Information Fusion Algorithms and Analysis for an Exemplar Detection of Intent Problem

Chris Lloyd, David Nicholson, and Mark Williams
BAE Systems Advanced Technology Centre
PO Box 5, Filton, Bristol BS34 7QW, UK
{chris.m.lloyd,david.nicholson2,mark.l.williams}@baesystems.com

Abstract – Information fusion algorithms for data association and inference are applied to a representative intelligence gathering problem in which signals of intent are monitored by multiple imperfect sensors over a period of time. Two sets of algorithms are developed: a brute force set which makes best use of the data but is not efficient, and an approximate set which sacrifices some performance for efficiency. The algorithms are applied to simulated data to generate evidence of intent. Then a Monte Carlo process and an associated metric are developed to evaluate the performance of the algorithms under different levels of uncertainty in the data. This analysis helped to validate the algorithms and it can also provide useful system design guidelines.

Keywords: Statistics; Data Association; Data Fusion; Monitoring; Surveillance; Security; Detection of Intent

1 Introduction

Intent is a signal to act in a way that could, for example, cause harm. It may be biological intent (to cause disease), environmental intent (to cause flooding), or military intent (to damage a high value asset). The source of intent may be natural (e.g. a virus, the weather), or human (e.g. a computer hacker or suicide bomber). Signals of intent are generally difficult to detect because they are embedded in a background of clutter generated by unrelated activities. Moreover, they are only partially observed through noisy sensor reports. Models of the signal, background, and noise sources are often uncertain and incomplete. Thus the problem of detecting intent is one of extracting a weak signal from a noise background under challenging conditions of data and model uncertainty.

Our goal is to develop a general mathematical framework for detecting intent. Therefore we focus on a *synthetic* problem and *simulated* data that contain some generic aspects of the problems noted in the opening paragraph. In particular, the presence of significant uncertainty creates a strong requirement for information fusion. This problem, known as Cogentry, was defined by scientists at the UK's Defence Science and Technology Laboratory (Dstl). It comprises a scenario description document and related data files. These were provided to the authors in support of their work to develop a mathematical framework for all-source fusion under the UK MOD's Multi-Intelligence Techniques (MIT) programme.

In this paper a Bayesian formulation of the Cogentry problem is described. Then it is used to derive algorithms which solve the problem and analyze its solutions. Specifically, the problem is decomposed into two phases, data association and inference, which are solved in turn. Two sets of algorithms were developed. The first set makes best use of the data and thus solves the problem very accurately, but it requires significant computational expense. The second set makes some approximations in order to solve the problem more efficiently, but with less accuracy. In practice, a pragmatic choice would depend on the scale of the problem, the quality of the data, and the performance requirement. The paper includes results from an empirical analysis which highlights a number of performance sensitivities and trade-offs. These provide checks on the algorithms as well as guidelines for designing systems to gather and process intelligence data.

The paper is organized as follows. Section 2 describes the Cogentry problem and its data. Section 3 presents our two-phase approach to deriving solutions. The accurate 'brute force' and approximate solutions are described in Sections 4 and 5, respectively. Section 6 contains the main results from an empirical analysis of the solutions and the paper closes with discussion and conclusions in Section 7.

2 The Cogentry Problem

Cogentry is a synthetic problem concerned with detecting breaches in quota (regarded as triggers of intent) for specific objects being manufactured at multiple factories. This section describes the problem scenario, its simulated datasets, and the main intelligence analysis requirements.

2.1 Scenario

In a town there are ten 'factories' ($f_1 \ldots f_{10}$) where up to five types of 'object' ($o_1 \ldots o_5$) may be manufactured. The factories are free to manufacture objects $o_3 \ldots o_5$ in any quantity, but there are controls on how many o_1 and

o_2 they produce. The quota of o_1 and o_2 for the current accounting period is full and the authorities want to monitor the town to check if its factories exceed quota in subsequent periods. This may signify an intention to carry out actions which the authorities would wish to prevent.

The authorities employ two CCTV monitoring techniques:

1. *Monitoring the roads into the town*. The CCTV operators may be able to identify the type of delivery vehicle $(v_1 \ldots v_3)$ entering the town and, if the vehicle is open-topped, its cargo type.
2. *Monitoring the factory entrance gates*. The CCTV feeds are lower quality and the operators are only able to identify the delivery vehicle type.

