Robust approaches for the data association problem

Hassene Aissi, Daniel Vanderpooten
Paris-Dauphine University
Place du Marchal de lattre de Tassigny
75775, Paris
France
{aissi,vdp}@lamsade.dauphine.fr

Jean Michel Vanpeperstraete
Thales Airborne Systems
1 Boulevard Jean - Moulin
78852, Elancourt
France
jean-michel.vanpeperstraete@fr.thalesgroup.com

Abstract—We consider the problem of data association which is central in multi-target tracking. Mathematically, this problem can be stated as a multidimensional assignment problem. Due to various sources of imprecision, the optimal solution of the multidimensional assignment problem can have low percentage of correct association. In order to improve the association correctness, we propose to evaluate the objective function coefficients by intervals instead of point values. This way, we obtain an interval programming problem. Different strategies are proposed to tackle this new problem. Numerical examples for demonstrating the effectiveness of our approach are presented.

I. INTRODUCTION

In the context of modern surveillance systems, the goal is to arrive at a precise assessment of the situation through the combination of data from multiple sensors offering various information. This process refers to the subject of information fusion. A central problem in information fusion is the data association problem which corresponds to partitioning observations into false alarms and tracks in such a way that their states can be estimated. A common surveillance configuration is described by a sequence of scans of a given surveillance region by several sensors. These observations are arranged into \( m \) sets of observations. The sensors provide kinematic information such as range, azimuth, and elevation.

Mathematically, the problem can be formulated as a multidimensional assignment problem where decision variables correspond to elementary associations and the objective is to maximize the likelihood that the associations correspond to targets [1]. Any feasible solution of this problem corresponds to a potential association hypothesis. For \( m \geq 3 \), the multidimensional assignment problem is \( NP \)-hard. Many heuristic algorithms have been proposed to find an approximate solution, such as Lagrangian relaxation [2], greedy rounding adaptive search (GRASP) [3], genetic algorithms [4] and linear relaxation and rounding techniques [5]. Moreover, in many cases, it is possible to resort to gating techniques [6], [7] which drastically reduce the number of decisions variables and make it possible to solve the problem optimally.

Even if a large part of the literature is devoted to this aspect, solving efficiently the multidimensional assignment problem is not the only challenge for data association problems. Indeed, the quality of near-optimal, or even optimal, solution may vary considerably depending on the context. In sparse configurations or with highly accurate sensors, the model behaves well and the optimal, or even an approximate solution, often has an acceptable percentage of correct associations. On the other hand, in medium or high density configurations or with sensors of low accuracy, the model behaves poorly; there is ambiguity due to similarity of kinematic likelihoods. The optimal solution can have a poor association correctness while the correct solution can be largely suboptimal.

The optimal solution of the model is supposed to be the most likely solution. As the complexity of the observed situations increases, the number of ambiguous elementary associations increases also. Since such associations get a high likelihood within the model, it usually happens that many solutions get an overall likelihood very close to the overall likelihood of the optimal solution. In such cases, any of these solutions, including the optimal one, could pretend to be the correct association hypothesis. Therefore, it seems more reasonable to try to build this hypothesis by considering several candidate solutions rather than by selecting only one solution, even if it has a slightly better likelihood. The general scheme underlying our robust approaches is based on the generation of a set of good candidate solutions in some sense. Moreover, we take into account the fact that due to various sources of imprecision, it seems more appropriate to evaluate the likelihoods, represented by cost coefficients in the model, using intervals instead of point values. This way, we obtain an interval programming problem. Several approaches to solve this problem will be discussed. Among these, we describe and experiment a min-max regret approach which is frequently used when dealing with interval programming problems to construct a robust solution [8], [9], [10]. Other robust approaches are based on the general idea of selecting several good candidate solutions and using a reliability index to derive an association hypothesis.

