Evaluation Metric of an Image Understanding Result

Baptiste Hemery, Helene Laurent, Bruno Emile, Christophe Rosenberger

Evaluation Metric of an Image Understanding Result

Baptiste Hemery[1]*, Helene Laurent[2], Bruno Emile[2], Christophe Rosenberger[1]

[1] Laboratoire GREYC, ENSICAEN - Université de Caen Basse-Normandie - CNRS

6 boulevard du Maréchal Juin, 14000 Caen - France

phone: (+33) 2 31 53 81 35 , fax: (+33) 2 31 53 81 10

{baptiste.hemery,christophe.rosenberger}@ensicaen.fr

[2] Laboratoire PRISME, INSA Centre Val de Loire - Université d'Orléans

88 boulevard Lahitolle, 18020 Bourges - France

phone: (+33) 2 48 48 40 00, fax: (+33) 2 48 48 40 40

helene.laurent@ensi-bourges.fr, bruno.emile@univ-orleans.fr

Abstract

Image processing algorithms include methods that process images from their acquisition to the extraction of useful information for a given application.

Among interpretation algorithms, some are designed to detect, localize and identify one or several objects in an image. The problem addressed in this article is the evaluation of interpretation results of an image or a video, given an associated ground truth. Challenges are multiple such as the comparison of algorithms, evaluation of an algorithm during its development or the definition of its optimal settings. We propose a new metric for evaluating an interpretation result of an image. The advantage of the proposed metric is to evaluate a result by taking into account the quality of the localization, recognition and detection of objects of interest in the image. Several parameters allow us to change the behavior of this metric for a given application. Its behavior has been tested on a large database and showed interesting results.

Keywords: Evaluation metric, Image understanding, Object localization, Object recognition.

1 Introduction

Image understanding is still a great challenge in image processing. Many applications are concerned such as target detection and recognition, medical imaging or video monitoring. Whatever the foreseen application may be, extracted information conditions the performance of the resulting process. For each object of interest, it is required for the localization to be as precise as

possible and with a correct recognition. Many algorithms have been proposed in the literature to achieve this task [1, 2, 3, 4], but it still remains difficult to compare the performance of these algorithms that extract the localization of objects of interest. In order to evaluate object detection and recognition algorithms, several research competitions have been created such as the Pascal VOC Challenge [5, 6] and the French Robin Project [7]. Given a manually made ground truth, these competitions use metrics to evaluate and compare the results obtained by different image understanding algorithms. If the metrics used for these competitions appeal to everyone's common sense (good correspondence between the ratio height/width or the size of the detected bounding box and of the ground truth), none of them puts the same characteristic forward. Most of proposed metrics in the state of the art quantify the similarity of two localization results with different distances For example, one metric might check if the center's localization of the detected object is correct, whereas an other one might compare the surface of the localized object to the ground truth one. The main objective of these competitions is to compare different image understanding algorithms by evaluating their behavior for different scenarios and parameters on a huge and significant database. It is therefore useful to have a reliable quality score of an interpretation result given the associated ground truth.

Many evaluation metrics initially proposed for various purposes such as image segmentation or image retrieval can be found in the literature and should reveal themselves relevant for the evaluation of image understanding results. In the proposed work, we intend to define a reliable quality score of an interpretation result that, for example, would be able to distinguish automatically the best result among the four synthetic examples presented in figure 1. Most existing metrics from the state of the art are limited as they are not able to evaluate at the same time both localization and recognition errors. For example, the metric proposed in [8] does not consider the class of detected objects but only the similarity of boundaries. Yang *et al.* [9] proposed an evaluation metric for analysing the behavior of an image understanding algorithm defined in the article, but this metric is not studied or validated . Most competitions such as the Pascal VOC challenge evaluate and compare image understanding algorithms on large databases. We think it could be interesting nevertheless to be able to compute the value of a single image understanding result taking into account both detection and recognition aspects.

