Adaptive Automatic Ground Truth Generation for Testing of Vehicle Detectors

C. Arthur MacCarley
Electrical Engineering Department
California Polytechnic State University
San Luis Obispo, California 93407 USA
Ph: 805 781-8461  FAX: 805 781-0768  amaccarl@calpoly.edu

Abstract. Field testing of vehicle presence detectors using non-aggregated data and metrics requires the generation of an accurate ground-truth record for each individual vehicle present at a test site. Performance results for individual detectors are assessed by comparison with the ground truth record. Ground truth has traditionally been determined either by use of a trusted reference detector, or by human observation, usually using time-coded video recorded images of the traffic during a test period. For meaningful multiple detector testing, the volume of required observations becomes unwieldy, defying accurate human verification. Further, no detector can truly be considered “trusted” for use as an absolute ground truth reference – all detectors have some limitations or potential sources of error. This paper describes an algorithm by which a weighted consensus of all detectors under test is used to generate a reliable ground truth record against which all detectors may be automatically compared for accuracy. The method employs a discrete adaptive (learning) algorithm that continuously adjusts a 'reliability coefficient' for each detector under test over many samples, assuring the most accurate possible consensus result. Limitations and convergence issues are discussed. The algorithm is currently in experimental use in the California Department of Transportation I-405 Detector Test Bed, in which multiple detectors of various technologies are being tested concurrently over six lanes of high-volume traffic.

BACKGROUND

Vehicle presence sensors are commonly used for traffic volume measurement and control of signalized intersections and ramp meters. In addition, vehicle speed and classification from length are important for automated incident detection, prediction of traffic demand, and assessment of real-time indications of traffic safety (1). In recent years, vehicle detection technologies have greatly expanded beyond traditional methods such as inductive loops, now including sensing methods based upon video, phased-array RADAR, ultrasonic, laser, and acoustic phenomena (2). Inherent to all sensing methods are some operational limitations and sources of error. The accuracy and effectiveness of roadway detection technologies has become an important concern in effective roadway monitoring, management, safety, and planning.

The accuracy of vehicle presence detectors has been traditionally assessed using aggregated metrics such as cumulative vehicle count over a fixed interval. Field-aggregated data reporting has largely been driven by computational and communications bandwidth limitations present at the time of implementation of much of the roadway management infrastructure, and has been considered adequate for basic traffic monitoring (3). But much more information can be gleaned from detector data if individual vehicle records are recorded and processed, both in real time and in post-processing. Further, if field-aggregated metrics such as vehicle count or average occupancy are used to evaluate the performance of a detector, test results are intrinsically unreliable since the actual accuracy of the detector is obscured in the aggregation process. For example, failures to detect and false detections are self-canceling errors. Current computational and communications technologies allow the comprehensive recording of individual detection events and the evaluation of detectors based upon non-aggregated performance metrics: correct detections, failures to detect, and false detections which properly reflect the actual performance of the detector. However, field testing using non-aggregated data and metrics requires the generation of an accurate ground-truth record accounting for each individual vehicle at a test site. Performance results for individual detectors are then assessed by comparison with the ground truth record.

Ground truth has traditionally been determined either by the use of a trusted reference detector, or by human observation, usually using time-coded recorded video of the traffic during a test period (4). For meaningful multiple detector testing, the volume of required observations can become unwieldy, defying accurate human verification. Further, no detector can truly be considered “trusted” for use as an absolute ground truth reference; all detectors have some limitations or potential sources of error.
The California Department of Transportation established the Advanced Traffic Management Systems Detector Testbed on Route 405 near the City of Irvine, California, to perform multiple concurrent detector evaluations under high-volume traffic conditions. Individual vehicle records are recorded, and non-aggregated performance metrics and statistics are generated in post-processing by the Video Vehicle Detector Verification System (V2DVS) (5). The V2DVS system records an event record, including a digital image acquired at the time of detection, for each vehicle detected by every detector under test. It also includes computer vision capabilities to provide a reference image and additional record of every vehicle in a standard detection zone. A location view of the ATMS Detector Test Bed is shown in FIGURE 1.

![FIGURE 1 Caltrans ATMS Detector Test Bed with V2DVS.](image)

Ground truth is generated automatically based upon a weighted consensus (voting) of all detectors under test, including the video detection record generated by the system itself. A means is provided for rapid human resolution of any reported detections for which a clear consensus does not exist. Finally, it compares the results of individual detectors to the ground truth record to generate an overall report of the performance of each detector under test. Without such a labor-saving system, 100% data verification is not considered practical, considering the large traffic volume (averaging 6000 vehicles per hour across six lanes) and as many as ten detectors tested concurrently. At maximum traffic capacity, as many as 96,000 records per hour may be generated, not including false detection records. The discussion below focuses on the automated data reduction and verification method, and the video processing techniques used for robust detection and measurement of the speed of each vehicle.

