space of the experiment, must be 1.0.

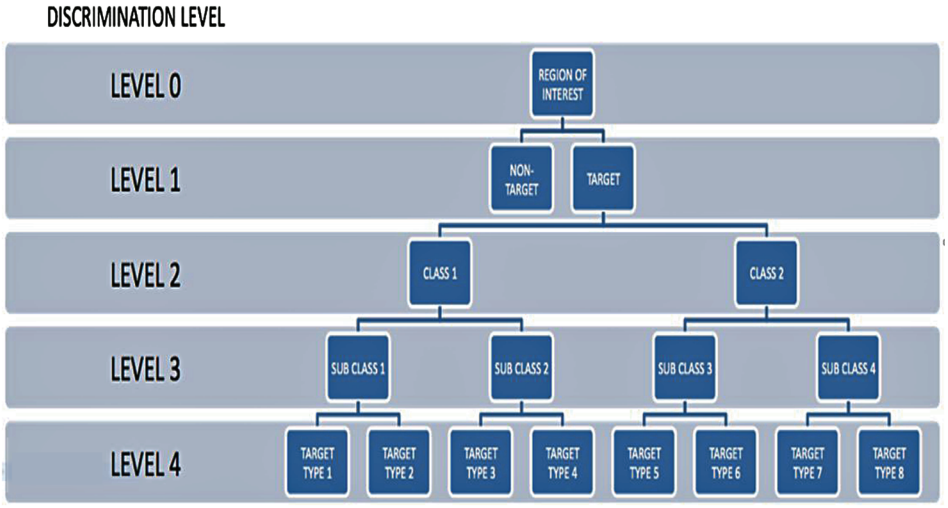
At the Target Class level of the decision tree, the ATR output is more specifically an *a posteriori* probability estimate vector, where each element of the vector corresponds to a target class. But, sometimes only the maximum



			Reported by ATR					
			Detect Accept					Detect Reject
			Class Accept				Class Reject	
			ADU	APC	Tank	Other		
Truth	Class	ADU	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}
		APC	s_{21}	s_{22}	s_{23}	s_{24}	s_{25}	s_{26}
		Tank	s_{31}	s_{32}	s_{33}	s_{34}	s_{35}	s_{36}
		Other	s_{41}	s_{42}	s_{43}	s_{44}	s_{45}	s_{46}
	Non Tgt	Clutter	s_{51}	s_{52}	s_{53}	s_{54}	s_{55}	s_{56}

Figure 1.15 Decision tree (top) and corresponding compound confusion matrix (bottom).

element of the vector is reported along with class label. These probability estimates are often based on the assumption that the *a priori* probabilities of all allowable target classes are equal and that the training data is in some sense



ATR Decision

Truth					TARGETS								Non Target
					CLASS 1				CLASS 2				
					SUB-CLASS 1		SUB-CLASS 2		SUB-CLASS 3		SUB-CLASS 4		
					Target 1	Target 2	Target 3	Target 4	Target 5	Target 6	Target 7	Target 8	
	Targets	CLASS 1	SUB-CLASS 1	Target 1									
				Target 2									
			SUB-CLASS 2	Target 3									
				Target 4									
		CLASS 2	SUB-CLASS 3	Target 5									
				Target 6									
			SUB-CLASS 4	Target 7									
				Target 8									
	Non Target												

Figure 1.16 Multilevel decision tree (top) and corresponding confusion matrix (bottom).

representative of the test data. For results to be statistically justifiable, both training data and test data must be random samples drawn from the same population. If the training data is not representative of the test data, it is not clear what should be expected of the ATR.

***a priori* probability**: The probability estimate prior to receiving new information (e.g., image data). The set of *a priori* probabilities are the priors.

For example, the *a priori* probability of database target classes {tank, truck, APC} would each be 0.333 under the nominal assumption.

The term *confidence* is often applied to an ATR score in a colloquial manner. Better definitions follow.

Confidence: The probability, based on a set of measurements, that the actual value of an event (e.g., target score) is greater than the computed and reported value.