The CCTV operators have varying levels of experience and may occasionally misidentify vehicles and cargo. They record each observation with a HIGH, MEDIUM or LOW precision indicator.

The authorities know what 'components' $c_1 \ldots c_{10}$ are required to produce each object. This information is presented in Table 1. Note the overlaps: different objects frequently share the same components.

The authorities also know what components each delivery vehicles is capable of carrying as cargo. This information is shown in Table 2. Again, note the significant overlaps: different vehicles carry very similar cargos.

Table 1: Components required for object manufacture

		Components									
		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
Objects	o_1	✓	✓		✓	✓		✓			✓
	o_2	✓	✓	✓		✓			✓		✓
	o_3	✓	✓	✓		✓	✓	✓		✓	
	o_4	✓	✓		✓	✓	✓		✓		✓
	o_5	✓	✓			✓				✓	

Table 2: Cargo (components) carried by each vehicle

		Components									
		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
Vehicles	v_1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	v_2	✓	✓	✓		✓	✓	✓	✓	✓	✓
	v_3	✓	✓	✓		✓	✓	✓	✓		

Finally, it is known that each factory complies with 'just in time' manufacturing standards and has no significant on-site storage facilities. However, no information is available about the factories normal delivery patterns, or the expected quantity of objects they manufacture, or the daily number of component deliveries they require.

2.2 Datasets

A dataset of CCTV operator reports was provided for a period of three calendar months. Each entry in the database contained the following records:

- Time-stamp
- CCTV operator identity code from set {1,2,…16}
- Road number or factory number from sets {1,2,…6} and {1,2,…10}, respectively
- Observed delivery vehicle identity from set {1,2,3}
- CCTV operator confidence about identity of delivery vehicle from set {1=HIGH,2=MEDIUM,3=LOW}
- Observed cargo identity from set {0='no cargo observed',1,2,…10}
- CCTV operator confidence about identity of cargo from set {0='no cargo observed',1=HIGH, 2=MEDIUM,3=LOW}

A further dataset was provided containing historical performance data for each CCTV operator. Each entry in the database contained the following records:

- CCTV operator identity code
- Number of previous HIGH confidence observations and number of these proven correct
- Number of previous MEDIUM confidence observations and number of these proven correct
- Number of previous LOW confidence observations and number of these proven correct
- Number of delivery vehicles missed

An initial inspection of these datasets revealed a number of challenging features: a surplus of vehicle sightings on the roads compared to factories (i.e. missing data), a significant population of MEDIUM and LOW confidence observations (i.e. uncertain data), and a large fraction of closed-top delivery vehicles (i.e. ambiguous data).

2.3 Intelligence Requirements

The primary intelligence requirement is to calculate how likely each factory is to be exceeding its quota, given the CCTV operator reports and any domain knowledge that is contained in the problem description. This requirement is challenging because of the dimensionality of the problem (multiple combinations of objects comprising multiple combinations of components may be manufactured at multiple factories), and the presence of ambiguity about the factories manufacturing processes (delivery schedules and production lines), and uncertainty in the CCTV data. The secondary intelligence requirement is to quantify confidence in the results concerning production at the factories, and probe the sensitivity of these results to

system variables, such as the fraction of LOW confidence operator reports, and the number of closed-top vehicles.

3 General Approach

First, a state space x must be defined for the problem. This is a joint state space over the delivery vehicles x^v, delivery times x^t, and production at each factory x_i^p.

Next, make some reasonable assumptions about the problem to factorise the state space and simplify algorithm development. The following assumptions were made:

- The production across factories is independent and within factories is independent of delivery vehicles and delivery times. This implies the factories are not colluding and there is no correlation between their productions and their delivery schedules.
- The observed delivery vehicle type, cargo type, and time stamp, depend only on their respective true values. This implies, for example, that the quality of an observation of a delivery truck or its cargo does not depend on when the observation was made.

Finally, decompose the problem into two separate phases. This is necessary because, in order to infer the production at each factory, information about which components are entering the factories is required. However, the low-grade factory CCTV cameras only observe delivery vehicle types and this information must be implied from the component information observed by the road CCTV cameras. The two sub-problems are therefore:

1. **Association**. Generate candidate pairings between observations of the delivery vehicles on the road and at the factory gates.
2. **Inference**. Given the data associations, estimate the expected number of objects being manufactured at each factory.