The reliability of any elementary association is computed on the basis of a robustness concept stating that elementary associations which belong frequently to solutions that are part of the set of good candidate solutions should be more reliable. The intuition is that elementary associations in the most likely solutions are the most likely to be correct. In this approach, we actually use a refined frequency index which takes into account the state estimation. Therefore, even if an association does not belong to a given solution, the presence of associations with similar states increases its frequency and consequently its reliability. This index will serve as an alternative criterion for the data association problem to construct robust solutions.
which are expected to obtain a better association correctness than the solution maximizing the likelihood criterion.

The focus of this paper is to describe robustness approaches, applicable for a large class of multiple-target tracking applications, that exploit in a different way the information available in the model in order to improve the association correctness. The running time is not our main concern in this paper, as opposed to the considerable interest showed, in recent years, in the design of approximate algorithms for the multidimensional assignment problem. In order to evaluate properly the relevance of the suggested approaches, we need to compute optimal solutions. It appears, however, that through the use of gating techniques, optimal solutions could be obtained in reasonable times for most of the approaches we tested.

In section II, we briefly present the classical problem formulation and an extension involving interval coefficients. Section III describes the use of a min-max regret approach as a classical approach for interval programming problems. In section IV, we present an alternative robustness approach. Section V describes experimental results on a 3 sensors passive data association problem. Finally, section VI provides concluding remarks.

II. PROBLEM FORMULATION

The goal of this section is to describe briefly the formulation of the data association problem as a combinatorial optimization problem. Good reviews can be found in [11] and [12]. Given m sets of observations, let \( Z^j = \{ z^j_i \} \) denote the \( j \)th set of observations where \( z^j_i \) is a noise contaminated observation. A dummy measurement \( i_j = 0 \) is added to model a missed detection or a false alarm which is an observation that does not match with any other observations. Let \( Z = \prod_{j=1}^{m} Z^j \) be the set of all elementary associations that can be derived from the \( m \) sets of observations. An elementary association \( z_{i_1, \ldots, i_m} \) is an \( m \)-tuple \( (i_1, \ldots, i_m) \) corresponding to a potential target.

A feasible solution \( x \) to the data association problem, corresponding to an association hypothesis, consists of a set of elementary associations in which every non dummy measurement from every set is assigned to exactly one association. There is no constraint on dummy observations.

A likelihood \( p(z_{i_1, \ldots, i_m}) \) is computed to determine the probability that the \( m \)-tuple \( (i_1, \ldots, i_m) \) fits a given target \( t \). The overall likelihood of a solution \( x \) is computed by the product of the individual association likelihoods.

The objective is to determine a solution corresponding to the maximum overall likelihood. In order to have a linear objective function, a logarithmic transformation is applied. The cost of association \( z_{i_1, \ldots, i_m} \) is then defined as

\[
c_i = -\ln p(z_{i_1, \ldots, i_m}) \quad (1)
\]

The multidimensional assignment formulation of the data association problem is obtained as follows

\[
\min \quad cx = \sum_{i_1=0}^{n_1} \sum_{i_m=0}^{n_m} c_{i_1, \ldots, i_m} x_{i_1, \ldots, i_m} \quad (2)
\]

subject to:

\[
\sum_{i_2=0}^{n_2} \ldots \sum_{i_m=0}^{n_m} x_{i_1, \ldots, i_m} = 1, \quad i_1 = 1, \ldots, n_1
\]

\[
\sum_{i_1=0}^{n_1} \ldots \sum_{i_{j-1}=0}^{n_{j-1}} \sum_{i_{j+1}=0}^{n_{j+1}} \ldots \sum_{i_m=0}^{n_m} x_{i_1, \ldots, i_m} = 1,
\]

\[
i_j = 1, \ldots, n_j, \text{ and } j = 2, \ldots, m-1
\]

\[
\sum_{i_1=0}^{n_1} \ldots \sum_{i_{m-1}=0}^{n_{m-1}} x_{i_1, \ldots, i_{m-1}} = 1, \quad i_m = 1, \ldots, n_m
\]

where \( n_j \) denotes the cardinality of set \( Z^j \) and \( x_{i_1, \ldots, i_m} \) are binary decision variables such that \( x_{i_1, \ldots, i_m} = 1 \) if association \( z_{i_1, \ldots, i_m} \) is associated with a potential target and \( x_{i_1, \ldots, i_m} = 0 \) otherwise.

