ROBIN Challenge
Evaluation principles and metrics

Emmanuel D'Angelo[1] Stéphane Herbin[2] Matthieu Ratiéville [1]

[1] DGA/CEP/GIP, 16 bis av. Prieur de la Côte d'Or, F-94114 Arcueil cedex
[2] ONERA, BP72 - 29 avenue de la Division Leclerc, F-92322 CHATILLON cedex

http://robin.inrialpes.fr

November 21, 2006

1 Introduction

This document is produced by the ROBIN Project Committee. Its objective is to describe the principle of the evaluation carried out in the ROBIN challenge.

The ROBIN challenge (*Recherche d'OBjets dans les Images Numériques* - Object Detection and Recognition in Digital Images) is part of the Techno-Vision program funded by the French Ministry of Defense and Ministry of Research. It aims to produce image datasets with associated ground truths as well as evaluation protocoles for object detection and recognition tasks. One of the main specifity of ROBIN is to orient evaluation toward operational applicability. Indeed, datasets are provided by industrial companies interested in the field of Computer Vision and Image Processing technologies :

- Bertin Technologies, ECA and SAGEM for infrared ground images;
- MBDA, EADS and THALES for aerial infrared and visible images;
- CNES for SPOT5 spatial images.

A companion document describes those data sets [1].

2 Functions evaluated

One of the first levels of image interpretation is to be able to state where the meaningful objects are, and what they are. The ROBIN competition is devoted to the evaluation of two fundamental functions of image understanding: object detection and categorization.

Each function uses, for its calibration, a series of annotated data describing the expected output requirements. The annotations or ground truths contain the location and type of objects occuring in each image, and various auxiliary information potentially available in each context (viewing conditions, pixel resolution...).

Annotated databases

An annotated database is a series of images associated with a description of the objects they contain. The annotation of an image I is a list of $N^*(I)$ elements:

$$\{(Y_1^*, Z_1^*), (Y_2^*, Z_2^*) \dots (Y_{N^*(I)}^*, Z_{N^*(I)}^*)\}$$

where

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

The space $\mathcal{Y}$ 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 $\mathcal{Z}$ 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.

It is assumed that, in general, two identical images have the same annotation, *i.e.* the database is only exceptionally inconsistent.

Detection

The detection function can be formulated the following way: assign to a given image I a list of candidate object locations $\{Z_1, Z_2 \dots Z_{N(I,\boldsymbol{\lambda})}\}$ where objects of interest and typical data are described by samples from a training database **B**. The function is controlled by a parameter $\boldsymbol{\lambda}$ 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 $N(I, \boldsymbol{\lambda})$ depends on the operating point.

The detection function is summarized as:

$$\text{Detection} : \mathbf{B}, I, \boldsymbol{\lambda} \mapsto \{Z_1, Z_2 \dots Z_{N(I,\boldsymbol{\lambda})}\}$$

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: (x_c, y_c)
- Bounding box: $(x_{min}, y_{min}, x_{max}, y_{max})$

An access point is a simple location assumed to be univocally 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 polygones or binary masks are expected to be provided by the participant also as bounding boxes.

Categorization

The function of categorization assigns to a given location Z in an image I a category Y. The labels of the categories are described further in this document for each dataset. Samples from the learning database **B** define empirically the category to be associated with each data: examples of object pictures will be provided for each category, and we assume that they perform a sufficient sampling of the different categories if one wants to build an object model from them. The function can be controlled by a parameter $\boldsymbol{\mu}$ able to set the operating point.

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

A categorization function has to solve two kinds of problems: discrimination and outlier rejection. Discrimination is the capacity of choosing a category in a predefined set of candidates. Outlier rejection is the capacity of detecting an unknown or novel category. In order to make a distinction between these two situations, the implemented categorization function may generate two special outputs: "Ambiguous" and "Other". An "Ambiguous" decision is provided when the algorithm is unable to make a reliable distinction between two or more categories. An "Other" decision means that the algorithm has been able to detect an object which is not defined in the learning database, *e.g.* a new model of car. 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;
- An "Other" decision.

Although related, discrimination and rejection capabilities will be evaluated on separate competitions.

3 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 given a specific input is;
- An **operational** requirement describes the global constraints that must be satisfied by the implemented function.

Functional requirement

The evaluation of functional requirement 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 **B** defining on samples:

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

A testing database **T** 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 can be characterized 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 characterize qualitatively his algorithm along the above operational features.