3.1 Association

The association problem is defined as follows: given a set of road and factory CCTV observations of the delivery vehicles, Z^r and Z^f respectively, generate the set a of associations between these observations. Following [1] and applying Bayesian probability theory, under a variety of reasonable statistical independence assumptions, results in an equation for the probability distribution over the association hypotheses:

$$p(a \mid Z^f, Z^r) \propto \prod_{i,j \in a} p({}_i z^f, {}_j z^r \mid a) \cdot \prod_{i \in a} p({}_i z^f, \bar{z}^r \mid a) \cdot \prod_{j \in a} p(\bar{z}^f, {}_j z^r \mid a) \quad (1)$$

This equation contains joint distributions over each vehicle observation at the roads and the factories, ${}_j z^r$ and ${}_i z^f$, as well as the missing observations, $\bar{z}^r$ and $\bar{z}^f$. These distributions are conditioned on the state space x.

Equation (1) can be mapped into a cost/score matrix and solved via linear programming. The mapping is given by,

$p({}_i z^f, {}_j z^r \mid a)$	$p({}_i z^f, \bar{z}^r \mid a)$
$p(\bar{z}^f, {}_j z^r \mid a)$	1

The upper left quadrant contains the association costs for observation pairs (i, j). This is a product of joint likelihoods for the vehicle and time-stamp observations under the association a:

$$p({}_i z^f, {}_j z^r \mid a) = p({}_i z_v^f, {}_j z_v^r \mid a) \cdot p({}_i z_t^f, {}_j z_t^r \mid a) \quad (2)$$

The likelihoods can be calculated by integrating over their respective state spaces, i.e.

$$p({}_i z_v^f, {}_j z_v^r \mid a) = \frac{\sum_{x^v} p({}_i z_v^f \mid x^v) p({}_j z_v^r \mid x^v) p(x^v)}{p({}_i z_v^f) p({}_j z_v^r)} \quad (3)$$

$$p({}_i z_t^f, {}_j z_t^r \mid a) = \frac{\iint p({}_i z_t^f \mid x_f^t) p({}_j z_v^r \mid x_r^t) p(x_f^t, x_r^t) dx_f^t dx_r^t}{\bar{v} p_0}$$

The double integral is represented by a gamma distribution, with parameters (k, θ), on the delay between road and factory observations. The denominator $\bar{v} p_0$ is an 'adaptive threshold': see [2] for details. The gamma parameters (k, θ) and the prior distribution $p(x^v)$ can be estimated from the data, but the details are omitted here.

The top right and lower left quadrants of the cost/score matrix contain the missed detection costs on the diagonals. These are:

$$\begin{aligned} p({}_i z^f, \bar{z}^r \mid a) &= 1 - P_D^r \cdot (1 - P_{FA}) \\ p(\bar{z}^f, {}_j z^r \mid a) &= 1 - P_D^f \end{aligned} \quad (4)$$

The probabilities of missed detection at the factories and the roads, P_D^f and P_D^r respectively, can be inferred from the operators historic performance dataset. The probability

of false-alarm, P_{FA}, is estimated from the data as the surplus of road observations relative to factory observations after accounting for any differences in probability of detection. The off diagonals in these quadrants are filled with negative infinities to prohibit their selection as an assignment. The lower right quadrant is populated with ones to counter-balance the scores in the upper left quadrant.

Given the score matrix the problem is to generate candidate pairings, or associations, between the observations. A 'brute force' accurate solution is presented in Sec. 4 and an approximate solution in Sec 5.

3.2 Inference

The inference problem takes output from the association algorithm (implied observations of components at the factories), and the known set of components required to manufacture each object, to infer probabilities for hypotheses about the state of production at each factory.

Consider the production hypothesis p at factory i:

$$x_i^p \in \{(o_i^1, o_i^2, \ldots o_i^5) \mid o_i^k \in M_0, \sum o_i^k \le \text{MAX}\} \quad (5)$$

This states that the factory production consists of up to MAX objects, comprising o_i^k objects of type k, where there are five types in total and o_i^k is a natural number. MAX was introduced to bound the number of hypotheses.

The set of required components for the production hypothesis x_i^p is a deterministic function $y_i^p = L(x_i^p)$:

$$y_i^p \in \{(c_i^1, c_i^2, \ldots, c_1^{10}) \mid c_i^k \in N_0, \sum c_i^k = N\} \quad (6)$$

Similarly, this states that c_i^k components of type k are required, where c_i^k is a natural number and N is the total number of components required.