**TEST METHODOLOGY**

For each actual vehicle present at the test site, the ground truth algorithm attempts to identify records of the same vehicle reported by different detectors, often having different detection zones and processing delays. The ability of the system to correctly correlate detections of the same vehicle occurring at different positions on the roadway is critical. Adjusted detection times are calculated from the distance (offset) of each zone from a baseline position, using the assumption of constant vehicle speed over the offset interval. Vehicle speeds are estimated by the video processing ability of the VTDS, and when available, also calculated using duplex loop detectors. When both sources are available, the values are averaged. The time adjustment for vehicle \( j \) reported by detector \( i \) is:

\[
 t_{i,j} = \frac{x_{i,j}}{v_{i,j}}
\]
where:  
$t_{i,j}$ is the pre-signal delay for the detection (sec.)
$x_i$ is the offset of the detection zone from the baseline (m.)
$v_{i,j}$ is the velocity of the vehicle (m/s)

Raw detection times for detection zones different than the baseline position are corrected by subtraction of the pre-signal delay to generate a compensated time of detection in a virtual common detection zone. A false detection may not have a corresponding speed measurement if it is not detected by either V2DVS or the duplex loop detector. In this case, the speed of the nearest proximate vehicle in that lane is used. After adjustment of the time of the detection, a reported detection is considered valid if it occurs within a user-defined admissible time/distance aperture in the virtual detection zone, centered about the weighted consensus of the detection times reported by all detectors. A 20 ft. (6.2m.) admissible detection window is illustrated as an overlay above a typical vehicle image in FIGURE 2.

![FIGURE 2 Typical detection image with adjusted admissible detection window overlay.](image)

The failure of a detector to report a vehicle within the aperture is considered a potential failure-to-detect error, unless a proximate detection is later associated with that grouping during manual verification. Detections occurring outside of an aperture, or multiple detections inside the same non-overlapping aperture, are considered potential false detections. Proximity to the consensus-derived center time is used to discriminate cases of nearly equal admissibility when apertures for closely following vehicles overlap.

A one-second aperture is illustrated in FIGURE 3 by lines overlaid on the manual verification window provided by the V2DVS application. This window graphically depicts on a common time line the results from all detectors for a given lane. Automatically correlated detections appear as blue dots, and false detections in red. One ambiguous detection appears as a yellow dot. Mouse-click selection of any dot brings up the corresponding image acquired at the time of that detection, that makes clear why a detection may not have been automatically correlated (grouped). The members of the currently selected group are highlighted in green.
Once manual verification has been completed for all lanes, the dataset is marked by the operator as closed, and V2DVS compiles test results for all detectors under test, including itself. A representative “Statistics” window is shown in TABLE 1.
TABLE 1. Sample V2VDS Report on Comparative Detector Performance

<table>
<thead>
<tr>
<th>Detector</th>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Total</th>
<th>Percent of Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUT 1</td>
<td>Correct</td>
<td>925</td>
<td>371</td>
<td>878</td>
<td>871</td>
<td>769</td>
<td>2</td>
<td>3,816</td>
</tr>
<tr>
<td></td>
<td>Fail</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Detector</th>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Total</th>
<th>Percent of Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUT 2</td>
<td>Correct</td>
<td>766</td>
<td>358</td>
<td>700</td>
<td>627</td>
<td>630</td>
<td>3</td>
<td>3,084</td>
</tr>
<tr>
<td></td>
<td>Fail</td>
<td>159</td>
<td>13</td>
<td>178</td>
<td>244</td>
<td>139</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Detector</th>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Total</th>
<th>Percent of Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUT 3</td>
<td>Correct</td>
<td>824</td>
<td>3</td>
<td>575</td>
<td>661</td>
<td>619</td>
<td>1</td>
<td>2,683</td>
</tr>
<tr>
<td></td>
<td>Fail</td>
<td>101</td>
<td>368</td>
<td>303</td>
<td>210</td>
<td>150</td>
<td>2</td>
<td>1,134</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td>132</td>
<td>0</td>
<td>216</td>
<td>363</td>
<td>425</td>
<td>1</td>
<td>1,137</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Detector</th>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Total</th>
<th>Percent of Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUT 4</td>
<td>Correct</td>
<td>905</td>
<td>353</td>
<td>823</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2,081</td>
</tr>
<tr>
<td></td>
<td>Fail</td>
<td>20</td>
<td>18</td>
<td>55</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td>1,426</td>
<td>3</td>
<td>839</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2,268</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Detector</th>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Total</th>
<th>Percent of Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUT 5</td>
<td>Correct</td>
<td>767</td>
<td>245</td>
<td>800</td>
<td>774</td>
<td>653</td>
<td>3</td>
<td>3,242</td>
</tr>
<tr>
<td></td>
<td>Fail</td>
<td>158</td>
<td>126</td>
<td>78</td>
<td>97</td>
<td>116</td>
<td>0</td>
<td>575</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td>107</td>
<td>2</td>
<td>58</td>
<td>25</td>
<td>7</td>
<td>2</td>
<td>201</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Detector</th>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Total</th>
<th>Percent of Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth</td>
<td>Correct</td>
<td>925</td>
<td>371</td>
<td>878</td>
<td>871</td>
<td>769</td>
<td>3</td>
<td>3,817</td>
</tr>
</tbody>
</table>

