Coarse-to-Fine Search and Rank-Sum Statistics in Object Recognition

Stuart Geman, Kevin Manbeck, Donald E. McClure
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

1995

1 Introduction

Imagine a collection of rigid objects, and imagine that we wish to find all instances of these objects that may appear in a given scene. The presentations of the objects in the image plane may be more-or-less constrained, precluding, for example, variations in scale and orientation, or they may be more-or-less free so that the objects may appear at arbitrary pose. The objects themselves may be three-dimensional, as with vehicles or furniture, or two-dimensional, as with characters or symbols.

This report summarizes a statistical approach, involving a coarse-to-fine search strategy and an imaging model based on rank statistics, that has been successfully applied in some applications to industrial automation. The approach is based on a crude model for the actual appearances of the objects in a given scene, and in fact depends only on object outlines, or silhouettes. Internal detail is ignored. Specifically, a given object at a given location and pose is represented as a set of transition values across pairs of suitably located pixels. A transition is the absolute value of the difference in intensity between two locations. An object/pose pair defines a region and the expectation is that transitions between pixels exterior to the region, but in the vicinity of the region, will tend to be smaller than transitions between pixels that straddle the region boundary. To emphasize this expectation, pixel pairs of the latter type are termed "transition pairs" whereas pixel pairs of the former type are termed "non-transition pairs."

By a "data model," we will mean an object and pose dependent distribution on a collection of transition and non-transition pairs. The data model developed here is designed to be robust to case-by-case variations in the actual signature of the object in a scene. This is achieved by using rank-based statistics, a common tool of nonparametric inference.

In some applications we have had good success by representing objects more simply in terms of raw pixel grey levels, rather than differences of grey levels between pairs of pixels. Pixels are chosen "on target" and "off target" (off target, but in the immediate vicinity of the region defined by a purported object/pose pair). If, for example, the object is lighter than the background, then the on-target pixels should generally have higher values than the off-target pixels. This is analogous to the expectation that transitions pairs will yield larger values than non-transition pairs, an in fact *the same data model developed herein could equally well be applied to collections of on- and off-target pixel values.* Furthermore, uncertainty in the relative contrast of an object (light on dark or dark on light) could also be easily accommodated, as will be evident from the statistical tests proposed below—in fact, this is simply a matter of substituting a two-tailed test for a one-tailed test. In any case, we will restrict our discussion to the analysis of transition and non-transition pairs, remarking only that many more-or-less straight-forward generalizations are possible.

2 Statistical/Computational Framework

The goal is to label each pixel of the image. A pixel may be designated as background, or as any one of the several object types. Under a suitable data model, which will be discussed shortly, the labeling will be the result of an effort to approximate the *maximum likelihood labeling,* constrained in such a way that no two objects overlap. (Of course, this is usually unrealistic. The framework suggested here can be generalized to allow for overlapping views—the issue becomes one of computational feasibility.)

The likelihood function is too complex to be maximized directly. Instead, a decision tree will be constructed that governs a series of hypothesis tests, to be performed at every pixel, the result of which is a list of candidate objects at given positions and poses. The likelihood is maximized over labelings consistent with this candidate list.

2.1 Data Models

The statistical framework rests upon a series of assumptions about the distributions of grey levels among background pixels, and among pixels on and in the vicinity of objects.

2.1.1 Object Models

Let T represent a particular object at a particular image location and pose (scale and orientation). Pairs of "transition" and "non-transition" pixels are chosen so that transition pairs straddle the object boundary, with the connecting line nearly normal to the boundary, and non-transition pairs are in the vicinity of the boundary, but outside of the object. The typical distance between a pair of transition pixels is about the same as the typical distance between a pair of non-transition pixels. The boundary is more or less densely covered with transition pairs, and there are equal numbers of transition and non-transition pairs.

Fixing T, let $A(T)$ be the collection of pixel locations comprising the transition and non-transition pairs. Let Z represent the entire array of pixel grey levels in the image, and let $Z_{A(T)}$ be the components of Z representing the grey levels of pixels in $A(T)$. A model distribution for $Z_{A(T)}$, conditioned on the hypothesis represented by T, will now be developed.