Each association a between the observed delivery vehicles on the roads and at a factory i implies observations of components at the factory (denoted Z_i^p), from which to potentially service production hypothesis p. The observation likelihood under association a is,

$$p(Z_i^p \mid x_i^p, a) = p(Z_i^p \mid y_i^p, a) \quad (7)$$

The data association phase generates a set of association likelihoods $p(a \mid Z^r, Z^f)$. These are used to weight the observation likelihoods above and a weighted sum is performed over the association set. Bayes Rule is then applied to determine the desired posterior probability of the production hypothesis:

$$p(x_i^p \mid Z_i^p, Z^r, Z^f) \propto p(x^p) \times \sum_a p(Z_i^p \mid x_i^p, a)\, p(a \mid, Z^r, Z^f) \quad (8)$$

This enables the probability that a number k of type j objects are being produced to be calculated as follows:

$$p(o_i^j = k \mid \bullet) = \sum_{x^p} p(o_i^j = k \mid x_i^p)\, p(x^p \mid \bullet) \quad (9)$$

where,

$$p(o_i^j = k \mid x_i^p) = \begin{cases} 1 & \text{if } o_i^j(x_i^p) = k \\ 0 & \text{otherwise} \end{cases}$$

and a flat prior is used for $p(x^p)$. The expected number of each object type j in production at factory i is then:

$$E[o_i^j \mid \bullet] = \sum_{k=0}^{\text{MAX}} k \cdot p(o_i^j = k \mid \bullet) \quad (10)$$

Like the data association problem, there are also 'brute force' and approximate solutions to the inference problem. These are described in Sections 4 and 5, respectively.

4 Brute Force Solutions

In theory the data association problem as posed in Section 3.1 can be solved exactly by enumerating and scoring all candidate pairings between observations. However, the number of possible associations for a score matrix of dimension n is $n!$. Murty's algorithm [3] allows the top m scoring associations to be determined in O(mn^3) time. This was achieved using publicly available software [4]. The data being associated in the Cogentry problem corresponds to classification data for only three different types of delivery vehicle. Also, the time window between observing a delivery vehicle on the road and at a factory is poorly constrained. These aspects of the data give rise to a relatively flat association hypothesis space with the m^{th} most likely association almost as likely as the 1st. Thus most of the probability mass is in $n! - m$ associations and the association likelihoods cannot be normalised to yield meaningful probabilities. Notwithstanding this, a pragmatic choice of *m* was used to generate associations.

Given the outputs from the association phase, the next step is to infer the expected production at each factory. The main challenge this poses is to compute the likelihood of

observing the components at factory i that are required to service a production p under association a, i.e. $p(Z_i^p \mid y_i^p, a)$. Suppose M_R road observations are associated with a factory and a total of N components are required. If $M_R > N$, $p(Z_i^p \mid y_i^p, a) = 0$. This is because the factory has no onsite storage facilities. Alternatively, if $M_R \le N$, we need to consider the multiplicity of ways a selection of observed components could be drawn to service the production hypothesis.

Let W be the total number of subsets $\{w_k, k = 1 \ldots W\}$, formed by permuting the observed components and let U be the number of M_R-sized subsets $\{u_j, j = 1 \ldots U\}$ corresponding to ways of selecting the required components from the observed components. Then,

$$W = M_R!; \quad U = \frac{N!}{M_R!(N - M_R)!} \tag{11}$$

The likelihood, $p(Z_i^p \mid y_i^p, a)$, is obtained by computing a product of observation likelihoods, summed for each permutation, and multiplied by the likelihood of making M_F observations at a factory from N required parts. Recall that $M_F < M_R$ because of missed detections at the factory. Thus,

$$p(Z_i^p \mid y_i^p, a) = \binom{N}{M_F} \cdot (P_D^f)^{M_F} \cdot (1 - P_D^f)^{N - M_F} \cdot \sum_{k=1}^{W} \sum_{j=1}^{U} p(Z_i^p \mid w_k)\, p(w_k \mid u_j)\, p(u_j \mid y_i^p, a) \tag{12}$$

Clearly this is a computationally demanding calculation and bounds on the number of observations, MAX, and the number of top-scoring Murty associations (m) carried forward into the inference phase, were required to solve it.

5 Approximate Solutions

The brute force solutions exhibit poor computational scaling and approximate solutions are desirable in general. These were achieved by developing a "soft-assign" solution for the data association problem and interfacing it with a model fitting solution for the inference problem.

