ROBIN: a platform for evaluating Automatic Target Recognition algorithms. Part 2: protocols used for evaluating algorithms and results obtained on the SAGEM DS database

D. Duclos, J. Lonnoy, Q. Guillerm SAGEM DS[1], F. Jurie CNRS-INRIA[2], S. Herbin ONERA[3], E.D'Angelo DGA/CEP[4]

ABSTRACT

Over the five past years, the computer vision community has explored many different avenues of research for Automatic Target Recognition. Noticeable advances have been made and we are now in the situation where large-scale evaluations of ATR technologies have to be carried out, to determine what the limitations of the recently proposed methods are and to determine the best directions for future works.

ROBIN, which is a project funded by the French Ministry of Defence and by the French Ministry of Research, has the ambition of being a new reference for benchmarking ATR algorithms in operational contexts. This project, headed by major companies and research centers involved in Computer Vision R&D in the field of Defense (Bertin Technologies, CNES, ECA, DGA, EADS, INRIA, ONERA, MBDA, SAGEM, THALES) recently released a large dataset of several thousands of hand-annotated infrared and RGB images of different targets in different situations.

Setting up an evaluation campaign requires us to define, accurately and carefully, sets of data (both for training ATR algorithms and for their evaluation), tasks to be evaluated, and finally protocols and metrics for the evaluation. ROBIN offers interesting contributions to each one of these three points.

This paper first describes, justifies and defines the set of functions used in the ROBIN competitions and relevant for evaluating ATR algorithms (Detection, Localization, Recognition and Identification). It also defines the metrics and the protocol used for evaluating these functions. In the second part of the paper, the results obtained by several state-of-the-art algorithms on the SAGEM DS database (a subpart of ROBIN) are presented and discussed.

Keywords: Evaluation, Pattern recognition, ATR, benchmark, database

1. INTRODUCTION

Evaluation and benchmarking campaigns are of prime interest for research funding agencies like the Délégation Générale de l'Armement of the French Ministry of Defence and for industrial companies. They allow measuring the state-of-the-art performances of a given technology, in order to assess the efficiency of their funding policy (by measuring progress) or to check the readiness of a technology for a marketed application. Objective, well-spread performance evaluation methodologies can even lead to certification procedures for performance critical applications, such as computer-assisted surgery, vehicle guidance, etc.

[1] SAGEM DS 72-74 Rue de la Tour Billy BP 72 95101 Argenteuil Cedex France (daniel.duclos@sagem.com, quentin.guillerm@sagem.com, jacques.lonnoy@sagem.com)

[2] INRIA LEAR Project-655 avenue de l'Europe - 38334 Saint Ismier Cedex – France (frederic.jurie@inrialpes.fr)

[3] ONERA Department of Modeling and Information Processing, BP 72, 29, av. de la Division Leclerc 92322 CHATILLON CEDEX, France (Stephane.Herbin@onera.fr)

[4] DGA/CEP (emmanuel.dangelo@etca.fr – emmanuel.angelo@cmla.ens-cachan.fr)

There are several ways to assess image processing algorithms. Evaluation can be done analytically using theoretical or empirical performance prediction models, or empirically by measuring the output of algorithms with some specific criteria on an evaluation images set. Today, the most common way is to measure the distance between the output of the algorithm and an ideal result or ground truth using a metric. This leads to data-driven evaluation campaigns, for which the choice of the evaluation datasets and metrics are critical.

The ROBIN platform (platform for evaluating Automatic Target Recognition algorithms) falls into this latter category. ROBIN is a project funded by the French Ministry of Defense and by the French Ministry of Research which aims at being a new reference for benchmarking ATR algorithms in operational contexts. This project, headed by major companies and research centers involved in Computer Vision R&D in the field of Defence recently released a large dataset of several thousands of hand-annotated infrared and RGB images of different targets in different situations (Presented in [3]).