The problem size can be reduced considerably by applying gating techniques in order to prevent implausible associations [6], [7]. More precisely, some associations are so unlikely that the corresponding variables are never considered (see section V for an illustration in the context of passive data association). Then, exact algorithms can be applied successfully in practice. However, even when computing an optimal solution, its association correctness can be poor, as illustrated by our experiments which show that for medium density situations (10 targets), the optimal solution of model (2) detects, in average, less than \( \frac{1}{4} \) of the correct associations (see table I).

In complex situations, the model is only partially relevant: point estimations of the objective coefficients are not reliable because sensors precisions are limited.

In this paper, we describe another formulation of the data association problem that aims at reflecting its imprecise nature. Instead of point values for the coefficients \( c_{i_1, \ldots, i_m} \), interval values \( [c_{i_1, \ldots, i_m}, \overline{c}_{i_1, \ldots, i_m}] \) are considered so as to integrate explicitly the imprecision in the estimation of the various model parameters (confidence interval of states of targets, probability of detection and the probability of false alarms). Thus, we obtain an interval programming problem where each association can have any cost value in its interval independently of the other associations. A possible assignment of cost values to each association is called a scenario. A standard approach to handle an interval programming problem is to solve its associated min-max regret version [9].

III. MIN-MAX REGRET APPROACH

Min-max regret optimization deals with optimization problems where some numerical parameters occurring in the objective function are imprecise or uncertain. It aims at finding a feasible solution minimizing, over all possible scenarios, the maximum deviation of the value of the solution from the optimal value of the corresponding scenario. Such a solution
is often referred to as robust [8], [9], [10]. More formally, let us consider an instance $I$ of a minimization problem $P$ and let $X$ and $S$ denote, respectively, the set of feasible solutions and the set of scenarios. The regret $R(x, s)$ of a solution $x \in X$ under a scenario $s \in S$ is defined as

$$R(x, s) = c^s x - c^s x^{**}$$  \hspace{1cm} (3)$$

where $x^{**}$ denotes the optimal solution corresponding to scenario $s$.

The **maximum regret** $R_{\text{max}}(x)$ of solution $x$ is defined as

$$R_{\text{max}}(x) = \max_{s \in S} R(x, s)$$  \hspace{1cm} (4)$$

The min-max regret optimization problem corresponding to $P$ consists of finding a solution $x$ minimizing the maximum regret $R_{\text{max}}(x)$, or equivalently

$$\min_{x \in X} R_{\text{max}}(x) = \min_{x \in X} \max_{s \in S} \{c^s x - c^s x^{**}\}$$  \hspace{1cm} (5)$$

In [13] a relaxation procedure to solve min-max problems is presented. In [14] and [15] this procedure is extended to solve min-max regret problem for linear programming problems with an interval objective function. The main idea is to solve at each iteration $h$ of the algorithm a relaxation of the min-max regret problem by considering a subset $S^h \subseteq S$. In this case (5) can be formulated equivalently as

$$\min_{x \in X} \max_{s \in S} R(x, s)$$  \hspace{1cm} (6)$$

subject to:

$$c^s x - c^s x^{**} \leq r, \forall s \in S^h$$

$$x \in X$$

Let $(r^h, \bar{R}^h)$ denote the optimal solution of (5) at iteration $h$. The scenario $s^-(\bar{R}^h)$ corresponding to the maximum regret of $\bar{R}^h$ is then computed, i.e.

$$s^-(\bar{R}^h) \in \arg \max_{s \in S} R(\bar{R}^h, s)$$  \hspace{1cm} (7)$$

The subset of scenarios $S^h$ is updated

$$S^{h+1} = S^h \cup \{ s^- (\bar{R}^h) \}$$  \hspace{1cm} (8)$$

and a new constraint, called cut 1, is generated and added to the relaxed problem at iteration $h+1$:

$$c^s (\bar{R}^h)x - c^s (\bar{R}^h)x^{**} (\bar{R}^h) \leq r$$  \hspace{1cm} (9)$$

Problem (6) is solved again until finding a solution $\bar{R}^n$ such that $R_{\text{max}}(\bar{R}^n) = r^{**}$. The number of iterations $n$ is finite.