The brute-force data association solution determines a ranked set of "hard" binary matches between input observations. Soft-assign is an alternative solution which determines a "soft" probability distribution over the association hypothesis space [5, 6]. The idea is to start with the score matrix, defined in Section 3.1, and apply an *iterative scaling* algorithm. The purpose is to transform the score matrix into a doubly-stochastic matrix and the algorithm is proven to converge. Moreover, the total number of iterations does not depend on the size of the matrix. This approach has connections to statistical physics and can be viewed as a mean-field approximation.

The brute-force inference solution involves summations of likelihood terms over permuted data sets. An approximation was found by representing these summations as a mixture of hyper-geometric distributions.

The hyper-geometric distribution (HGD) is defined as [7]:

$$p_H(y_1, \ldots, y_n \mid x_1, \ldots x_n) = \frac{\prod \binom{x_i}{y_i}}{\binom{\sum x_i}{\sum y_i}} \tag{13}$$

It can be implemented as a continuous generalisation of the binomial coefficient, i.e.

$$\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)} \tag{14}$$

Thus an approximation for the likelihood (12) is,

$$p(Z_i^p \mid y_i^p, a) \approx \binom{N}{M_F} \cdot (P_D^f)^{M_F} \cdot (1 - P_D^f)^{N - M_F} \cdot p_H(E[c^1 \mid Z_i^p], \ldots, E[c^{10} \mid Z_i^p] \mid c_i^1, \ldots c_i^{10} + \Delta c_i^p) \tag{15}$$

The Δc_i^p term is introduced to ensure the *y*'s are never less than the *x*'s in the HGD. This requires,

$$\Delta c_i^p = \max\left(E[c^1 \mid Z^p] - c_i^p, \ldots E[c^{10} \mid Z^p] - c_i^p\right)$$

In the mixture of HGDs approximation the number of mixture components grows exponentially. This was counteracted by guessing the parameters of a single HGD.

The HGD is calculated with respect to the expected number of components arriving at the factory given observation *and* association uncertainty,

$$E[c^k \mid Z_i^p] = \sum_{z_j^r \in Z^p} p(z_j^p \in Z_i^p)\, p(z_j^r \mid c^k) \tag{16}$$

The first term under the sum is calculated by summing over the soft-assign association probabilities between the road and factory observations,

$$p(z_j^p \in Z_i^p) = \sum_{{}_i z_k^f \in Z^p} p(({}_i z_k^f, z_j^r) \in a) \qquad (17)$$

Although no direct references to the HGD approximation that we have used could be found in the literature, our empirical analysis showed it performed reasonably well. Further work will be required to analyse it in more detail.

As problem size grows the accuracy of the brute force solutions will be compromised by the requirement to limit computational cost. The approximate solutions are likely to be just as good in this regime. However, further analysis will be required to explore their mutual performance as a function of the problem dimension. Our results only provide a point solution for a 'Cogentry-sized' problem.

6 Results and Analysis

This section presents results and analysis in support of the intelligence requirements defined in Sec. 2.3. The main result is expected production numbers for each object at each factory. These inform the intelligence analyst about whether a breach in quota (intent) has been detected. This result is augmented by a confidence analysis which also explores how performance is affected by various parameters which influence the intent signal-to-noise ratio.

6.1 Results

Three months of CCTV surveillance data was analysed in monthly batches. Some of the data for one of these batches is displayed in the timeline shown in Fig.1.

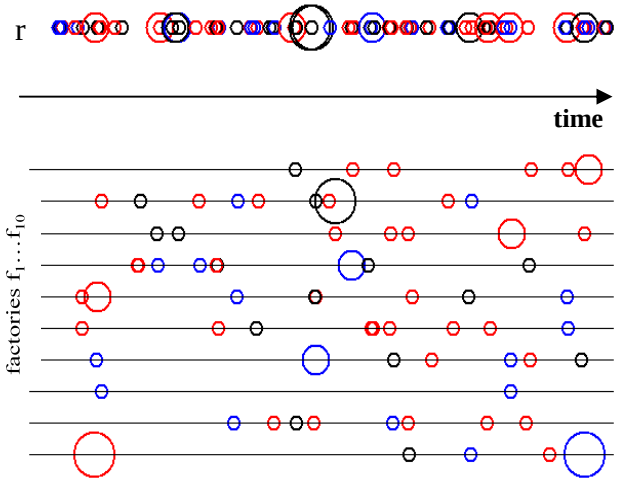


Fig.1 Timeline for CCTV observations of delivery vehicles at the road ('r') and the factories ('f_1…f_{10}'). The observations are coloured according to the type of vehicle that was observed and scaled according to the confidence of the observations (HIGH=small, LOW=large).