Let x_i, $i \in \{1, ...n\}$ be the absolute difference in intensities of the two pixels defining the $i'th$ transition pair, and let y_i, $i \in \{1, ...n\}$ be the corresponding absolute difference for the $i'th$ non-transition pair. (For mathematical and notational convenience, we are assuming an equal number, n, of transition and non-transition pairs. But there are more-or-less straightforward generalizations.) Let N be the total number of these absolute-difference observations ($= 2n$). Notice that $Z_{A(T)}$ has $2N$ components. If in fact object T is present, then it is expected that a typical transition value, x_j, will be larger than a typical non-transition value, y_j. The *rank sum* statistic is a robust measure of the extent to which this expectation is realized. In fact, the rank sum is invariant to a broad class of transformations of the data. If R_i is the rank of x_i among the N numbers $x_1, ...x_n, y_1, ...y_n$ (assigning rank 1 to the largest of the N values, rank 2 to the next largest, and so on [1]), then the rank sum is

$$R = \sum_{i=1}^{n} R_i.$$

It is easy to see that $R \geq n(n+1)/2$.

[1] This is unconventional but convenient. Usually, small values are assigned small ranks, but the definition adopted here makes for simpler notation. In any case, the two conventions lead to the same algorithm.

It is assumed that R has exponential distribution:

$$P(R=r)=(1-e^{-\alpha})e^{-\alpha(r-r_o)}, \quad r=r_o, r_o+1, ...$$

where $r_o = n(n+1)/2$, and α is a constant which determines the extent to which large rankings of transition pairs (i.e. unexpectedly small transition values) are unlikely. Actually, there is also an upper limit on the possible values of R (namely $(3n^2+n)/2$), which can be ignored without consequence.

A more intuitive parameterization can be derived in terms of a "flip probability" p, which is the probability that a randomly chosen transition is *smaller* than a randomly chosen non-transition. This provides a natural way to characterize the noise level; the value of p is a measure of the degree of degradation. One way to compute α as a function of p is to compute the expected value of R in two ways: as a function of α alone and as a function of p alone. In the former case,

$$E[R]=\sum_{r=r_o}^{\infty} rP(R=r) = r_o + \frac{e^{-\alpha}}{1-e^{-\alpha}}.$$

In the latter case, one observes first that $E[R] = nE[R_1]$, and then that $E[R_1]$ is one plus the expected number of other transition tests with smaller ranks, namely $(n-1)/2$, plus the expected number of non-transition tests with smaller rank, namely np. Thus

$$E[R]=n(1+\frac{n-1}{2}+np)=r_o+n^2p.$$

Equating the two formulas for $E[R]$ yields

$$\alpha(p)=\ln\frac{n^2p+1}{n^2p}.$$

Put this, finally, back into the formula for $P(R = r)$:

$$P(R = r) = (1 - \frac{n^2 p}{n^2 p + 1})(\frac{n^2 p}{n^2 p + 1})^{r - r_o}, \ r = r_o, r_o + 1, ... \tag{1}$$

It is, of course, the components of $Z_{A(T)}$ which are actually observed. To get to a distribution on these, it is further assumed that, under the hypothesis represented by T, the distribution on $Z_{A(T)}$ *depends only on the rank sum* of the transition values. Writing $r(Z_{A(T)})$ $(= r(x_1, ...x_n, y_1, ...y_n))$ for the rank sum, it then follows that:

$$P(Z_{A(T)}; T) = \frac{1}{C(r(Z_{A(T)}))} P(R = r(Z_{A(T)})), \tag{2}$$

where $C(r)$ is the number of ways to arrange the $2N$ grey-level intensities so as to arrive at a rank sum value of r.

