A Bayesian Framework for the Assessment of Vision-based Weed and Fruit Detection and Classification Algorithms

Tristan Perez*, Inkyu Sa*, Christopher McCool* and Christopher Lehnert*

Abstract—This paper proposes new metrics and a performance-assessment framework for vision-based weed and fruit detection and classification algorithms. In order to compare algorithms, and make a decision on which one to use for a particular application, it is necessary to take into account that the performance obtained in a series of tests is subject to uncertainty. Such characterisation of uncertainty seems not to be captured by the performance metrics currently reported in the literature. Therefore, we pose the problem as a general problem of scientific inference, which arises out of incomplete information, and propose as a metric of performance the (posterior) predictive probabilities that the algorithms will provide a correct outcome for target and background detection. We detail the framework through which these predicted probabilities can be obtained, which is Bayesian in nature. As an illustration example, we apply the framework to the assessment of performance of four algorithms that could potentially be used in the detection of capsicums (peppers).

I. INTRODUCTION

Robotic vision algorithms can be used for detecting and classifying weed species as well as fruit and fruit quality grading in agricultural applications [1]–[3]. The output of these Detection and Classification Algorithms (DCA) provides information used in a decision problem. In other words, we can think of these algorithms as a perception service provided to a decision agent. For example, in the case of weed management, the information as to what kind of weed has been detected can be used by an autonomous agent to select whether to act on the weed with either a chemical agent or with an alternative method such as thermal or mechanical [4]. In the case of the harvesting of a horticulture crop, the output of a detection and classification algorithm is used as information to decide whether to pick a piece of fruit from a particular location in space with an autonomous system like a robotic arm [3].

A common figure of merit for detection and classification algorithms is the area under the precision-recall curve [5],

\[ AUC = \frac{TP}{TP + FP} \]

where

- Precision = \( \frac{TP}{TP + FP} \),
- Recall = \( \frac{TP}{TP + FN} \),

in which \( TP \) is the count of true events that have been identified as true by the algorithm, \( FN \) is the count of true events that have been identified as false by the algorithm, and \( FP \) is the count of false events that have been identified as true by the algorithm (error). This type of curve is obtained by varying the threshold value on the probabilities—internally handled by the algorithms—which is used to provide an outcome. These curves are generally used for tuning DCA by choosing the threshold value for which precision and recall attain the same value [5].

Although the area under the precision-recall curve provides a figure of merit—a larger area under the curve is better—the difference in score is hard to interpret. In addition, there is no characterisation of uncertainty of the results. This paper, thus, focuses on the characterisation of performance based on probabilities, which we take as a description of our uncertainty about a hypothesis related to the performance of the algorithms. Understanding this uncertainty is a key aspect for the process of choosing a particular algorithm and also for making decisions given the uncertainty about its performance or reliability [6]. Hence, in this context, we can consider the proposition or hypothesis

\[ H = \{ \text{The DCA provides the correct information} \} \]

and its complement \( \bar{H} \), and we seek to assess the predicted probability \( p(H|D,B) \), where \( D \) is a proposition that stands for the data related to the outcomes of a sequence of tests made on the particular algorithm, and \( B \) is a proposition that stands for background information.

In our approach, we follow a Bayesian formulation for the testing of the hypothesis \( H \). That is, a hypothesis \( H \) is a proposition that can either be true or false, and \( p(H|D,B) \) describes our state of knowledge about the truth of falsity of the hypothesis based on our background information and the data obtained from the tests. In other words, we make no reference to random phenomena or consider probabilities as a frequency. That is, we treat the problem as a general problem of scientific inference, which arises out of incomplete information rather than randomness—we work under the banner of [7], [8], [9], [10].

II. A PROBABILISTIC MEASURE OF PERFORMANCE

The detection and classification problem of weeds or crops based on robotic vision considers the presence, or lack thereof, of a target object in an image. We can then consider two propositions:

\[ O = \{ \text{Target object is present in the image} \} \]
\[ A = \{ \text{Algorithm accuses the presence of the target object} \} \]

where the target object can be a particular weed species in the problem of weed management or a fruit in the problem of crop identification for harvesting.
Then, we can define two hypotheses or propositions:

\[ H_1 = (A|O), \quad H_2 = (\overline{A}|\overline{O}), \]  

which state that the detection and classification algorithm provides the correct information for target and for background respectively. In relation to standard terminology, [5], the proposition \( H_1 \) is associated with a true positive, and \( H_2 \) with a true negative.