The paper is organized as follows: Metrics for the evaluation of image understanding in the literature are presented in section 2. A comparative study of these metrics is also briefly discussed. Section 3 presents the proposed

quality score of an image understanding result [10] while section 4 is dedicated to the validation tests. Some illustrations are presented in section 4.2.5 showing how the proposed metric can be used to compare several image understanding results. At last, conclusions and perspectives of this work are given in section 5.

2 State of the art

Image understanding has for objective to automatically extract the maximum amount of information on objects present in an image I. We can extract some information about localization and/or about recognition of predefined classes. We present in this section an overview of existing evaluation metrics for image understanding in a supervised context, that is to say when a ground truth is available.

2.1 Evaluation of a localization algorithm result

The supervised evaluation of a localization algorithm consists in comparing two images: the ground truth and the localization result. Many evaluation metrics initially proposed for various image processing algorithms can be found in the literature [11, 12, 13, 14, 15, 8] and should reveal themselves relevant for localization evaluation. Simply the existence of all these met-

rics expresses the lack about a a widely accepted evaluation algorithm for localization. Moreover, there are three representations of a localization result. The simplest one is the bounding box containing the object. This type of localization is simply represented by the coordinates of the rectangle. The second one consists in using red pixels for representing the contour of objects. The last representation of a localization result is a mask image, where black pixels correspond to the background and white ones to the interior of the object. An illustration of the different representations of a localization result is given in figure 2.

We present for each representation an example of a localization metric. For the French Robin project [7] which aims at evaluating localization and recognition algorithms providing bounding boxes as localization outputs, three metrics have been developed to evaluate a localization result:

$$ROB_{loc}(BB_l, BB_{gt}) = \frac{2}{\pi} arctan(max(\frac{|x_l - x_{gt}|}{w_{gt}}, \frac{|y_l - y_{gt}|}{h_{gt}})) \quad (1)$$

$$ROB_{com}(BB_l, BB_{gt}) = \frac{|\mathcal{A}_l - \mathcal{A}_{gt}|}{max(\mathcal{A}_l, \mathcal{A}_{gt})} \quad (2)$$

$$ROB_{cor}(BB_l, BB_{gt}) = \frac{2}{\pi} arctan(|\frac{h_l}{w_l} - \frac{h_{gt}}{w_{gt}}|) \quad (3)$$

where BB_l is the output of the localization algorithm, $\{x_l, y_l\}$ are the coordinates of the center of the bounding box, $\mathcal{A}_l$ is the area covered by the bounding box and $\{h_l, w_l\}$ are the height and width of the bounding box. Variables with $_{gt}$ subscript correspond to the ground truth. These three metrics evaluate different characteristics of the localization result: ROB_{loc} evaluates the localization of the center of the bounding box, ROB_{com} evaluates the size of the bounding box and ROB_{cor} quantifies the ratio height/width of the bounding box. These metrics evaluate different characteristics of the localization result corresponding to the localization of the center, the size and the ratio height/width of the bounding box.

Concerning the contour representation, several metrics have been proposed, initially for image segmentation evaluation. They can be easily extended to the localization evaluation. For example, the Figure of Merit (FOM) proposed by Pratt [16] is an empirical distance between image I_l that contains the localized object contour and the corresponding ground truth I_{gt}:

$$FOM(I_{gt}, I_l) = \frac{1}{MP} \sum_{k \in I_l^C} \frac{1}{1 + \alpha * d(k, I_{gt}^C)^2} \quad (4)$$

where

$$MP = \max\left(Card(I_{gt}^{C}), Card(I_{l}^{C})\right) \tag{5}$$

and I_l^C are contour pixels of the localized object, α is a constant set to $\frac{1}{9}$ by the authors [16], $d(x, I) = \min_{y \in I} d(x, y)$ and $Card(I^C)$ denotes the cardinal, i.e. the number of contour pixels. One problem of this metric is that it is not sensitive to under-localization errors whereas it is to over- and miss-localization errors. Moreover, it is not sensitive to the shape of miss-localization, which is a problem for the evaluation of localization result.