The algorithm performance can be drastically improved by replacing cut (9) in (6) by the following cut:

$$\tau x - c^- (x)\bar{R}^{**} (\bar{R}^h) \leq r$$  \hspace{1cm} (10)$$

where

$$c^- (x) = \underline{c} + (\bar{c} - \underline{c})x$$  \hspace{1cm} (11)$$

and costs $\tau$ and $\underline{c}$ correspond, respectively, to the scenario obtained by assigning to all variables costs corresponding to the upper bound, respectively lower bound, of their respective costs intervals. This cut will be called cut 2.

**IV. AN ALTERNATIVE ROBUST APPROACH**

An alternative idea to handle imprecision in model (2) is to determine a set $C \subseteq X$ of candidate good solutions. Let $k$ denotes its cardinality. The reliability of all elementary associations is then computed on the basis of their respective frequency in $C$. This index will serve as an alternative criterion to the data association problem and will be used to construct robust solutions.

In section IV-A we show various strategies to generate set $C$ and in section IV-B we explain how to construct the reliability index.

**A. Construction of a set of candidate solutions**

The intuition behind constructing the set $C$ of good candidate solutions is to try to include solutions which are as legitimate as the optimal solution considering the imprecision inherent in the data process. A natural strategy to generate multiple solutions is to consider the $k$ best solutions to the multidimensional assignment problem. In [16], approximate $k$ best solutions are computed based on the algorithm described in [17]. The latter is general and can be applied to other combinatorial optimization problems as well. It requires the resolution of several restricted instances of the original problem which are solved in [16] using a Lagrangian relaxation heuristic.

In the interval data case, the strategy consists of generating a set of scenarios representing the cartesian product of all cost intervals. The best solutions to these scenarios constitute potential solutions since each scenario corresponds to an admissible cost assignment to each association. An easy way to do this is to draw for each coefficient $c(i_1, ..., i_m)$ a value from a uniform distribution in $[\underline{c}(i_1, ..., i_m), \bar{c}(i_1, ..., i_m)]$ and by repeating this to obtain a set of $k$ scenarios. The dispersion of the resulting scenarios may, however, be rather poor. In order to overcome this, we first generate a large number $K$ of scenarios, and filter them in order to obtain the most dispersed scenarios among these $K$ scenarios. Forward filtering techniques have been devised for this purpose [18]. Among the techniques discussed there, we used the first point outside the neighborhood technique (FiPO).

**B. A reliability index**

In [16], an index quantifying the reliability of the associations is computed using a JPDA-like technique. More precisely, this index depends on the frequency of the association in the $k$ best solutions and the exponential of the cost deviations of these solutions from the best one. If target density is high or sensors are imprecise, false associations can have much better costs than the correct ones, and some correct associations may not belong to the $k$ best solutions. Furthermore, if a correct association belongs to bad ranked solutions, its robustness, when computed this way, is expected to be very small.

Since the final goal of data association is to estimate the states of targets, we consider a new reliability index $R(z_{i_1, ..., i_m})$ of each association $z_{i_1, ..., i_m}$ which is computed
on the basis of the frequency of association $z_{i_1,...,i_m}$ in the set $C$ of candidate solutions and its state estimation. More precisely, considering any solution $x \in C$, if an elementary association $z_{i_1,...,i_m}$ belongs to $x$ then its occurrence indicator verifies $\mathcal{R}(z_{i_1,...,i_m}, x) = 1$. Otherwise, it has a value in $[0, 1]$ depending on the similarity between the state corresponding to $z_{i_1,...,i_m}$ and the states corresponding to all elementary associations in $x$. Formally, let $\rho_{i_1,...,i_m}$ denote the state vector of dimension $d$ corresponding to elementary association $z_{i_1,...,i_m}$ and $\Sigma_{i_1,...,i_m}$ its covariance matrix. The occurrence indicator $\mathcal{R}(z_{i_1,...,i_m}, x)$ is defined as follows:

$$\mathcal{R}(z_{i_1,...,i_m}, x) = Pr(\eta \geq d^*(z_{i_1,...,i_m}, x))$$  \hfill (12)

where $\eta$ corresponds to a random variable having a $\chi^2$ distribution with $d$ degrees of freedom and

$$d^*(z_{i_1,...,i_m}, x) = \min_{z_{i_1',...i_m'} \in x} d(\rho_{i_1,...,i_m}, \rho_{i_1',...i_m'})$$  \hfill (13)

where

$$d(\rho_{i_1,...,i_m}, \rho_{i_1',...i_m'}) = (\rho_{i_1,...,i_m} - \rho_{i_1',...i_m'})^T(\Sigma_{i_1,...,i_m} + \Sigma_{i_1',...i_m'})^{-1}(\rho_{i_1,...,i_m} - \rho_{i_1',...i_m'})$$  \hfill (14)

refers to the Mahalanobis distance [19]. The reliability $\mathcal{R}(z_{i_1,...,i_m})$ of association $z_{i_1,...,i_m}$ is then given as follows:

$$\mathcal{R}(z_{i_1,...,i_m}) = \sum_{x \in C} \mathcal{R}(z_{i_1,...,i_m}, x)$$  \hfill (15)

By construction, this index varies between 0 and $k$.

After evaluating the reliability of each association, we can construct the optimal solution of a new multidimensional assignment problem where the objective coefficients correspond to the reliability index and the goal is to maximize the global reliability. This solution will be called robust solution.

V. EXPERIMENTAL RESULTS

The robustness approaches, as described in the previous sections, are applied to the passive data association problem with three passive sensors (see Fig. 1). Each measurement corresponds to a line-of-sight angle from a sensor to a potential target. The presence of multiple targets creates a number of false associations, called ghosts, which make the data association problem difficult. Configurations presented in [1] were tested and 50 random runs were made. The average angular targets separation to measurement error standard deviation is fixed at 5. Two angular configurations were tested: a difficult one with 10 targets and an easier one with 5 targets. The probability of detection is set to 1 and the probability of false alarm is set to 0. Thus, the association difficulties arise mainly from ghosting phenomena and the density of targets.

In order to alleviate the resolution of the multidimensional assignment problem, gating tests were performed to eliminate unlikely association $z_{i_1,i_2,i_3}$ if

$$\sum_{j=1}^{3} \left(\frac{z_{i_j} - \hat{\theta}_{i_j}}{\sigma_j}\right)^2 \geq \delta$$  \hfill (16)

where $\hat{\theta}_{i_j}$ refers to an estimation of the line-of-sight angle from sensor $j$ to the potential target, $\sigma_j$ corresponds to error standard deviation of sensor $j$, and $\delta$ verifies:

$$Pr(\eta \leq \delta) = 1 - \alpha$$  \hfill (17)

where $\eta$ is a random variable having a $\chi^2$ distribution with 1 degree of freedom and $\alpha$ is the desired precision.

For the determination of interval costs, the lower bound $\xi_{i_1,i_2,i_3}$ corresponds to the cost obtained using maximum likelihood estimator and the upper bound $\tau_{i_1,i_2,i_3}$ is computed as follows:

$$\tau_{i_1,i_2,i_3} = \xi_{i_1,i_2,i_3} + \delta'$$  \hfill (18)

where $\delta'$ verifies

$$Pr(\eta \leq \delta') = 1 - \alpha$$  \hfill (19)

where $\eta$ is a random variable having a $\chi^2$ distribution with 2 degrees of freedom.