The combinatorial factor $C(r)$ is computationally intractable. Fortunately, an approximation can be derived through an application of the central limit theorem. Suppose, for the moment, that the joint distribution of the $2N$ components of $Z_{A(T)}$ were iid uniform on the intensity scale $\{0, 1, ...255\}$. It is not hard to show that, in this case, the rank sum R is asymptotically (large n) normal with mean

$$\mu_o = n(\frac{N + 1}{2}),$$

and variance

$$\sigma_o^2 = n^2(\frac{N + 1}{12}).$$

Clearly, given $R = r$, all assignments of values to the components of $Z_{A(T)}$, which result in a rank sum of r, are equally likely (recall the assumption of uniformity). Thus

$$\begin{aligned}
(\frac{1}{256})^{2N} &= P(Z_{A(T)}) \\
&= P(Z_{A(T)}|R = r(Z_{A(T)}))P(R = r(Z_{A(T)})) \\
&= \frac{1}{C(r(Z_{A(T)}))}P(R = r(Z_{A(T)})) \\
&\approx \frac{1}{C(r(Z_{A(T)}))}\frac{1}{\sqrt{2\pi\sigma_o^2}}\exp(-\frac{1}{2\sigma_o^2}(r(Z_{A(T)}) - \mu_o)^2).
\end{aligned}$$

From which follows the approximate formula

$$C(r(Z_{A(T)})) = 256^{2N}\frac{1}{\sqrt{2\pi\sigma_o^2}}\exp(-\frac{1}{2\sigma_o^2}(r(Z_{A(T)}) - \mu_o)^2). \qquad (3)$$

2.1.2 Null (Background) Model

Rejection regions for hypothesis tests performed during operation of the decision tree are derived under the hypothesis that one of a certain (node-dependent) class of objects is present (see §2.2), rather than under the background, or null, hypothesis. Hence, the null model does not affect the operation of the decision tree.

On the other hand, the intention here is to (approximately) maximize the entire image likelihood relative to the placement of object types in the image. (Recall that the purpose of the decision tree is merely to highlight candidate objects.) Up until now, the data distribution has only been specified at certain pixels in the neighborhood of the object boundary. Other locations will be assumed to behave like background, and therefore the full data likelihood does involve the null model. Ideally, the null model would accommodate non-object structures, especially those structures ("clutter") likely to be confused with objects, *but no usable and believable models of this type are known.* Instead, a simple iid model governing background grey-level intensities has been adopted. Any distribution for the marginals would be manageable; we have used the uniform distribution on $\{0, 1, ...255\}$.

2.1.3 Independence-Type Assumptions

Given a configuration of objects, and given an observed grey-level image, the complete data-likelihood is the probability of the array of grey-level values representing the observed image. The likelihood is viewed as a function of the object configuration. A maximum likelihood labeling is an object configuration at which this function achieves a (global) maximum. The goal of the approach described here is to identify a maximum likelihood labeling.

For any object configuration, the likelihood involves probabilities of background pixels under the null data model as well as probabilities of object pixels under the object data model. Certain assumptions about the dependencies among the image intensities are made under which the problem of computing good approximations to the maximum likelihood labeling is rendered manageable. More specifically, under various assumptions of conditional independence, the problem can reasonably be attacked *locally*—candidate objects are evaluated based purely on image intensities in the purported-object vicinities. The necessary assumptions are made explicit in the following paragraphs.

Consider first the data likelihood, given an object configuration consisting of t objects. Let T_i, $i \in \{1, 2, ...t\}$, represent the type, location, and pose (scale and orientation) of the $i'th$ object in the configuration. As mentioned earlier, the objects are assumed to be non-overlapping. Associated with each object is a collection of pixel locations from which the transition and non-transition statistics are collected. Given an object T, let $A(T)$ be used to represent this collection of pixel locations. (If there are n transition pairs, and n non-transition pairs, then $A(T)$ contains $4n$ pixel locations.) Following earlier notation, Z will represent the entire array of pixel grey levels, and given a subset A of pixels locations, Z_A will represent the corresponding pixel grey levels. Let $C = \mathcal{C}(\bigcup_{i=1}^{t} A(T_i))$, which is the set of all pixel locations *not* associated with an object.