This data has to be associated to determine the components being delivered to the factories since this information is not directly available from the factories CCTV cameras. The result of the soft-assign data association algorithm is shown in Fig.2. This is a doubly-stochastic matrix populated by association probabilities. For perfect association the matrix would be occupied by one cell (of unit probability) in each row and each column, while for completely uninformative data every cell would have the same small probability. The Cogentry data is reasonably well associated, as indicated by the crisp nature of the soft-assign matrix with only a small amount of blurring evident.

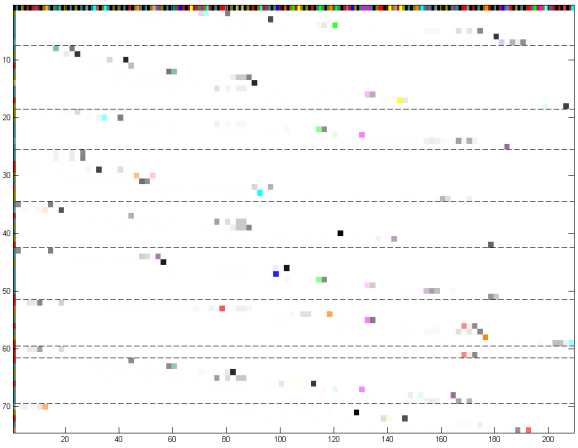


Fig.2 Soft-assign output for the data association phase. Each cell is the probability that a road observation (running down the columns) is associated with a factory observation (partitioned along the rows). The colours correspond to the delivery vehicle/component types.

The data association solution (from either the 'brute force' or approximate algorithms) is passed forward into the inference phase which determines the expected number of objects being produced at the factories, according to (10). An upper limit of MAX=2 was placed on the maximum number of vehicles of any type a factory could build during the observing period. The results are displayed in Fig.3 for the 'brute force' algorithm combination. No evidence of breach of quota was found for the objects of interest to the authorities. This was also the case for other batches of data and the results did not significantly change when the approximate algorithms combination was used.

6.2 Analysis

A Monte-Carlo simulation process was used to analyze the information fusion solutions and their performance under different conditions. The process involved creating multiple random realizations of the Cogentry datasets under specified values for probabilities of detection, ratio of LOW/HIGH confidence observations, etc. This allowed us to inject a known production of objects into the data and test the ability of our algorithms to infer it.

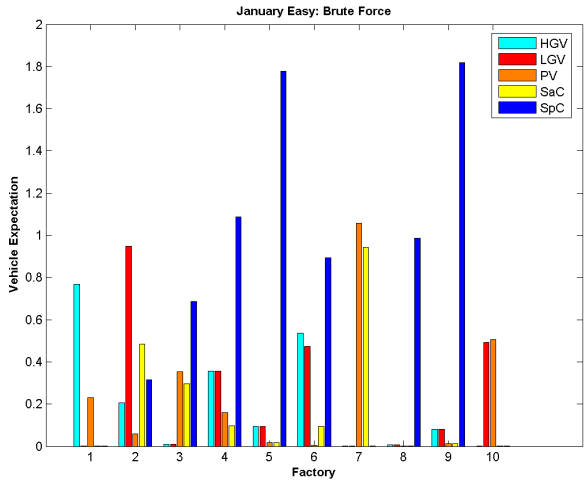


Fig.3 Expected number of objects produced at each factory. The 'HGV (cyan)' and 'LGV (red)' objects, of interest to the authorities, are not found to exceed quota.