For example, there may be 50% confidence that the actual score of the T-72 is greater than the reported score of 0.8.

Confidence interval: The probability, based on a set of measurements, that the actual value of an event (e.g., target score) resides within a specified interval.

For example, the probability could be 80% that the actual T-72 score falls between 0.7 and 0.9. Note that confidence bounds contradict the interpretation of the ATR score as a probability since a probability must lie within the bounds of [0.0, 1.0] and confidence bounds are typically not so restricted.

Confidence bounds could also be placed around ATR performance results as a whole. For example, the probability could be 80% that the P_{ID} performance of the ATR falls within the bounds of (0.7, 0.9). Although confidence bounds are sometimes provided for ATR performance results, the required rigorous statistical model justifying them is generally lacking. Furthermore, such bounds would only apply to a carefully designed closed experiment, not the infinitely varying real world.

Various competing theories provide ways of measuring confidence. Each of these approaches has major proponents as well as its own terminology and equations. For example, opinions can be said to come in degrees, called degrees of belief or credences.

Degree of belief [bel(A)]: The degree of belief, given to event A, is the sum of all probability masses that support the event A without supporting another event.

Degree of plausibility [pl(A)]: The degree of plausibility quantifies the total amount of belief that might support an event A. The plausibility is the degree of support that could be attributed to A but can also support another event.

Knowledge imprecision of probability estimate: The two quantities bel(A) and pl(A) are often interpreted as a lower and upper bound of an unknown probability measure P on A. The difference pl(A) – bel(A) is an indicator of the degree of knowledge imprecision on $P(A)$.

Pignistic (Latin for betting) probability: Pignistic probability Bet P is the quantification of a set of beliefs into final form for decision making. The value of a pignistic probability falls between that of a belief and a plausibility.

1.6 Experimental Design

An *experimental design* is a blueprint of the procedure used to test the ATR and reach valid conclusions. A good experimental design is critical for

understanding the performance of the ATR under all conditions likely to be encountered. The experimental design is *internally valid* if it results in a fair test with no extraneous factors. The experiment is *externally valid* if it is sufficiently well designed for test results to generalize to operational conditions.

Factors jeopardizing **internal validity** of competitive ATR tests include:

- unequal access to training data;
- unequal information on, or ability to negotiate or alter, the experimental design;
- unequal access to *a priori* probabilities for target types or operating conditions;
- unequal access to test data, data very similar to test data, or number of times that the test is taken on the same data; and
- blind test data that is not 100% unseen by some organizations taking the test.

From a scientific viewpoint, it would be nice for all competitive ATR tests to be fair tests. From a business perspective, participants in the test might each seek an advantage.

Internal validity is at the center of all cause–effect inferences that can be drawn from a test. If it is important to determine whether some operating condition (OC) causes some outcome, then it is important to have strong internal validity. Essentially, the objective is to assess the proposition if X, then Y. For example, if the ATR is given data for operating condition X, then the outcome Y occurs.

However, it is not sufficient to show that when the ATR is tested with data from a certain OC, a particular outcome occurs. There may be many other reasons, other than the OC, for the observed outcome. To in fact show that there is a causal relationship, two propositions must be addressed: (1) if X, then Y and (2) if *not* X, then *not* Y.

As an example, X may refer to daytime images and *not* X to night images. Evidence for both of these propositions helps to isolate the cause from all of the other potential causes of the outcome. The conclusion might be that when solar radiation is present, the outcome Y occurs, and when it is not present, the outcome Y doesn't occur. This may be just a first step. To better understand the effect, the ATR could be tested on OC bins for each hour of the diurnal cycle with and without cloud cover.

Factors adversely affecting **external validity** or generalizability of ATR tests include:

- test data that is not representative of true operational sensor data (e.g., not considering dead bugs on window in front of IR sensor, not considering platform vibration or motion);
- very limited set of OCs in training and testing data sets;
- unreasonable hardware size, weight, power, or cost requirements;

- ignoring possibilities of enemy changing tactics, countermeasures, and decoys;
- ancillary information (metadata) provided for test not matching that (in type, update rate, or accuracy) which would be available from a relevant operational system.