4 Evaluation

The main objective of ROBIN project is the quantitative evaluation of functional requirements for detection and categorization. We adopt an empirical point of view, *i.e.* compute statistical quantities on samples from a testing database **T**.

4.1 Detection

4.1.1 Formulation of the Detection Task

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

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

Evaluation is based on counting the true detections (or True Positives - TP) from the two lists. It depends on the input data I and the control parameter $\boldsymbol{\lambda}$ assumed to be fixed during a whole evaluation session. It is defined the following way:

1. Define a geometrical acceptance criterion G between two locations:

$$G : (Z, Z^*) \rightarrow \{0, 1\} \tag{1}$$

   The geometrical acceptance criterion G is defined for each competition and should take into account the relevant tolerance values specific to each context.

2. Define an association matrix $\mathrm{m} \in \{0,1\}^{N(I,\boldsymbol{\lambda}) \times N^*(I)}$ between Z_j and Z_i^* ensuring that at most one detection is assigned to a ground truth location and reciprocally. Formally:

$$\begin{aligned} & G(Z_j, Z_i^*) = 0 \Rightarrow \mathrm{m}(j,i) = 0 \\ & \sum_{j=1}^{N(I,\boldsymbol{\lambda})} \mathrm{m}(j,i) \leq 1 \\ & \sum_{i=1}^{N^*(I)} \mathrm{m}(j,i) \leq 1 \end{aligned} \tag{2}$$

3. Compute the number of true detections as:

$$TP(I, \boldsymbol{\lambda}) = \sum_{i,j} \mathrm{m}(j,i)$$

   If several association matrices satisfy conditions (2), choose the one giving the maximum number of true detections.

This definition can also be interpreted as the size of a *maximal matching* in a bipartite graph whose edges are the items of the two location lists and whose adjacency rule is given by the acceptance criterion G.

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

$$\text{Precision}(\boldsymbol{\lambda}) = \frac{\sum_{I \in \mathbf{T}} TP(I, \boldsymbol{\lambda})}{\sum_{I \in \mathbf{T}} N(I, \boldsymbol{\lambda})} \tag{3}$$

$$\text{Recall}(\boldsymbol{\lambda}) = \frac{\sum_{I \in \mathbf{T}} TP(I, \boldsymbol{\lambda})}{\sum_{I \in \mathbf{T}} N^*(I)} \tag{4}$$

The three numbers $TP(I, \boldsymbol{\lambda})$, $N(I, \boldsymbol{\lambda})$ and $N^*(I)$ will be used to detect difficult images by comparing their values to the average performance given by the ratios (3) and (4).

The role of control parameter $\boldsymbol{\lambda}$ 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^*
- Precision for maximal Recall: P^*
- Equal Precision and Recall, often called also Equal Error Rate: EER.

They characterize different operational conditions. It is expected that each participant is able to tune the parameters of his algorithm in order to approximately provide those three operational points.

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.

4.1.2 Choice of the acceptance criterion G in the case of object and people detection

Except for the CNES dataset, which also deals with extended objects like roads and highways, all the ROBIN Detection challenges aim at finding man made objects and persons. These objects of interest have a limited spatial extension and can be contained in a bounding box. Therefore, the expected output of the algorithms will be a point or (preferred) a bounding box.

Furthermore, since for 5 datasets (Bertin-ECA, EADS, MBDA, Sagem and Thales) the Detection Task is alike, we can propose a unique definition of the geometric acceptance criterion G (see 1) valid over these 5 datasets. In this case, the measures proposed below aim at describing the quality of a given detection result, assessing how close (in location, geometric area and shape) to a ground truth bouding box the detection result is.

- the first measure m_1 is a function of the distance between the centers of the detection and of the reference object. This distance is compared to the dimensions of the ground truth bounding box, because we assume that in this patch extraction task an error of 1 pixel will not have the same consequences if it happens on a 5 or 6 pixel long object than on a 50 pixel long one. It is chosen to vary between 0 when both centers are superimposed, and 1 when they are at an infinite distance;
- the second measure m_2 is the relative deviation between the area of the two bounding boxes. Its value varies between 0 and 1;
- the third measure m_3 compares the aspect of the two bounding boxes. Since we only have to deal with rectangles, we use the height-over-width ratio to characterize them.

The goal of these three measures is to quantify if the system has found something in the right place (m_1), with the good scale (m_2) and a roughly correct shape (m_3).