The first independence-type assumption is that, given the configuration of objects $T_1, ...T_t$, the random vectors $Z_{A(T_1)}, ...Z_{A(T_t)}, Z_C$ are independent. Formally:

$$P(Z; T_1, ...T_t) = (\prod_{i=1}^{t} P(Z_{A(T_i)}; T_i))P_o(Z_C),$$

where $P_o(\cdot)$ is the null-model probability distribution. This is a strong assumption, substantially wrong unless $A(T)$ comprises most or all pixel locations relevant to the hypothesis represented by T. In particular, not only is an object's internal detail ignored, but it is further assumed to be governed by the null data model.

A second independence-type assumption concerns the background data model. In the absence of objects, $P_o(Z)$ becomes the image grey level distribution. The assumption is that grey-level values associated with connected and "substantially-sized" disjoint regions are independent under the null distribution. This is, and will remain, somewhat imprecise. It is meant to apply to regions such as the collection $A(T_1), ...A(T_t), C$ entertained in the previous paragraph. More formally, given disjoint sets of pixel locations $B_1, ...B_t$,

$$P_o(Z) = (\prod_{i=1}^{t} P_o(Z_{B_i}))P_o(Z_D),$$

where $D = \mathcal{C}(\bigcup_{i=1}^{t} B_i)$, the set of pixel locations not included in at least one of the sets $B_1, ...B_t$.

2.2 Hypothesis Testing for Identification of Candidate Objects

Basically, the procedure, *which is executed at each pixel in the image,* is coarse-to-fine in the space of all possible object/pose pairs. More specifically, a binary decision tree is constructed in such a way that the root node corresponds to a hypothesis test for the compound hypothesis "object present" versus the null, "background," hypothesis. If the test succeeds at a given pixel, then tests associated with each of the two daughter nodes are performed. These correspond to somewhat more specific hypotheses, dividing the hypothesis "object present" into two disjoint sets of possible object/pose pairs. If the test associated with a daughter node succeeds, then the tests associated with each of the two daughter nodes below the successful node are performed. This procedure continues down the tree—the terminal nodes correspond to testing for single object/pose pairs. The null hypothesis is accepted at a pixel if no terminal node is reached. Typically, the null hypothesis is accepted at the root node, and therefore no additional tests are performed.

A clustering procedure is used to define the binary decision tree. Each node represents a registered collection of object/pose pairs. Each terminal node represents a single object/pose pair; the collection of terminal nodes represents the collection of meaningfully-distinct object/pose pairs. Parent nodes are generated, recursively, by combining the object/pose pairs of two "similar" daughter nodes. This is the clustering procedure. It is continued recursively until arriving at a single (root) node representing all possible object/pose pairs. (It may be that the entire collection of object/pose pairs can not be registered in such a way as to produce a substantial common interior. In this case the clustering is terminated prematurely, thereby producing two or more trees, each with a single root node.)

For each node, and any given location in the image, a statistic is derived for testing the compound hypothesis that one of the associated object/pose pairs is present, versus the "background" or "null" alternative. When applied to the terminal nodes, this amounts to testing for the presence of a particular object at a particular pose.

The tests are based upon the object data model described earlier. This entails defining a node-specific collection of transition and non-transition pixel pairs. Associated with each node is a deformed annulus, or "ribbon," separating the plane into three sets of points: those in the common interior of the objects in the collection; those in the common exterior of the objects in the collection; and the ribbon locations, representing the remaining, ambiguous, points.

Each ribbon is associated with a set of "transition" pixels and a set of "non-transition" pixels. The former are pairs of points, one immediately inside the ribbon and one immediately outside, with the line connecting these points roughly normal to the ribbon. There are an equal number of non-transition pairs, these being outside the ribbon, but in the vicinity of the ribbon. The pixels in a typical non-transition pair are about as far apart as the pixels in a typical transition pair.