If the test set represents the intended mission and scenario along all degrees of freedom (such as range, target type, aspect angles, weather conditions, clutter environment, platform, and target motion), then the performances measured on sufficiently large and representative OC bins, when properly interpreted, will point toward expected mission performance.

1.6.1 Test plan

A test plan provides a tangible description of the experimental design so that an entire Integrated Product Team (government, industry, and occasionally university) can work toward the same goals. A test plan should ideally be carefully developed and agreed on by all stakeholders. It can cover such items as how the test site will be restored after tanks tear up the ground, how and by whom data will be sequestered, and aircraft airworthiness certification. A test plan can be 100 pages long. A good test plan is critical to the success and smooth operation of a field test. The test plan should anticipate equipment failure and suggest workarounds. The author has seen tests where dozens of participants from different organizations had to be sent home because one piece of equipment was broken. The following components make up the test plan:

1. Specification of the product to be tested.
2. Scope of the testing.
3. Safety issues as well as issues concerning security, privacy, ethics, environmental, etc.
4. Test lead/manager and team member responsibilities.
5. Entry and exit criteria.
6. Descriptions of items and features to be tested, and list of items and features not to be tested.
7. Applicable requirements and requirements traceability, and key performance parameters (KPPs).
8. Test procedures and guidelines.
9. Test schedule. (Note: Planning a test of soldiers engaged in specified activities requires almost the precision of choreographing a ballet.)
10. Responsibilities for supplying test resources. Staffing and training needs.
11. Measurement & Analysis: Guidelines for in-progress and post-test analysis and reporting. Pass/fail criteria. Contingency for retest.

Safety concerns are considerable when working around heavy machinery such as helicopters and tanks. Testing with live munitions is obviously more

dangerous. Testing is often performed in dangerous locations such as fire-prone California in the summer, tire-wrecking Yuma desert, or Alaska in the winter.

1.6.2 ATR and human subject testing

It may be desirable to compare ATR performance with that of human test subjects. Human subjects can also serve as test targets. The first consideration when using human subjects is whether approval is required from an independent Institutional Review Board (IRB). An IRB is designated to approve, monitor, and review research involving humans. Its function is to protect the rights and welfare of the research subjects. Government contracts often require IRBs when human testing is involved. Human testing can be as innocuous as test subjects looking at a monitor and pointing to targets, with performance recorded. It may be as serious as conditions involving live munitions or active sensors. Whether an IRB is required depends on what is written in a contract, the precise nature of the testing, the relationship between the test subjects and the organization doing the testing, and where the tests take place. Rules about the need for an IRB keep changing.

The IRB must have at least five members. The members must have enough experience, expertise, and diversity to make an informed decision as to whether the research is ethical, informed consent is sufficient, and appropriate safeguards are in place.

IRBs appraise research protocols and related materials (e.g., informed consent documents). The chief objectives of an IRB protocol review is to assess the ethics of the research and its methods, to promote fully informed and voluntary participation by prospective subjects, and to maximize the safety of subjects.

Contracting an IRB, submitting the protocols and other paperwork, and obtaining approval can take up to one year. Keep this in mind when planning a test involving human subjects! If a test is to be conducted at different locations, say different military bases, a separate IRB might be required for each location. A project involving government, industry, and university can potentially have several IRBs, with an agreement needed for one IRB to take the lead.

Some issues that can arise in ATR-related human subject testing include:

- Safety: Will individuals serving in the role of targets be around live munitions, experimental aircraft, hazardous material, or maneuvering vehicles? Is an active sensor used, such as a laser, LADAR, radar, or terahertz camera?
- Privacy: Are faces discernable in a database? Could pictures of minors be inadvertently captured in a data collection? Can the sensor see through clothes? Are the privacy laws of the state in which the test takes place being followed?

- Coercion: Are the test subjects competing against the ATR? Are the test subjects being coerced in the direction of particular results?
- Monetary reward: Are the test subjects, for example, being given a trip to Hawaii to take part in a field test?