The objective of this paper is to give an overview of the protocols defined for this evaluation campaign, as well as the results obtained by the competitors (datasets are presented in [3]). Having good specifications of the tests, having relevant metrics and evaluation procedures are the cornerstones of any benchmarking campaign. If the images are too far from the operational context, then the lessons learned during the evaluation will not allow for a accurate technology readiness assessment. If images are too difficult to process by all competitors, then the only lesson will be that technology is not yet ready. The gold reference (or ground truth) must also be carefully created, since errors would lead to biased results.

Metrics should also be chosen according to the task. For example, for autonomous ground vehicle guidance it is preferable to miss 10% of the road in a picture than to find a road 10% too big, since this latter case means confusion between the road and its environment and could lead to vehicle damages.

The paper is organized as follows. In the next section we show that the evaluation of ATR algorithms can be divided up into two tasks, *detection* and *categorization*. General definitions are also given. In section 3 and 4 we explain how to measure the performance obtained by algorithms on these 2 tasks. Section 6 is devoted to the organization of the competitions. At last, a short presentation of the results is given in Section 7, immediately followed by our conclusions.

2. EVALUATION OF OBJECT RECOGNITION

Detection and categorization

One of the fundamental questions of image interpretation is to devise where and what are the objects of interest in the observed scene. Although the two questions of object localization – the *where* part – and characterization – the *what* part – are strongly linked, it may appear useful for analytical purposes to divide this overall function into two conditional functions: detection and categorization.

Object detection is a function describing the possible locations of objects of interest, whatever they are. Categorization is the function giving a label to a given location in an image, this label included in a given list of candidates.

The ROBIN competition is devoted to the evaluation of these two fundamental functions of image understanding in several operational contexts.

Databases

The evaluation relies on a series of annotated databases dedicated to the evaluation of problems with various levels of difficulty. Each evaluated problem is associated with a specific database. Data is associated with ground truths describing the location and type of objects occurring in each image, and various auxiliary information available in each context (viewing conditions, pixel resolution...).

An annotated database is a series of images alongside a description of the objects they contain. The annotation of an image I is a list of $N^*(I)$ elements $\{(Y^*_1, Z^*_1), (Y^*_2, Z^*_2)\ldots (Y^*_{N^*(I)}, Z^*_{N^*(I)})\}$ where:

- $Y_i \in \mathbf{Y}$ is the category of object i;
- $Z_i \in \mathbf{Z}$ is a description of its location and geometry in the image.

The space $\mathbf{Y}$ defines the set of all possible categories. It may be structured in a hierarchy, although the evaluation of decision making among structured categories is not the main objective of ROBIN.

The space $\mathbf{Z}$ describes the type of geometric description used to characterize object location and, when available, object extension and pose.

In the ROBIN competition, two different databases are involved: a learning database **B** used to specify the type of data and objects to be detected or categorized, and a testing database **T** used for the evaluation of the implemented functions.

Evaluation and fulfillment of requirements

Performance evaluation is a procedure designed to quantify the disagreement between a given implemented function and several specified requirements.

Requirements can be divided into two types:

- A *functional* requirement describes what the expected value produced by the function when is given a specific input;
- An *operational* requirement describes the global constraints that must be satisfied by the implemented function.

Functional requirements

The evaluation of functional requirements agreement is the main objective of ROBIN. A functional requirement is satisfied if the algorithm is able to reproduce the behavior of a function. In ROBIN, the function is known from a training database defining samples:

- The input space = structure and density;
- The output space = set of target categories or object locations;
- The association of input and output values.

A testing database is used to measure the empirical adequacy of the implemented function with its requirement. Its main goal is to provide samples to measure the quality of the association and the algorithm rejection ability.

Operational requirement

To become operational, the implementation of an algorithm must meet several requirements. An implemented function should be characterized, at least, by the following features:

- Hardware features
  - Computation time
  - Memory
- Ease of evolution or adaptation
  - Learning or parameter estimation complexity
  - Increase/modification of category type
  - Modification of input context
  - Parameter tuning
- Robustness
  - Sensitivity to parameter setting
  - Outlier management
- Flexibility
  - Missing information or data management
- Interactivity
  - Functional point tuning
  - Possibility of human intervention or correction

All those requirements are difficult to quantify. In ROBIN, each participant is invited to fill in a questionnaire in order to characterize qualitatively his algorithm along the above operational features.