*DUT = DETECTOR UNDER TEST.
ADAPTATIVE COEFFICIENT OPTIMIZATION

Adaptive sensing based upon a consensus of multiple independent observations has been utilized in many applications to achieve the most reliable possible estimate of the measurement sought, for example (6). The initial (TRB07) ground truth consensus process was based upon fixed “reliability coefficients” associated with each detector under test, with detectors known to be more accurate affecting larger influence on the consensus results. In further work reported herein, an adaptive method has been developed which automatically adjusts these coefficients based upon the cumulative reliability of each contributing detector. Reliability, in this case, is continuously assessed by the algorithm based upon agreement with the consensus.

Coefficient adaptation is performed with any detection event, correct or incorrect, the result of a reported detection by any detector under test:

\( i = \text{detector index, } i = 1,2,\ldots,n \) Typically, \( n=8 \)

\( k = \text{event (possible vehicle) index } k = 1,2,\ldots,m \) A detection by any of the \( i \) detectors can cause a detection event. Typically, a dataset of an hour would yield \( m=2000 \) for a centrally located traffic lane.

\( d_i(k) = [0,1] \) binary conclusion of detector \( i \) for any detection event \( k \). 0 means that detector \( i \) did not report a detection in the time or distance window surrounding this event. 1 means that detector \( i \) reported a detection in the time or distance window surrounding this event.

\( c_i(k) = 1 \) if detector \( i \) agrees with the consensus conclusion for event \( k \) , 0 if it disagrees. Thus a correct detection (agreement with the consensus) would yield \( c_i = 1 \) while either a failure to detect of a false detection would yield \( c_i = 0 \). Thus \( c_i(k) = d_i(k) \oplus G(k) \) where \( \oplus \) is the logical exclusive OR.

\( g(k) = \text{ground truth analog conclusion for } i^{th} \text{ detection window (possible vehicle)} \)

\( G(k) = [0,1] \) binary conclusion for \( k^{th} \) detection event (possible vehicle)

The consensus conclusion for event \( k \) is found by thresholding a weighted average

\[
g(k) = \frac{\sum_{i=1}^{n} a_i d_i}{\sum_{i=1}^{n} d_i}, \quad a_i = \text{weighting coefficient for detector } i
\]

\[
G(k) = \begin{cases} 
0 & g(k) < \gamma_{lower} \\
\beta & \gamma_{lower} \leq g(k) < \gamma_{upper} \quad \text{where typically } \gamma_{lower} = 0.4, \gamma_{upper} = 0.6 \\
1 & g(k) \geq \gamma_{upper}
\end{cases}
\]

“0” means that the consensus for that event was that no vehicle was actually present.
“1” means that the consensus for that event was that a vehicle was actually present
“\( \beta \)” implies indeterminate, which means that the consensus for that event was not sufficiently certain to establish a definitive ground truth conclusion. \( \alpha \) remains a variable until the system user manually verifies the validity of the detection by inspection of the images acquired at the time of detection of each detector. This is referred to as “manual verification” for indeterminate detection events. If the detection is manually verified to be false, \( \beta = 0 \). If the detection is manually verified to be true, \( \beta = 1 \).