We then propose that the performance of a DCA is characterised by the predictive probability of success in one future test: \( p(H_i|D, B) \) \((i = 1, 2,\ldots)\), where \( D \) is a proposition that stands for the data or outcomes of a sequence of tests made to the algorithm based on images for which it is known whether the target object is present in the image, and \( B \) is a proposition that stands for background information.

If the robustness of a DCA is to be taken into account, we can consider the fact that the data of the test corresponds to different environment conditions \( E_j \) with the associated probabilities \( p(E_j|B) \). The latter describes the probability of encountering \( E_j \) during the nominal operation of the system. For example, \( E_j \) can refer to data collected during the sunny day, overcast day, night, etc. Then, we can consider the following data set:

\[ D = \{d_1, d_2, \ldots, d_N\}, \]  

where

\[ d_k = \begin{cases} 1 & \text{if the DCA provides the correct identification,} \\ 0 & \text{otherwise.} \end{cases} \]

If we are considering \( H_1 \), then correct refers to the identification of the target; if we are considering \( H_2 \), then correct refers to the identification of the background. The data \( D \) thus depends on the hypothesis being tested, but for ease of notation, we will omit the subindex \( i = 1, 2 \).

This scenario could be used to specify a model with a parameter \( \theta_i \) such that at each test we have, independently of anything we know about other tests, a probability \( \theta_i \) for a correct identification; and conversely, a probability \( (1 - \theta_i) \) for a failing. Thus, each test is a Bernoulli trial.

The probability of obtaining the data set \( D \) is

\[ p(D|\theta_i, B) = \prod_{k=1}^{N} P(d_k|\theta_i) = \prod_{k=1}^{N} \theta_i^{d_k}(1 - \theta_i)^{1-d_k}, \]

\[ = \theta_i^R(1 - \theta_i)^{N-R}, \]

where \( R \) is the number of successes in \( N \) tests. Here, we are assuming the independence of the outcomes of the test. This assumption implies that we are not informing the algorithm about its success rate and therefore limits its capability to learn during the trials. Note that \( N \) and \( R \) depend on the hypothesis being tested, but for ease of notation we have omitted the subindex \( i = 1, 2 \).

### B. Inference

The parameter \( \theta_i \), assumed constant, is uncertain. We can describe this uncertainty using a prior distribution for the parameter \( p(\theta_i|E_j, B)^1 \). Note that within a Bayesian approach, a prior distribution for a parameter does not mean that the parameter is random. The parameter is constant, and the distribution describes our uncertainty about its value—what is distributed is the probability not the parameter [9].

Using the data (3), we can update a prior distribution for the parameter to a posterior distribution using the Bayes’s Theorem:

\[ p(\theta_i|D, E_j, B) = \frac{p(D|\theta_i, E_j, B)p(\theta_i|E_j, B)}{p(D|E_j, B)}, \]

where the likelihood function is given by (5), and \( p(D|E_j, B) \) is simply a normalisation constant.

The elicitation of the prior \( p(\theta_i|E_j, B) \) in (6) has been the subject of vehement attacks to the Bayesian framework, for which much of the literature calls it subjective. Rather than a hindrance, this is an advantage of the method, which allows one to incorporate background information. Further, it is the nature of science that when data is analysed starting from different states of knowledge, this could lead to different results. For an eloquent discussion, we refer to the work of [7], [8], [9], [10], and [6].

For the particular problem of interest in this paper, we advocate the use of a uniform prior distribution \( p(\theta_i|E_j, B) \) in (6). The adoption of a uniform prior distribution, or Bayes-Laplace prior, for the parameter \( \theta_i \) reflects our unassuming attitude towards the attainable performance of the algorithm in a particular operational condition—that is, \( \theta_i \) could take any value in the range from 0 to 1. This follows from the Maximum Entropy Principle given that there may be no testable information about the algorithm performance before the test [9], [10]. Further details as to why this a good choice of a prior for the type of experiments being considered in this paper can be found in [11].