Operating point setting

Detection and categorization are decision functions, i.e. they make a choice between various potential hypotheses. They both have to trade between various types of errors or risks associated with each individual hypothesis. Very often, the algorithm implementing the function is characterized or controlled by several parameters which, when they vary, are able to set various operating points, i.e. the way the algorithm realizes the global error tradeoff. Characterizing the range of reachable operating points is essential in an evaluation protocol.

Various operating points are of interest for practical applications:

- Detection
  - Good detection only: the algorithm only proposes outputs where an object of interest is present;
  - No missed detection: the algorithm points to every object of interest;
  - Tradeoff
- Categorization
  - Good choice only: the algorithm only proposes adequate hypotheses;
  - Forced choice: the algorithm always proposes a hypothesis;
  - Tradeoff

In practical situations, it is often interesting to evaluate the algorithm's ability to detect anomalies, i.e. data that is clearly outside the operational domain. The evaluation of such a requirement is beyond the scope of ROBIN competitions.

A usual control parameter is a confidence value associated with each hypothesis. This overall decision scheme is a generalization of the standard likelihood ratio test: a decision is produced using a threshold test, i.e. a hypothesis is issued if its confidence is above a given threshold. Each threshold value generates a new operating point.

Formal definitions of the above operating points are presented in the following sections.

3. DETECTION FUNCTION PRESENTATION

The detection function can be formulated the following way: assign to a given image I a list of candidate object locations $\{Z_1,Z_2 \ldots Z_{N(I,\,\lambda)}\}$ where objects of interest and typical data are described by samples from a training database **B**. The function is controlled by a parameter λ for operating point setting. This parameter may be of any type, scalar or multidimensional. If the decision requires a final thresholding, we assume that the only control parameter is a scalar equal to the threshold value.

The number of object locations provided by the algorithm $N(I, \lambda)$ depends on the operating point. The detection function is summarized as:

$$\text{Detection} : \mathbf{B}, I, \lambda \rightarrow \{Z_1,Z_2 \ldots Z_{N(I,\lambda)}\}$$

The output format may be different than the annotation in the training database, e.g. annotations contain precise object boundary whereas algorithm output is the object center in the image.

In ROBIN, acceptable object location descriptions are:

- Access point: (xc, yc)
- Bounding box: ($x$min, $y$min, $x$max, $y$max)

An access point is a simple location assumed to be unequivocally associated with the presence of an object, e.g. the center of a bounding region, the position of a characteristic part.

The bounding box, however, is the favored format output since most of the ground truth descriptions will use it. More detailed outputs such as closed polygons or binary masks are expected to be provided by the participant as well as bounding boxes.

Figure 1: two examples of erroneous detections. Ground truth is in green, detections are in red. Left image: One detection encompasses two regions, but only one object should be counted as correctly detected (TP = 1, N = 1, N* = 3). Right image: two detections are superimposed on a same object. One is counted as true, the other as a false positive (TP = 1, N = 2, N* = 3).

Evaluation metric

In our formulation, evaluating the function of detection is equivalent to comparing two lists:

- $\{Z_1, Z_2 \ldots Z_{N(I,\lambda)} | I, \lambda\}$: the list of $N(I,\lambda)$ locations output by the implemented algorithm and controlled by parameter λ.
- $\{Z^*_1, Z^*_2 \ldots Z^*_{N^*(I)} | I\}$: the list of $N^*(I)$ ground truth locations.

Evaluation is based on counting the true detections (or True Positives) $TP(I,\lambda)$ from the two lists. It depends on the input data I and the control parameter λ assumed to be fixed during a whole evaluation session.

A careful treatment of multiple detections has been carried out in the definition of the evaluation metric (cf. Figure 1). It was found useful to define the number of true positives as the maximum matching size of a bipartite graph whose edges connect the items of the two location lists with an adjacency rule given by a geometric acceptance criterion. A formal definition of the computation scheme and geometric criterion specification can be found in [2].