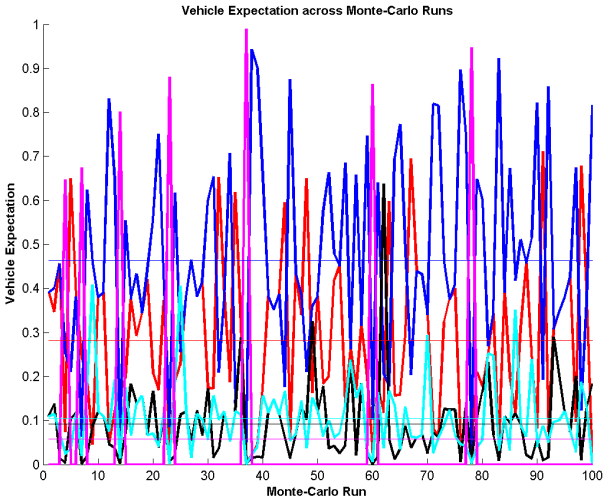


Fig.4. Expected object production at factories calculated for 100 Monte-Carlo runs. The colours represent different object types. Blue corresponds to the actual object being produced and its mean value (the straight line) is higher than the mean values for the other objects/colours

In the first test of performance only a single object type was actually being produced at the factories and the expected production was computed across all object types. This was done for 100 separate Monte-Carlo runs. The results are displayed in Fig.4. If a moderate threshold of 0.6 is set on the expected production, then across the Monte-Carlo runs there is a higher incidence of the actual object crossing threshold relative to the other objects. However, there are a significant number of missed detections, and a few false-alarms generated by one of the other objects in particular. A performance measure is required to required to quantify this further.

The derived performance measure is an error norm defined as follows:

$$\frac{\sum_{i=1}^{10}\sum_{j=1}^{5}\left\|\begin{bmatrix} o_1^1 & \cdots & o_1^5 \\ \vdots & \ddots & \vdots \\ o_{10}^1 & \cdots & o_{10}^5 \end{bmatrix}-\begin{bmatrix} E[o_1^1] & \cdots & E[o_1^5] \\ \vdots & \ddots & \vdots \\ E[v_{10}^1] & \cdots & E[v_{10}^5] \end{bmatrix}\right\|}{\sum_{i=1}^{10}\sum_{j=1}^{5}\left\|\begin{bmatrix} o_1^1 & \cdots & o_1^5 \\ \vdots & \ddots & \vdots \\ o_{10}^1 & \cdots & o_{10}^5 \end{bmatrix}-\begin{bmatrix} 0.2 & \cdots & 0.2 \\ \vdots & \ddots & \vdots \\ 0.2 & \cdots & 0.2 \end{bmatrix}\sum_{i=1}^{10}\sum_{j=1}^{5} o_i^j\right\|}$$

The numerator is the difference between the actual number of objects being produced at the factories and the expected number, averaged over all objects and all factories. The denominator is the averaged difference between the actual number of objects and a non-informative estimate which assigns an equal expected production number to each object. Consequently the error norm has a value of zero when information in the data is perfect and a value of one when there is no information in the data or it is ignored.

The error norm was calculated across 100 Monte-Carlo datasets for different parameter values relative to their base settings in the actual Cogentry dataset. This is shown in Fig.5. The error norm has a mean value of ~0.5 for the actual data, so it is reasonably informative. By allowing perfect data association, the error value was only reduced very slightly, implying a good data association solution as indicated by Fig.2. However, allowing perfect identification of cargo by the road CCTV cameras significantly reduced the mean error to ~0.2. The error was further reduced by increasing the number of HIGH confidence detections by the road and factory CCTV cameras. This sort of analysis provides a 'sanity check' on our solutions but also provides potentially useful guidelines to an intelligence monitoring system designer.

Analysis was also performed to determine the performance gap between the brute force and approximate solutions. This is displayed in Fig.6 which displays how the error norm varied with the road CCTV cameras probability of detection. The performance differential between the brute force and approximate solutions is reassuringly small. The gap is expected to close further as the size of the dataset increases and computational resources are insufficient to generate 'brute force' solutions for the complete dataset

7 Discussion and Conclusions

This paper has presented a Bayesian probabilistic framework for extracting partially observed signals of intent embedded in clutter. The framework was evaluated on a synthetic problem, Cogentry, which is representative of real problems in terms of the level of uncertainty in the data. The framework was used to develop a set of brute force and approximate algorithms for carrying out two main information fusion functions: data association and

inference. The algorithms were applied to the Cogentry datasets. A Monte-Carlo process and associated metric was established to evaluate their performance under different levels of uncertainty. This validated the algorithms and provided useful system design guidelines.