Using gating tests, the running time to compute an optimal solution to the maximum likelihood criterion (model 2) is, in average, 0.69 seconds for configurations with 5 targets and 0.72 seconds for configurations with 10 targets. On the other hand, the running time of the relaxation procedure for solving the min-max regret version using cut 2, as described in section III, is in average 1.23 seconds for configurations with 5 targets and 8.16 seconds for configurations with 10 targets. When using cut 1, the running time is in average 23.09 seconds for configurations with 5 targets and exceeds 500 seconds for configurations with 10 targets. All these computations were carried out on a Pentium4 2.5 GHz machine with 512MB RAM using software ILOG CPLEX 8.1.

Tables I-III show the performance in terms of percentage of correct associations of the various approaches described earlier. The optimal solution to the maximum likelihood criterion will serve as a reference. In these tables, COR represents the average percentage of correct associations for each type of solutions, and IMP represents the percentage of improvement over the reference solution. All averages are computed over 50 randomly generated configurations.

The optimal solution to the maximum likelihood criterion has an acceptable association accuracy in configurations with
low target density. However, when target density gets higher, the optimal solution includes ghosts with low costs and the correctness is poor, as shown in Table I. The optimal solution to the min-max regret version of the three-dimensional assignment problem improves the correctness both in easy (5 targets) and complex (10 targets) configurations (13%). These first results suggest that using interval costs reflecting imprecision is beneficial.

Tables II-III exhibit the correctness of the robust solutions, which are computed, as described in section IV-B, after generating a set \( C \) of candidate solutions of size \( k = 30, 50 \) or 100 and computing the reliability of all associations that passed the gating test. We can see that, except for the approach based on the \( k \) best solutions, almost all these strategies improve the association correctness. Furthermore, these results are stable and do not depend much on \( k \) since the correctness is almost the same in the three cases.

### TABLE I

<table>
<thead>
<tr>
<th>% Association correctness</th>
<th>5 targets</th>
<th>10 targets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COR</td>
<td>IMP</td>
</tr>
<tr>
<td><strong>Optimal</strong></td>
<td>62</td>
<td>-</td>
</tr>
<tr>
<td><strong>Min-max regret</strong></td>
<td>70</td>
<td>13</td>
</tr>
</tbody>
</table>

### TABLE II

<table>
<thead>
<tr>
<th>% Association correctness (5 targets)</th>
<th>30 sol.</th>
<th>50 sol.</th>
<th>100 sol.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COR</td>
<td>IMP</td>
<td>COR</td>
</tr>
<tr>
<td><strong>k Best</strong></td>
<td>52</td>
<td>-16</td>
<td>40</td>
</tr>
<tr>
<td>Uniform</td>
<td>69</td>
<td>11</td>
<td>71</td>
</tr>
<tr>
<td>FiPO</td>
<td>70</td>
<td>13</td>
<td>70</td>
</tr>
</tbody>
</table>

### TABLE III

<table>
<thead>
<tr>
<th>% Association correctness (10 targets)</th>
<th>30 sol.</th>
<th>50 sol.</th>
<th>100 sol.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COR</td>
<td>IMP</td>
<td>COR</td>
</tr>
<tr>
<td><strong>k Best</strong></td>
<td>35</td>
<td>52</td>
<td>34</td>
</tr>
<tr>
<td>Uniform</td>
<td>34</td>
<td>48</td>
<td>37</td>
</tr>
<tr>
<td>FiPO</td>
<td>32</td>
<td>39</td>
<td>30</td>
</tr>
</tbody>
</table>

From these first experiments, it appears clearly that robustness approaches involving the selection of a set of candidate solutions based on random cost generation give stable results and dominate the usual, min-max regret, and \( k \) best approaches.

### VI. Conclusions

We investigated alternative approaches to the usual strategy consisting of taking as association hypothesis the (near) optimal solution to a multidimensional assignment problem where the objective is to maximize the likelihood that all the elementary associations in this solution are correct. Using the same information as in the original model, but exploiting interval costs, we can substantially improve the quality of association, especially in difficult configurations. Numerical experiments show an improvement with respect to the classical optimization lying between 10% and 61%.

Our methodology is general enough and can be applicable to a large class of target tracking problems including the case where costs are based on different types of information.

### Acknowledgements

This work has been partially funded by grant CNRS/CGRI-FNRS number 18 227.

### References