Global ratios averaging the testing database known as *Precision* and *Recall* are computed from the number of true detections as:

$$\text{Recall}(\lambda) = \frac{\sum_I TP(I,\lambda)}{\sum_I N^*(I)} \qquad \text{Precision}(\lambda) = \frac{\sum_I TP(I,\lambda)}{\sum_I N(I,\lambda)}$$

Recall measures the ratio of detected objects; *Precision* measures the ratio of good decisions. The series of couples (Precision(λ), Recall(λ)) generated for various values of the control parameters λ produces a manifold which is usually represented on a common graph (Figure 2).

Operating points

The role of control parameter λ is to modify internal setting of the algorithm able to select different values of *Precision* and *Recall*. We are especially interested in computing three points:

- *Recall* for maximal *Precision*: R^* characterizing the quality of a "good detection only" regime;
- *Precision* for maximal *Recall*: P^* characterizing the quality of a "no missed detection" regime;
- Equal *Precision* and *Recall*, often called also Equal Error Rate: *EER* characterizing the quality of a good tradeoff regime.

It is expected that each participant is able to tune the algorithm parameters in order to approximately provide those three operating points. Figure 2 shows the positions of those three points on a typical *Precision* vs. *Recall* graph.

The Area Under the Curve (AUC) or Mean Average Precision (MAP) will be computed as another global indicator on the basis of a whole precision/recall curve when available, or on its approximation using several operating points. Its role is to average out all the possible tradeoffs.

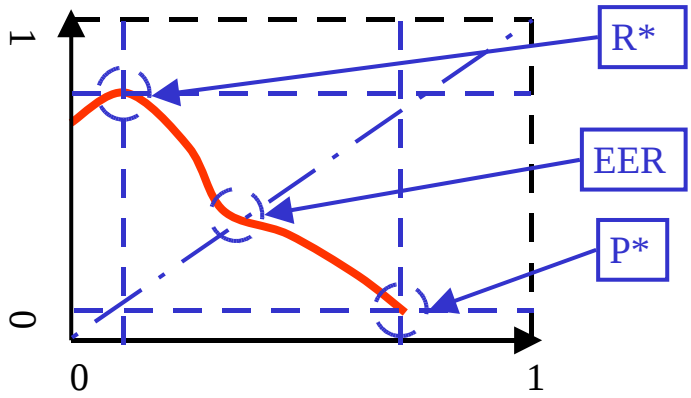

Figure 2: Typical Precision vs. Recall graph and corresponding extreme operating points.

4. CATEGORIZATION FUNCTION PRESENTATION

The function of categorization assigns a category *Y* to a given location *Z* in an image *I*. Samples from the learning database **B** define empirically the category to be associated with each data. The function can be controlled by a parameter **μ** able to set the operating point.

$$\text{Categorization} : \mathbf{B}, I, Z, \boldsymbol{\mu} \rightarrow Y$$

A categorization function is assumed to be able to choose a category in a predefined set of candidates. An "Ambiguous" decision is provided when the algorithm is unable to make a reliable distinction between two or more categories. The background is interpreted as a regular category, with a corresponding label.

Acceptable categorization outputs *Y* are:

- A category from the training database, including *background*;
- An "Ambiguous" decision;

A categorization function may also be able to reject outliers, i.e. the observation of an unknown or novel category. Although related, discrimination and rejection capabilities should be evaluated on separate competitions. The SAGEM database provided in the ROBIN challenge does not include any rejection competition.

Evaluation metric

The evaluation of categorization relies on counting label differences: Y and Y*. We consider various cases according to the type of category labels issued by the algorithm: class index including a *background* category or ambiguous ("A"). It is assumed in the competitions that all target classes included in the testing database have samples in the learning database ("closed world" assumption).

Discrimination ability is measured by counting misclassification for categories available in the training database from a normalized *confusion matrix* depending on the control parameter **μ**: $\eta\,(c, c^*, \boldsymbol{\mu})$. This matrix describes the distribution of output decisions for each ground truth category.

The diagonal coefficient $\eta\,(c^*, c^*, \boldsymbol{\mu})$ measures the categorization performance for each class in the training database. The coefficient $\eta\,(A, c^*, \boldsymbol{\mu})$ where *A* is the "Ambiguous" category, measures the algorithm level of decision making.