According to the data model introduced earlier, the presence of any one of the object/pose pairs associated with a node is characterized by an exponential distribution on the rank-sum statistic of the transition intensities. A natural test statistic for the compound hypothesis "one of the node-specific object/pose pairs present," is, therefore, the rank sum; large values can be interpreted as evidence against the hypothesis. This suggests a rejection region of the form $\{R \geq \gamma\}$, where the threshold, γ, is chosen so as to achieve a user-specified probability, β, of missed detection:

$$P(R \geq \gamma) = \beta.$$

The probability is calculated under the object-model distribution—see equation 1. It follows easily that

$$\gamma = n(\frac{n+1}{2}) - \frac{\log \beta}{\log\{\frac{n^2p+1}{n^2p}\}}.$$

There are only two parameters for the entire tree: the "flip probability" p, and the probability of missed detection, β. It should be pointed out, however, that even if the underlying probabilistic assumptions were exactly true, the actual probability of missed detection would be higher than β. This is simply because many tests, at rejection probability β, are performed sequentially before a hypothesis is accepted. In principle, node-specific thresholds could be calculated in such a way as to achieve, for each object, a rejection probability of β, while at the same time maximizing a measure of power (such as the probability that no object is detected given the null hypothesis). This is a problem in sequential decision theory, and it may well be tractable. (The tree structure lends itself to dynamic programming, a principal tool in sequential analysis.) Nevertheless, the calculation would appear to be unwarranted, given the numerous assumptions and approximations upon which the probabilistic model is based. A simple and expedient alternative is to choose β so as to reach a favorable setting on the ROC curve. This has been the practice, so far.

2.3 Maximum Likelihood Labeling

The procedure rests upon the assumption that the true maximum of the data likelihood, over all allowable (i.e. non-overlapping) labelings, can be achieved by restricting to objects identified by the decision tree. Obviously, this is not always the case. The decision tree does occasionally reject a correct object/pose pair, sometimes, in fact, by failing to detect any object at all. However, these failures are fairly rare, and it is reasonable to proceed with a maximum likelihood calculation that is restricted to the output from the decision tree.

Recall that there is an assumption of no overlap of object signatures. Therefore, the problem of maximizing the data likelihood, restricted to detections made within the decision tree, becomes one of choosing among candidate objects at those locations for which there are multiple overlapping candidates.

A primary difficulty is that the object-data models are distributions on object-dependent subsets of the pixel array. Specifically, under the hypothesis T, representing a particular object at a particular location and pose, the data model is a distribution restricted to the grey levels of pixels in $A(T)$, i.e. restricted to $Z_{A(T)}$. Two candidate objects, T and T', therefore can not be meaningfully compared by simply examining $P(Z_{A(T)};T)$ and $P(Z_{A(T')};T')$. The independence-type assumptions (see §2.1.3) provide a mechanism for properly normalizing the data probabilities, as follows:

$$\begin{aligned}
& \arg \max_{t,\{T_1,T_2,...T_t\}} P(Z;T_1,T_2,...T_t) \\
= \; & \arg \max_{t,\{T_1,T_2,...T_t\}} \frac{P(Z;T_1,T_2,...T_t)}{P_o(Z)} \\
= \; & \arg \max_{t,\{T_1,T_2,...T_t\}} \frac{(\prod_{i=1}^{t} P(Z_{A(T_i)};T_i))P_o(Z_C)}{(\prod_{i=1}^{t} P_o(Z_{A(T_i)}))P_o(Z_C)} \\
& \left(\text{recall that } C = \mathcal{C}(\bigcup_{i=1}^{t} A(T_i))\right) \\
= \; & \arg \max_{t,\{T_1,T_2,...T_t\}} \prod_{i=1}^{t} \frac{P(Z_{A(T_i)};T_i)}{P_o(Z_{A(T_i)})}
\end{aligned}$$