\(^1\)We use a standard abuse of notation where \( p(\cdot) \) denotes either a probability of a proposition or a probability density function for a continuous parameter.
If we adopt a uniform prior distribution for \( \theta \), namely,
\[ p(\theta_i | E_j, B) = 1, \quad 0 \leq \theta \leq 1, \tag{7} \]
that is the least committed prior in a maximum entropy sense, then the posterior (6) is a Beta distribution:
\[ p(\theta_i | D, E_j, B) = \frac{(N + 1)!}{R!(N - R)!} \theta_i^R (1 - \theta_i)^{N-R}. \tag{8} \]

**C. Prediction**

From the posterior distribution for the parameter \( \theta_i \), namely (6), the question that arises is *what is the probability of obtaining a certain number of successes in a number of future operations?*

If we knew the true value of \( \theta_i \), then the probability of having \( m \) successes in \( n \) operations can be modelled by the Binomial distribution:
\[ P(m|\theta, n) = \binom{n}{m} \theta^m (1 - \theta)^{n-m}. \tag{9} \]

Through the inference process described in the previous section, we only know the posterior (6) and not the actual value of \( \theta_i \). We could pick a point estimate, and plug it in (9), but this would ignore the uncertainty about \( \theta_i \) that the posterior \( p(\theta_i | D, E_j, B) \) describes, and we would be throwing away information. To make a better use of the information available, the predicted probability of \( m \) given the data \( D_i \) can be computed by marginalisation:
\[ P(m|D, E_j, B) = \int p(m, \theta_i | n, D, E_j, B) \, d\theta_i, \tag{10} \]
\[ = \int P(m|\theta_i) p(\theta_i | n, D, E_j, B) \, d\theta_i. \]

By doing this integration, we take into account the uncertainty about \( \theta_i \).

The metric of performance we seek to use, is the predicted probability of one success in the next test. These can be computed by taking \( m=1 \) and \( n=1 \) in (9), in which case (10) reduces to
\[ P(H_i|D, E_j, B) = \int \theta_i p(\theta_i | D, E_j, B) \, d\theta_i, \tag{11} \]
which is the posterior mean. If we adopt a uniform distribution for the prior \( p(\theta_i | E_j, B) \) in (6), as suggested, then
\[ P(H_i|D, E_j, B) = \frac{R + 1}{N + 2}, \tag{12} \]
where \( R \) and \( N \) depend on the hypothesis being tested. For example if we use this analysis for pixel detection, then there usually is much more data related to the background than to the target in a particular image.

**IV. EXAMPLE - CAPSICUM DETECTION**

We consider the use of four algorithms for detection of capsicum (pepper) crops. These algorithms have been tuned to the particular application, and we seek to test these algorithms with new test data. In particular we consider, algorithms that use different techniques for feature extraction to detect capsicum [12]:

- Algorithm 1 - Combines colour and IR information;
- Algorithm 2 - Histogram of Oriented Gradients (HOG);
- Algorithm 3 - Sparse Auto Encoder (SAE)
- Algorithm 4 - Local Binary Pattern (LBP)

Figure 1 shows, as an example, one of the 10 images available for testing all collected under the same environmental conditions. This figure shows both the colour (left) and NIR (right) data. The test images have been hand labeled at a pixel level for either fruit or background. Figure 2 shows the ground truth and the prediction of Algorithm 1, which combines colour and IR information [12].

We use the ground-truth data of all ten images and the output of the algorithms to compare pixel by pixel. This is used to generate the sequence of Bernoulli trials (3) related to the two hypothesis \( H_i \) being considered. The ten images have been collected for a single environmental condition \( E_j \). This information is then processed to compute the posterior distributions \( p(\theta_i | D, E_j, B) \) for the four algorithms. Figure 3 shows the posterior distributions for \( \theta_1 \) and \( \theta_2 \) for the different algorithms. From these distributions, we can finally compute the sought predicted probabilities \( p(H_i | D, B) \) for the different algorithms using (11); however, since we used uniform priors, we can proceed directly to use (12). Table 1 shows the results.