Formally, a detection result Z (bounding box of center (x_d, y_d), height h_d, width w_d, area $\mathcal{A}_d$) is a candidate good detection if there exists a reference bouding box ground truth object Z^* of center (x_{gt}, y_{gt}), height h_{gt}, width w_{gt}, area $\mathcal{A}_{gt}$ such as :

$$\begin{cases} m_1(Z, Z^*) = \frac{2}{\Pi} \arctan\left(\max\left(\frac{|x_d - x_{gt}|}{w_{gt}}, \frac{|y_d - y_{gt}|}{h_{gt}}\right)\right) & \leqslant \varepsilon_1 \textit{ (localization)} \\ m_2(Z, Z^*) = \frac{|\mathcal{A}_d - \mathcal{A}_{gt}|}{\max(\mathcal{A}_d, \mathcal{A}_{gt})} & \leqslant \varepsilon_2 \textit{ (completeness)} \\ m_3(Z, Z^*) = \frac{2}{\Pi} \arctan\left(|\frac{h_d}{w_d} - \frac{h_{gt}}{w_{gt}}|\right) & \leqslant \varepsilon_3 \textit{ (correctness)} \end{cases}$$

ε_1, ε_2 and ε_3 are three thresholds varying from 0 to 15%.

All the qualities (localization, completeness, correcteness) are valued between 0 (for the best) and 1 (for the worst results) and can be interpreted in terms of distance. The final decision is made on the basis of these three results: a candidate good detection is kept if it satisfies the three criteria of localization, completeness and correcteness simultaneously. This allows us to define the function G we will use :

$$G(Z, Z^*) = \begin{cases} 1 \text{ if } m_1 \leqslant \varepsilon_1, m_2 \leqslant \varepsilon_2 \text{ and } m_3 \leqslant \varepsilon_3 \\ 0 \text{ otherwise} \end{cases}$$

Notes

1. The validation of a detection result as a candidate good detection relies on three thresholds ε_1, ε_2 and ε_3. We will make them vary in order to evaluate their influence (and the bias induced by their values) on the quality of the results. In practice two sets will be used:
   - a rough acceptance set: $\varepsilon_1 = 0.15$, $\varepsilon_2 = 0.5$ and $\varepsilon_3 = 0.15$;
   - a precise acceptance set: $\varepsilon_1 = 0.05$, $\varepsilon_2 = 0.2$ and $\varepsilon_3 = 0.05$.
2. In the case the algorithm only output is a **point**, the validation of a detection result as a candidate good detection will be based only on the localization criterion m_1 and threshold ε_1, *i.e.* $G(Z, Z^*) = 1$ if $m_1 \leqslant \varepsilon_1$ and 0 otherwise.

4.1.3 Detection using confidence coefficient

An optional coefficient α can be added to qualify a priori the confidence in the delivered decision for each candidate detection Z. This coefficient may have various significations such as a posteriori probability or likelihood ratio. If such a coefficient is available, it is assumed that output detections result from a thresholding operation with value $\boldsymbol{\lambda} \in \mathbb{R}$, *i.e.* a detection is output if $\alpha > \boldsymbol{\lambda}$. In this case, the threshold $\boldsymbol{\lambda}$ is the unique control parameter, generating this way a precision/recall curve.

4.2 Categorization

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, ambiguous ("A") or other("O"). Remind that a category is considered "Other" if it is not in the training database **B**.

Categorization is evaluated along two directions, discrimination and rejection.

4.2.1 Discrimination

In a discrimination problem, we are interested in issuing the right category given the set of possible categories. It is assumed that all data contain objects from a given set of categories ("closed world" assumption). In this setting, "Other" and "Ambiguous" outputs are treated equally.

Discrimination ability is measured by counting misclassification for categories available in the training database from a *confusion matrix* depending on the control parameter $\boldsymbol{\mu}$. This matrix describes the repartition of output decisions for each ground truth category. Define its coefficients as:

$$\eta(c, c^*, \boldsymbol{\mu}) = \frac{1}{\mathcal{N}(c^*)} \sum_{n \in \mathbf{T}} \mathbf{1}_c(Y_n).\mathbf{1}_{c^*}(Y_n^*)$$

where Y_n is the category generated by the algorithm on data n, and $\mathbf{1}$ is the indicator function and $\mathcal{N}(c^*)$ is the number of samples in the testing database $\mathbf{T}$ associated with the category c^*,

$$\mathcal{N}(c^*) = \sum_{n \in \mathbf{T}} \mathbf{1}_{c^*}(Y_n^*).$$

The diagonal coefficient $\eta(c^*, c^*, \boldsymbol{\mu})$ measures the categorization performance for each class in the training database. The coefficient $\eta(\mathrm{A}, c^*, \boldsymbol{\mu})$ 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 on class c^*:

$$\pi(c^*) = \frac{\mathcal{N}(c^*)}{\sum_c \mathcal{N}(c)}$$

Define similarly U as the average uncertainty rate:

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

measuring the algorithm ability to postpone decision under uncertainty conditions. In general, favoring ambiguous decision rate goes with increasing discrimination capacity.