The problem is thereby reduced to one of comparing competing hypotheses via the *likelihood-ratio score function*:

$$\frac{P(Z_{A(T)};T)}{P_o(Z_{A(T)})}.$$

For the numerator, combine equations 1, 2, and 3:

$$P(Z_{A(T)};T) = \frac{(1 - \frac{n^2p}{n^2p+1})(\frac{n^2p}{n^2p+1})^{r(Z_{A(T)})-r_o}}{256^{2N}\frac{1}{\sqrt{2\pi\sigma_o^2}}\exp(-\frac{1}{2\sigma_o^2}(r(Z_{A(T)})-\mu_o)^2)},$$

where $r_o = n(n+1)/2$, $\mu_o = n(\frac{N+1}{2})$, and $\sigma_o^2 = n^2(\frac{N+1}{12})$. As discussed earlier, the null-data distribution is simply iid uniform on $\{0, 1, ...255\}$, $P_o(Z_{A(T)}) = (1/256)^{2N}$. Hence the score function is

$$\frac{P(Z_{A(T)}; T)}{P_o(Z_{A(T)})} = \frac{(1 - \frac{n^2 p}{n^2 p+1})(\frac{n^2 p}{n^2 p+1})^{r(Z_{A(T)})-r_o}}{\frac{1}{\sqrt{2\pi\sigma_o^2}}\exp(-\frac{1}{2\sigma_o^2}(r(Z_{A(T)}) - \mu_o)^2)}. \tag{4}$$

3 Suggestions for Improving Performance

1. Recall that the parameter p represents the probability that a random pair of transition pixels has smaller absolute difference in grey levels than a random pair of non-transition pixels, when in fact an object is present. Currently, p is treated as a global parameter. In particular, a single value is assumed to apply to each object in the image.

   It is perhaps more natural, and it may be more effective, to treat the flip probability as a nuisance parameter, at least for the purposes of computing the likelihood ratio (score function) of candidate objects. (It is common practice to replace nuisance parameters with their maximum-likelihood estimates when computing likelihood ratios. There is a sound theoretical justification for this when the likelihood ratio is to be used as a statistic for hypothesis testing.)

   Given an observed rank sum, r, the maximum likelihood estimator for p is easily derived from equation 1:

$$\hat{p} = \frac{r - r_o}{n^2} = \frac{r - \frac{n(n+1)}{2}}{n^2}.$$

The expression has a simple interpretation: The numerator is the number of times that a non-transition pair yields an absolute difference that exceeds the absolute difference of a transition pair. The denominator is the number of such comparisons that are observed. Thus $\hat{p}$ is the empirical relative frequency of such flips.

Experiments should be performed in which p appearing in the score function (equation 4) is replaced by $\hat{p}$.

The empirical flip probability provides an intuitive goodness-of-fit measure. The average background value of $\hat{p}$ is .5; a value of 0 can be taken as strong evidence for a given hypothesis. It might be useful to append the retention lists with estimated p values for each candidate object.

2. As explained earlier, null and object data models are based upon observations of absolute differences, of the form $|Z_s - Z_t|$, where s and t are locations of purported "transition" and "non-transition" pairs. It is possible that better performance would be realized if $|Z_s - Z_t|$ were replaced by the more specific (and therefore more informative) statistic

$$\max_{l \in L(s,t)} |\mathrm{grad}(Z_l)|$$

   where $L(s,t)$ is the set of pixel locations lying along the line segment between s and t, and $\mathrm{grad}(Z_l)$ is a discrete gradient of the image intensity array at l.

3. Quite clearly, the data models, both object and null, are at best gross approximations of reality. There are several possible directions for improvement.

Concerning the object data models, it is likely that the use of the rank sum as a sufficient statistic actually goes *too far* in accommodating distortions of the grey-level scale. By restricting the likelihood to depend only on the *order* of the absolute differences associated with transition and non-transition pairs, an object with good separation between the transition and non-transition populations is not preferred to one with arbitrarily small separation. In essence, the coverage of the object data model is too broad. One would expect that a more favorable ROC curve could be achieved with a data model that caters more accurately and more specifically to real object signatures.

Of course the task of crafting good data models is difficult. Here is a concrete suggestion that may improve performance. It is designed to be robust to outlying values and to be scale invariant, but to otherwise reward good separation between transition and non-transition tests.