The experimental design of a test of human subjects must consider different items compared to a test of an ATR. For example:

- What length of time is the test subject given to make and record a decision?
- What are the many display issues involving type of display: display size, distance from the display, room lighting, ability of test subject to adjust display parameters, etc.?
- Is fatigue an issue?
- Are there noise and other distractions?
- Is the test setup realistic, for example, using a flight simulator or ground station, or less realistic such as using the test subject's own computer?

A typical ATR's software forgets the last set of data that it has been tested on unless programmed otherwise. A human test subject can't help but learn the particulars of a data set during a test and, hence, perform better on future data with similar characteristics.

1.7 Characterizations of ATR Hardware/Software

Several key performance parameters of an ATR hardware/software system are as follows:

- size
- weight
- power requirement
- latency
- cost
- mean time between failure
- security level
- number of source lines of code (SLOC count).

The main cost in ATR system development is often the cost of collecting sufficiently comprehensive training and testing data sets. Air-to-ground data collection costs are measured by the flight hour. Costs include fuel, air crew, and equipment. Setting up an array of ground vehicles might include renting the military test site, renting foreign military vehicles, paying for drivers of the vehicles, and keeping a ground crew in place. Likewise, undersea data collections and testing involve considerable expenses. An ATR is procured and deployed only if there is a budget item for doing so. Once the ATR is deployed, there are costs associated with the logistics trail. If the ATR uses the

same electronic boards that are used throughout a platform or series of platforms, the logistics cost is low. Another consideration is the number of years that the chips and other components of the ATR will be available. Some chip and board manufacturers guarantee seven years of availability. If one compares that to the longevity of a platform such as the B-52 bomber, that doesn't seem like a long time. It is common to buy and warehouse all potentially needed spare parts in advance. Keeping counterfeit parts, used parts, and lower-grade parts out of a system are critical issues. Another issue is the maintenance cost of the ATR as target sets, rules of engagement, sensor inputs or computer operating systems keep changing. After the ATR is delivered, who is going to train it on new target types, and where is the budget to do this? Software has to be kept under configuration control. Supposing that the ATR software reports errors, there needs to be a maintenance plan to address these errors as well as an upgrade plan. How will software changes be made when the operating system is no longer supported or when the original development team is disbanded? Many of these issues are not unique to ATR and are well understood by the government and defense contractors.

References

1. T. D. Ross, L. A. Westerkamp, R. L. Dilsavor, and J. C. Mossing, "Performance measures for summarizing confusion matrices: the AFRL COMPASE approach," *Proc. SPIE* **4727**, 310–321 (2002) [doi: 10.1117/12.478692].
2. S. G. Alsing, "The Evaluation of Competing Classifiers," Ph.D. dissertation, Air Force Institute of Technology, Wright-Patterson AFB, Ohio (2000).
3. C. B. Bassham, "Automatic Target Recognition Classification System Evaluation Methodology," Ph.D. dissertation, Air Force Institute of Technology, Wright-Patterson AFB, Ohio (2002).
4. A. Mahalanobis, S. R. F. Sims, and A. Van Nevel, "Signal-to-clutter measure for measuring automatic target recognition performance using complimentary eigenvalue distribution analysis," *Opt. Eng.* **42**(4), 1144–1151 (2003) [doi: 10.1117/1.1556012].
5. K. Fukunaga and W. L. G. Koontz, "Representation of random processes using the finite Karhunen–Loeve transform," *J. Opt. Soc. Am.* **72**(5), 556–564 (1982).
6. M. Self, B. Miller, and D. Dixon, "Acquisition Level Definitions and Observables for Human Targets, Urban Operations, and the Global War on Terrorism," Technical Report No: AMSRD-CER-NV-TR-235, RDECOM CERDEC (2005).
7. A. L. Magnus and M. E. Oxley, "Theory of confusion," *Proc. SPIE* **4479**, 105–116 (2001) [doi: 10.1117/12.448337].