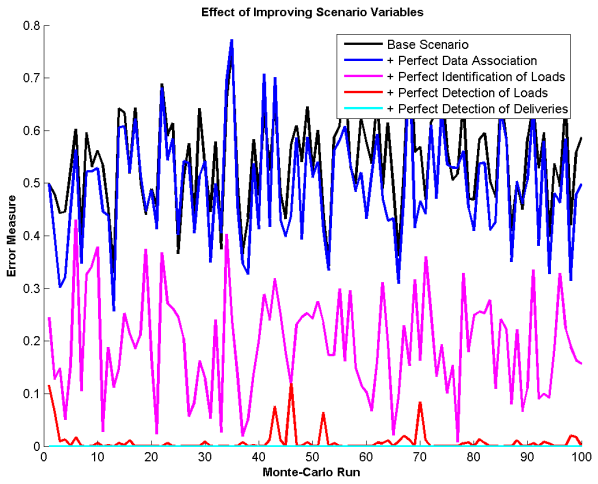
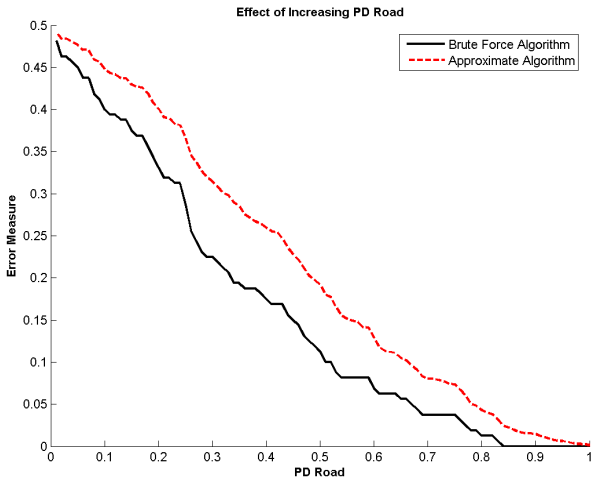


Fig.5 Error norm across 100 Monte-Carlo datasets for the base scenario (black) and various incremental improvements: perfect data association (blue), perfect identification of cargo (magenta), perfect detection of cargo (red), perfect detection of delivery vehicles (cyan).



.

Fig.6. Error norm dependency on the probability of detection for the road CCTV cameras. The error norm has been calculated with the brute force set of algorithms (black curve) and the approximate set (red curve)

The algorithms and analysis presented here are underpinned by a number of assumptions. The Bayesian framework makes these assumptions explicit and that is one of its core strengths. However, the algorithms chose not to assume anything about the order in which components are required to produce objects at the factories. If a known order could be specified, the computational challenge is relaxed because the number of hypotheses that have to be tracked is reduced. This also opens up the solution to well-established sequential estimation algorithms, such as HMMs and particle filters.

The computational framework we have developed is powerful and flexible, but ultimately it would need to be run in the background, supporting human intelligence analysts. This would require the algorithms and software to be embodied in application-specific software tools with the elaborate mathematical details of the algorithms hidden from the analyst. This effort should form a key part of any subsequent work which builds on the results of this paper.

Acknowledgements

The authors would like to thank Dstl (in particular Rob Young, Jonathan Barker, and Paul Thomas) for supplying the Cogentry datasets and providing useful feedback on our results and analysis.

References

[1] M.B. Hurley, *Track association with Bayesian probability theory*, MIT Lincoln Laboratory Technical Report 1085, October, 2003.

[2] L.D. Stone, T.M. Tran, and M.L. Williams, *Improvement in track-to-track association from using an adaptive threshold*, Proceedings of the 12th International Conference in Information Fusion (Fusion 09), Seattle WA, July 2009

[3] M.L. Miller, H.S. Stone, and I.J. Cox, *Optimizing Murty's ranked assignment method*, IEEE Trans. On Aerospace and Electronic Systems, 33(3), pp. 851-862, 1997

[4] Ranked Assignment Code. Available from http://www.cs.ucl.ac.uk/staff/ingemar

[5] S. Gold and A. Rangarajan, *A graduated assignment algorithm for graph matching*, IEEE Trans. On Pattern Analysis and Machine Recognition, 18(4):377-388, 1996

[6] J. Shin, L. Guibas, and F. Zhao, *Distributed algorithm for managing multi-target identities in wireless ad-hoc sensor networks*, 2nd Int'l Workshop on Information Processing in Sensor Networks (IPSN), pp. 223-238, 2003

[7] M.H. DeGroot and M.J. Schervish, *Probability and Statistics*, 3rd Edition, Addison Wesley, Chapter 5, 2001.