Following earlier notation, T will represent an object hypothesis by specifying the object type, object location, and object pose; $A(T)$ is the collection of pixels involved in transition and non-transition pairs; $Z_{A(T)}$ represents the corresponding components of Z; $x_1, ...x_n$ and $y_1, ...y_n$ are the absolute differences of intensities of transition and non-transition pairs, respectively. Let m_x and m_y be the median values of the transition differences and non-transition differences, respectively. Define the ramp function

$$\phi(x, \mu, \nu) = \left[\frac{x-\mu}{\nu-\mu}\right]^+$$

where $[x]^+$ is x when $x > 0$, and 0 otherwise. The distribution on $Z_{A(T)}$

is assumed to depend only on $x_1, ...x_n, y_1, ...y_n$, and the distribution of $x_1, ...x_n, y_1, ...y_n$ is assumed to be of the form

$$P(x_1, ...x_n, y_1, ...y_n; T) =$$

$$\frac{1}{\mathcal{Z}_\xi} \exp\{-\xi \sum_{i=1}^{n} \phi(x_i, m_x, m_y) - \xi \sum_{i=1}^{n} \phi(y_i, m_y, m_x)\} 1_{m_x > m_y}.$$

The parameter ξ determines signal-to-noise (much like p of the current model), and $\mathcal{Z}_\xi$ is the normalizing constant (partition function). Zero mass is assigned to configurations in which $m_x \leq m_y$, by virtue of the indicator function $1_{m_x > m_y}$.

Given T, and given $x_1, ...x_n, y_1, ...y_n$, all values of $Z_{A(T)}$ that are compatible with $x_1, ...x_n, y_1, ...y_n$ are assumed to be equally likely. Let $D(x_1, ...x_n, y_1, ...y_n)$ be the number of such arrangements of $Z_{A(T)}$. The object data distribution is, then,

$$P(Z_{A(T)}; T) = \frac{1}{D(x_1, ...x_n, y_1, ...y_n)} P(x_1, ...x_n, y_1, ...y_n; T).$$

The combinatorial factor, $D(x_1, ...x_n, y_1, ...y_n)$, is easy to compute in closed form, and the partition function, $\mathcal{Z}_\xi$, admits to a manageable analytic approximation (through a somewhat tedious calculation).

Critical regions for hypothesis testing (for use in the decision tree) as well as a likelihood ratio score function are easily arrived at. The natural statistic for hypothesis testing is the "energy"

$$H = \sum_{i=1}^{n} \phi(x_i, m_x, m_y) + \sum_{i=1}^{n} \phi(y_i, m_y, m_x).$$

Large values are evidence against the object, T, hypothesis. From the well-known relations $E[H] = -\frac{d}{d\xi} \mathcal{Z}_\xi$ and $Var[H] = \frac{d^2}{d\xi^2} \mathcal{Z}_\xi$, and the above-mentioned analytic expression for $\mathcal{Z}_\xi$, $E[H]$ and $Var[H]$ can be shown to be well approximated by n/ξ and n/ξ^2 respectively. These observations suggest the critical region $\{H > (n/\xi) + \theta\sqrt{n/\xi^2}\}$ for testing T against the null hypothesis, with θ used to adjust the probability of missed detection.

The uniform distribution could again serve as the null model, and competing hypotheses could again be scored via the likelihood ratio. It has been suggested (item 1 above) that $\hat{p}$ replace p in the currently-used score function. Similarly, the proposed new score function may be most effective with an estimated, rather than fixed, value for ξ. The expression $E[H] = n/\xi$ suggests the moments estimator $\hat{\xi} = n/H(x_1, ...x_n, y_1, ...y_n)$.

4. Very little effort has gone into devising reasonable null (background) data models. Certainly the uniform model can be improved, possibly to good effect. The assumption of independence is convenient, but the use of a uniform marginal distribution leaves much room for improvement. A first step would be to experiment with Gaussian marginals. The empirical mean and empirical variance would be used in computing the likelihood ratio score function, again using the generalized likelihood ratio as a guideline.

   A further step in this direction would be to model dependent noise, which produces background *structure*, and which thereby tends to promote false alarms. It might be possible to do this effectively by adopting *object fragments* as a working definition of clutter, and then by devising a null model which includes random placements of these fragments.