A first order autoregressive filter is used to gradually adapt the weighting coefficient for the \( i^{th} \) detector based upon its agreement or disagreement with the consensus for each event \( k \). The weighting coefficient \( a_i(k) \) is slightly modified with each detection event.

\[
a_i(k+1) = a_i(k)(\alpha) + (1-\alpha)c_i(k) \quad \text{where } \alpha = \text{typically 0.95} \quad k+1 \text{ is the next event after event } k.
\]
CONVERGENCE CONSIDERATIONS AND OBSERVATIONS

Convergence of the algorithm is achieved when the reliability coefficients stabilize about mean values that tend to vary minimally and equally in a positive or negative sense. Note that if every detection is correct (agrees with the consensus), \( a_i(k) \) asymptotically approaches unity, while if the detector consistently fails to detect or falsely detects \( a_i(k) \) asymptotically approaches zero. Consequently, the contribution to the consensus asymptotically approaches zero for consistently inaccurate detectors, as the algorithm continuously reduces them with every successive disagreement with the consensus. Reliance upon consistently accurate detectors increases with every agreement with the consensus to which they increasingly contribute.

As with most iterative optimization strategies, the terminal values of the weighting coefficients (reliability coefficients) are susceptible to seeking turning points which could represent either local or global maximum accuracy, or maximum inaccuracy, dependent upon the initial values of the coefficients. The relatively simple adaptive approach described above assumes that the detectors assigned higher initial values (assumed higher reliability) first, yield a majority contribution to the weighted consensus, and second, are at least initially correct most of the time. Failing this assumption admits the possibility that the initial consensus is incorrect. In such a case, detectors disagreeing with the incorrect conclusion will be de-emphasized with each successive detection event. The convergence behavior of the algorithm is therefore potentially unstable – a characteristic of consensus-seeking adaptive strategies. A reasonable prior assumption of relative detector reliability is therefore important to achieving the most accurate ground truth data set. These assumptions are embedded in the form of initial coefficient values for each detector under test. By default, all detectors start with coefficients equal to 0.5.

In practice, we have observed that detectors exhibiting higher error rates (failures to detect or false detections) also are the least consistent, e.g., they tend to be incorrect as often as correct. However, detectors that tend to be more accurate in a given test scenario also tend to be the most consistent, consistently affecting the consensus result in a way that favors the correct result for each detection event. Therefore, in cases where all reliability coefficients have been nominally set to the default value, the net effect observed is an increase in the time to convergence of the coefficients, but not a failure to ultimately converge to optimal values that maximize the accuracy of the ground truth data set. Thus the algorithm appears to be robust, but this is difficult to quantify without restrictive assumptions about the traffic data and detectors under test.

Convergence speed is function of the value of the adaptive modification parameter \( \alpha \). The suggested value above (0.95) is probably not optimal for fastest convergence, but has been found to be acceptable for avoiding internal instability of the algorithm. In most trial cases, not yet well quantified, an acceptable level of convergence is achieved after between 50 and 100 consensus-verified vehicle detection events. Since the optimization process can be performed entirely in post processing of existing datasets, it is possible to do a first “pass” on the dataset to optimize the coefficients, then a second pass to generate the actual ground truth record. For longer data sets, adaptation of the coefficients can and should continue during the processing run to compensate for the effects of changing illumination, traffic density and other test conditions on the relative accuracy of the detectors under test.

In preliminary testing with eight detectors under test and 1600 actual vehicles, after coefficient convergence approximately 99% of vehicle detections were properly automatically classified as correct, false, or failure-to-detect by the automated process, leaving only 1% for manual verification. This may be compared with the prior reported result (TRB07) of 97% achieved using fixed coefficients based upon “best guesses”. The same data set was used in both cases, and all other factors and parameters (such as the size of the admissible time/distance aperture) remained the same. For the eight detectors under test, no coefficient stabilized uniquely at either 0 or 1, since even the least accurate detector occasionally is correct, and even the most accurate detector occasionally makes an error. Obviously, better “best guesses” for fixed detector coefficients could have achieved the same results, but an adaptive algorithm of this type probably represents the only quantitative means to determine these.

More sophisticated criteria than simple agreement with consensus and first order autoregressive adaptation are certainly possible. To date, variations of the adaptive update algorithm have not yielded noticeable improvements either in accuracy of the ground truth data set or the convergence time. Convergence characteristics and alternative adaptive algorithms remain under study at this time in the test bed.

ACKNOWLEDGEMENTS AND DISCLAIMER

This work was supported by the California Department of Transportation, Division of Innovation and Research. Statements and results reported herein are the responsibility of the authors exclusively. This report does not constitute a standard, specification or regulation.
REFERENCES