As we can see from Figure 3 and Table I, the LBP algorithm performs best at detecting the fruit when the fruit is present, but it is less reliable than the combined algorithm at detecting the background—that is, the LBP algorithm gives more false positive target detections. Over all the algorithm that combines the different features provides the best overall performance.

These results characterise performance in simple terms; that is, in terms of predictive probabilities that the algorithms will correctly identify pixels as either background or fruit, based on what we have learned during the testing. These probabilities provide an intuitive figure of measure.

Note that due to the availability of the data we have for the ground truth in this example, we have analysed the performance at a pixel level. Since in the images used there is much more background pixels than capsicum pixels, the posterior distributions \( p(\theta_2 | D, E_j, B) \) are more sharply concentrated about their mean values (mind the scale of the plots in Figure 3).

The same method can be used with lower granularity in the image data, and simply count the number of fruit in the image and then check the number of fruits that the algorithm has correctly identified in order to generate the sequence (3). If this is done, the number data in the sequence (3) will be significantly reduced since it is related to the number of fruit and the number of false positives that the algorithm may accru. In this case, the posterior distributions shown in Figure 3 will be more diffused and asymmetric. Hence, instead of the predictive probability, it may be more convenient to use the full posterior information as a basis of comparison, or compute the high-posterior density 95% interval.
The comparison of the results based on varying granularity outside is the scope of this paper. Table II shows the corresponding area under the precision-recall curve (AUC) performance metric for the different algorithms. As we can see, these results show the same ordinal characteristic of the performance as the Bayesian approach, but the difference in the scores of AUC are harder to interpret than the proposed predictive probabilities.

### Table I
**Predicted Probabilities of Success for Capsicum.**

| \( p(H_1|D, B) \) | Combined | HOG | SAE | LPB |
|----------------|----------|-----|-----|-----|
| 0.741342       | 0.309616 | 0.480995 | 0.780485 |
| 0.987430       | 0.966439 | 0.971216 | 0.959869 |

Fig. 1. Colour and NIR image of capsicums (peppers). These figures show an instance of an experiment scene that is complex and reasonably cluttered. Some capsicums are nicely located at the centre whereas others are highly-occluded by leaves and capsicums.

Fig. 2. Left - ground truth produced by hand labelling each pixel as either fruit or background. Right - output of a particular CDA.

### Table II
**Area under the Precision–Recall Curve (AUC).**

<table>
<thead>
<tr>
<th>AUC</th>
<th>Combined</th>
<th>HOG</th>
<th>SAE</th>
<th>LPB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8120</td>
<td>0.2340</td>
<td>0.3880</td>
<td>0.730</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Posterior probability density function for the parameter \( \theta_1 \) for each DCA considered.

As an example, we consider the assessment of performance of four algorithms that could potentially be used in the detection of capsicums (peppers). We show how we can use pixel-level ground truth images to generate a sequence of pixel-based Bernoulli trials from which we can compute the predicted probabilities that the algorithms will correctly identify pixels as either background or fruit, based on what we have learned during the testing. We compare our results with a standard metric of area under the precision-recall curve. Both metrics indicate the same ordinal performance, but the proposed metric of predictive probabilities offer a much easier interpretation of the scores. This was the motivation for putting forward the probabilistic assessment.

We also discuss the possibility of analysing robustness to changes in the environmental conditions. We show how this can be incorporated into the computation of the predictive probabilities, and discussed how the results can be used either to improve the algorithm developments or to restrict the operation.

### V. Conclusion

Vision-based algorithms for detecting and classifying weed species as well as fruit and fruit quality grading are a key enabling capability for future agricultural applications. In order to make a decision as to which algorithm may be preferable for a particular application, it is necessary to test the performance against ground truth data. In this paper, we pose the testing problem as a general problem of scientific inference, which arises out of incomplete information, and propose as a metric of performance the (posterior) predictive probability of the hypotheses that the algorithms will provide a correct outcome.

### References