The confusion matrix coefficients are averaged to define the measure $D(\boldsymbol{\mu})$ accounting for a global discriminating capacity:

$$D(\boldsymbol{\mu}) = \sum_{c^*} \pi(c^*)\,\eta(c^*, c*, \boldsymbol{\mu})$$

where $\pi(c)$ is the prior for category c.

Define similarly $U(\boldsymbol{\mu})$ as the average uncertainty rate:

$$U(\boldsymbol{\mu}) = \sum_{c^*} \pi(c^*)\,\eta(\mathrm{A}, c*, \boldsymbol{\mu})$$

measuring the algorithm's ability to postpone a decision under uncertain conditions. In general, favoring an ambiguous decision rate goes with increasing discrimination capacity.

Operating points

The two numbers $D(\boldsymbol{\mu})$ and $U(\boldsymbol{\mu})$ describe the categorization behavior for various operating points controlled by $\boldsymbol{\mu}$. We are interested, as in the detection case, in specific operating points:

- Discrimination at minimal uncertainty rate: $D*$ characterizing the quality of a "good choice only" regime;
- Uncertainty at maximal discrimination rate: $U*$ characterizing the quality of a "forced choice" regime;
- Equal discrimination and uncertainty rate: *EDU* characterizing the quality of a good tradeoff regime.

The discrimination behavior at minimal uncertainty rate ($U \approx 0$) can be analyzed in a confusion matrix revealing the inter-category misclassification errors.

5. ORGANISATION OF ROBIN COMPETITIONS

The organization of Robin competition took more than two years and included different phases with associated tasks.

- Phase 1. The goal of this phase, which lasted about 4 months, was to define the specification of the datasets. It was done by small group of experts from research labs, engineers from the companies and representatives of funding agencies.
- Phase 2. Immediately following the specification of the datasets, the second phase was oriented towards the collection and annotation of the images. All of the data has been specially collected for the competitions (no re-use of existing materials), solving intellectual and industrial property issues. This has been the most difficult, the longest and the most expensive phase of the project. It took more than 14 months from the very first acquisitions to the delivery to competitors. It has been very difficult to obtain clean and consistent annotations; targets can be occluded, small and multiple targets can overlap (like parked cars). We adopted a common XML framework for representing annotations and provided the competitors with tools for reading them.
- Phase 3, done in parallel with Phase 2. The definitions of the metrics were the main task of the third phase. It took 2 months to discuss and formalize them. As we mainly used conventional metrics, they were easy to define and well accepted.
- Phase 4. Once the datasets had been produced, the organizers selected how to use and combine them across different competitions. Each competition was designed to address specific issues (influence of target size, influence of image quality, etc.).
- Phase 5. The competitions. 4 months were given to the competitors for optimizing their software to the data and to produce results. Ground truths for test data were not released to the competitors, and the aim was for the algorithms to find what they supposed to be the ground truth.
- Phase 6. Final workshop. In June 2007, a workshop of about 60 people came to present the results and to comment on competitions.

6. RESULTS OBTAINED ON SAGEM DS DATABASE

The SAGEM database competitions consisted of three object detection challenges (person, car and landmark) and one object categorization challenge (three models of car) [3]. Three competitors tested their approach on this database. Car

detection was the only challenge evaluated on more than one competitor. Since the aim of the challenge was not to rank various teams but rather give a hint of the state of the art algorithms in operational object recognition, the competitor names are not mentioned in this article.

Detection

The results of the detection challenge are shown on .

An object is assumed correctly detected if the center of its bounding box coincides with the actual object with a tolerance in localization equal to ±25% of the ground truth bounding box size.

Only one of the competitors gives acceptable results on the vehicle detection challenge. The others provide results with erratic locations which we suspect is a lack of parameter tuning.

The images contained in the database show very different viewing conditions and contexts (Figure 3): object width varies from 20 to 250 pixels, viewing direction is ground based or aerial, some scenes have many occluded objects, other scenes have one single object… For algorithms relying on a heavy learning phase, such a variety seems difficult to handle in a single "one pass" algorithm.