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^*
- Uncertainty at maximal discrimination rate: U^*
- Equal discrimination and uncertainty rate: EDU

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

4.2.2 Discrimination using confidence coefficient

A category decision may result from a thresholding operation on a priori confidence. Let $\boldsymbol{\beta} = \{\beta_c\}_{c\in\mathcal{Y}}$ be a vector of confidence for each candidate category in $\mathcal{Y}$. The confidence can be interpreted as a posteriori or class conditional likelihood.

The category is chosen according to the thresholding policy:

$$Y = \begin{cases} \arg\max_{c\in\mathcal{Y}} \beta_c & \text{if} \quad \max_{c\in\mathcal{Y}} \beta_c > \boldsymbol{\mu} \\ \text{A} & \text{if} \quad \max_{c\in\mathcal{Y}} \beta_c \leq \boldsymbol{\mu} \end{cases}$$

where label A refers to the "Ambiguous" decision. The threshold $\boldsymbol{\mu} \in \mathbb{R}$ is the unique control parameter.

Threshold variation generates various operating points. It is the responsibility of the participant to conform to this type of decision where a single scalar is used for thresholding uniformly confidences for all categories.

Note that with this type of policy, an "Other" decision cannot be issued. Rejection ability will not be evaluated in ROBIN if decisions result from a thresholding policy.

4.2.3 Rejection

Rejection is the ability of detecting an unexpected data. It is a binary decision which can be characterized using an average capacity based on a counting policy defined as:

$$\text{Rejection}(\boldsymbol{\mu}) = \frac{1}{\mathcal{N}(\text{O})} \sum_{n\in\mathbf{T}} \mathbf{1}_{\text{O}}(Y_n).\mathbf{1}_{\text{O}}(Y_n^*) \tag{5}$$

which measures the algorithm ability to detect outliers or new categories at an operating point controlled by $\boldsymbol{\mu}$. $\mathcal{N}(\text{O})$ is the number of samples with ground truth category not in the learning database **B**.

As a binary decision, rejection can be characterized globally with a feature such as an "Equal Rejection Rate" where false rejection (*i.e.* erroneously assigning label "Other") and false acceptance (*i.e.* erroneously assigning any label but "Other") have equal rates.

4.2.4 Categorization as multiple detections

Categorization algorithms are sometimes organized as a series of object *detectors* dedicated to the interpretation of a specific category, usually against background. This kind of structure is often used when a priori knowledge is introduced in the decision process, leading to rather different detector designs.

The evaluation of such categorization algorithms cannot be made according to the criteria defined above since they do not integrate in general any specific multiclass category discrimination step: two object detectors may respond positively. If such a structure is chosen, evaluation will be reduced to measure performances using detection indices based on (3) and (4).

4.3 Summary

4.3.1 Measures

Evaluation will be based on the following measures:

- Detection
  - Recall for maximal Precision: R^*
  - Precision for maximal Recall: P^*
  - Equal Precision and Recall: EER
  - Area under the curve: AUC
- Discrimination
  - Discrimination at minimal uncertainty rate: D^*
  - Uncertainty at maximal discrimination rate: U^*
  - Equal discrimination and uncertainty rate: EDU
  - Confusion matrix at maximal uncertainty: $\eta(c, c^*)$
- Rejection
  - Equal Rejection Rate: ERR

4.3.2 Protocole

In ROBIN, evaluation will carefully look at operational point setting. Two cases will be considered, depending on the availability of a confidence coefficient associated with each decision.

Decision with confidence coefficient Detection, rejection or discrimination outputs are given with a coefficient characterizing the a priori confidence on the decision provided. Various operating points are generated by varying a single threshold as control parameter on these coefficients.

Decision without confidence coefficient When no a priori confidence can be generated by the algorithm, the generation of various operating points is the responsibility of the participant. It is expected for each competition at least 5 different series of output in order to approximate the operating points used to define the above measures.

References

[1] “ROBIN Challenge: Competitions”, Emmanuel D’Angelo, Stéphane Herbin and Matthieu Ratiéville