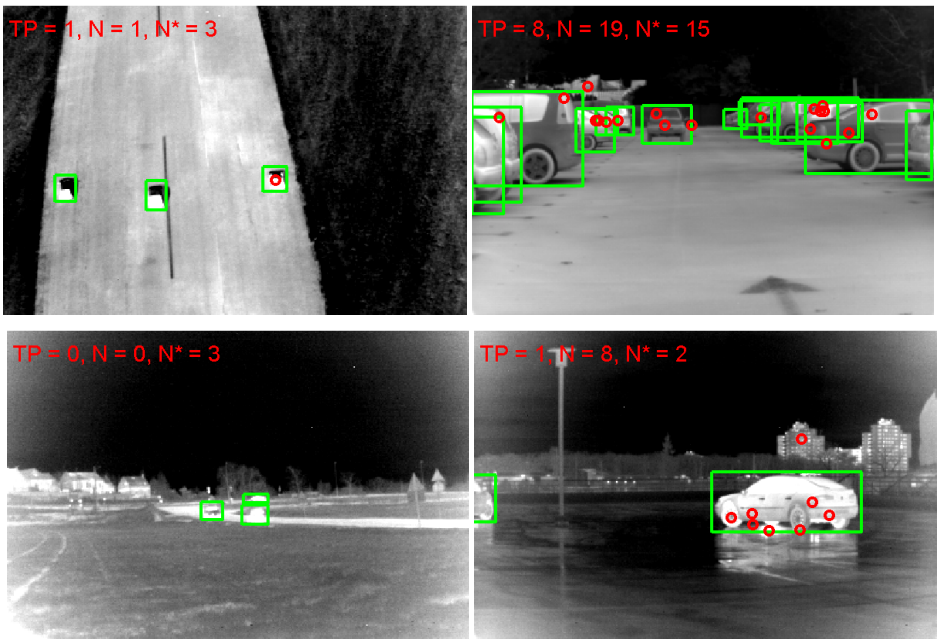

Figure 3: samples of the SAGEM database and car detection results of one of the competitors.

Table 1: results of detection competitions (in percentage).

Object	*Recall(P*)*	*EER*	*AMP*
Vehicle(1)	71.4	34.2	26.6
Vehicle(2)	17.4	10.0	2.3
People	0	0	0
Landmark	1.7	1.7	< 0.1

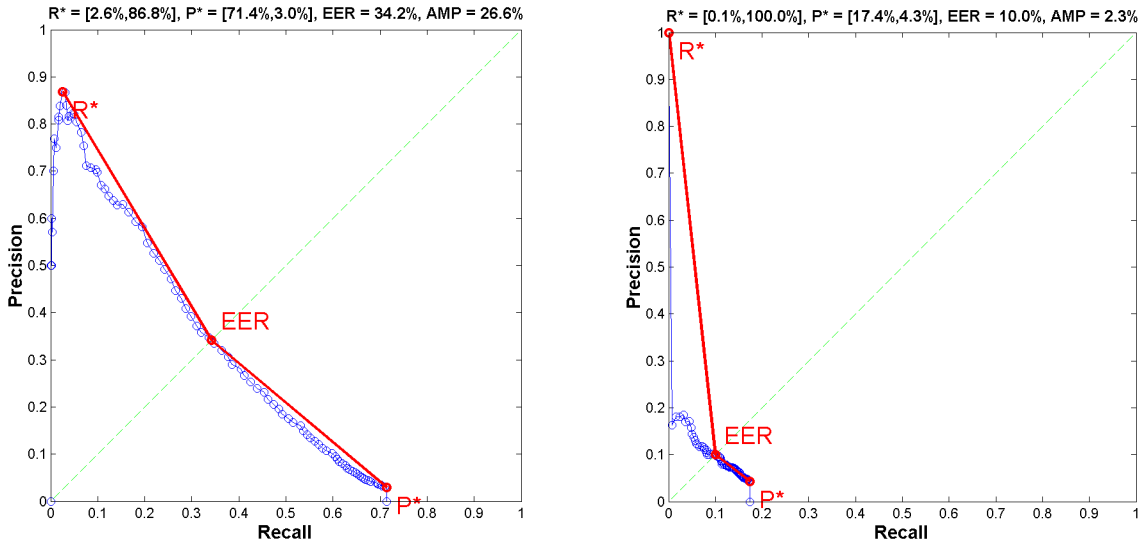

Figure 4: Results of the car detection challenge for competitors 1 and 2.

Categorization

Contrary to most of the other object characterization challenges which are category based (Pascal VOC, Caltech 101 categories), the SAGEM competition is designed to evaluate an algorithm ability to identify a specific object model in various viewing conditions. Indeed, objects show a large variety of appearance, position, size and occlusion level, making the recognition problem very challenging.

The categorization competition was challenged by a single team. Their results are show on Figure 5.

Peugeot 106	**69.6**	10.0	20.4
Citroen xsara	53.0	**29.7**	17.2
Renault express	28.2	2.9	**68.9**

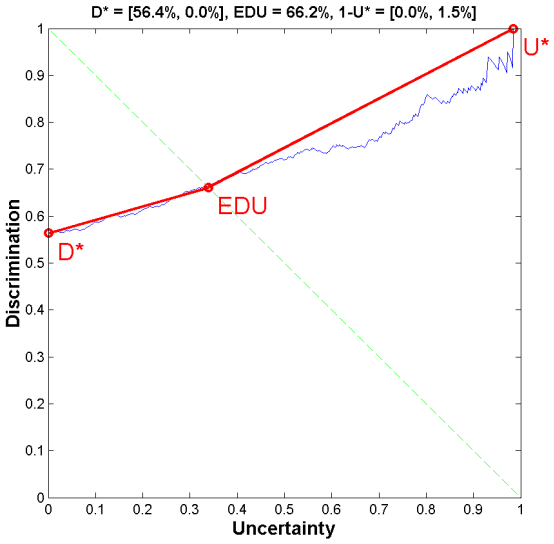

Figure 5: results of the car categorization challenge. Left: discrimination vs. uncertainty graph. Right: confusion matrix at minimal uncertainty rate.

7. CONCLUSION AND PERSPECTIVES

Despite the recent advances that have been seen in the area of object detection and Automatic Target Recognition, it is still very difficult to have resources available for the evaluation of algorithms. Robin is a contribution to this need for comparing and measuring performances of state-of-the-art algorithms. By producing six publicly available different datasets for a total of more than 80.000 images, by providing handmade annotations for all the images, by defining training and tests sets as well as the corresponding metrics, ROBIN has become one of the most advanced resources for the evaluation of ATR algorithms.

The first round of evaluation, which involved more than 40 runs from different competitors, gives us very positive feedback on the datasets. It has also been possible to draw several conclusions on the algorithms themselves. First and most surprisingly, we noticed that algorithms design for standard gray level images gave comparable results on infra red images. Secondly, state-of-the-art algorithms have been found to be still below the requirements of operational situations, especially for small targets, leaving room for improvement.

We believe that resources such as those provided by ROBIN will contribute to the improvement of algorithms in operation situations, and we plan to measure this improvement through other rounds of competition. In 2008 the second round will take place, and we invite all of the potential competitors to take part in the competition.

8. REFERENCES

[1] E. D'Angelo, S. Herbin and M. Ratiéville, *ROBIN Challenge: Competitions*, http://robin.inrialpes.fr/robin_evaluation/downloads/ROBIN_competitions_v2.pdf

[2] E. D'Angelo, S. Herbin and M. Ratiéville, *ROBIN Challenge: Evaluation principles and metrics*, http://robin.inrialpes.fr/robin_evaluation/downloads/ROBIN_metrics_v6.pdf

[3] D. Duclos, J. Lonnoy, Q. Guillerm, F. Jurie, S. Herbin, E. D'Angelo, *ROBIN: a platform for evaluating Automatic Target Recognition algorithms. Part 1: Overview of the project and presentation of the SAGEM DS competition*, SPIE Defense and Security Conference, 2008.