The mask or region representation is, for example, used in the Pascal VOC Challenge [5]. A simple metric is then defined to evaluate the localization result of an object.

$$PAS(I_{gt}, I_l) = \frac{Card(I_{gt}^{r} \cap I_{l}^{r})}{Card(I_{gt}^{r} \cup I_{l}^{r})} \tag{6}$$

where I_l^r corresponds to region pixels of the localized object, $I_{gt}^r \cap I_l^r$ corresponds to the object pixels correctly localized and $I_{gt}^r \cup I_l^r$ corresponds to object pixels from the ground truth or from the localized object. This metric equals 1 when $I_{gt}^r \cap I_l^r = I_{gt}^r \cup I_l^r$, that is to say when $I_{gt}^r = I_l^r$.

Two other metrics have been proposed by Martin [13] for the evaluation of a localization result. These metrics work for several objects localized in a single image result. These metrics use the local refinement error between two images I_1 and I_2 defined as:

$$E(I_1, I_2, k) = \frac{Card(I_{1\backslash 2}^{r(k)})}{Card(I_1^{r(k)})} \tag{7}$$

where $r(k)$ corresponds to the region containing a pixel from object k, $I_1^{r(k)}$ corresponds to the objets pixels from object k present in image 1, and $I_{1\backslash 2}^{r(k)}$ corresponds to the object pixels from object k present in image 1 and not present in image 2. We can notice that this local error measure is not symmetric and only measures a refinement from image I_1 to image I_2. Martin *et al.* use this local refinement error to create two metrics called global consistency error (GCE) and local consistency error (LCE):

$$\begin{aligned} &MAR_{gce}(I_{gt}, I_l) = \\ &\frac{1}{Card(I)} \min\big(\sum_{k\in I} E(I_{gt}, I_l, k), \sum_{k\in I} E(I_l, I_{gt}, k)\big) \end{aligned} \tag{8}$$

$$MAR_{lce}(I_{gt}, I_l) = \frac{1}{Card(I)} \sum_{k \in I} \min(E(I_{gt}, I_l, k), E(I_l, I_{gt}, k)) \tag{9}$$

with I the common support image of I_{gt} and I_l. We can notice that both metrics are symmetric. Moreover, it is clear that MAR_{gce} is tougher than MAR_{lce}, since MAR_{gce} forces all local refinement to be in the same direction (either from I_l to I_{gt} or from I_{gt} to I_l) whereas MAR_{lce} allows refinement in both directions.

In a previous work [17], we aimed at evaluating the reliability of metrics for the evaluation of localization results. We first referenced up to 33 different metrics from the literature allowing the evaluation of a localization result. In order to evaluate these metrics, we created a synthetic database with 16 ground truths. We used different alterations to create synthetic localization results: translation, scale change, rotation and perspective. We can see on figure 3, some examples of alterations. We finally obtained a total of *118,080* synthetic localization result images.

We computed the values of the metrics between the ground truth and each

synthetic result. We computed a curve showing the behavior of each selected metric face to alterations (an example is given in Figure 10). From these curves, we verified if all the referenced metrics fulfill some properties. The chosen properties that a metric should fulfill to correctly evaluate localization results are the following:

1. Symmetry: a metric should equally penalize two results with the same alteration, but in opposite directions (example, translations of the localization result +5 or -5 pixels horizontally),

2. Strict Monotony: a metric should penalize the results the more they are altered,

3. Uniform Continuity: a metric should not have an important gap between two close results,

4. Topological Dependency: a metric result should depend on the size or the shape of the localized object.

The more properties it fulfills, the better is the metric. The conclusion of [17] was to use region based metrics, and more particularly PAS [5, 6], MAR_{lce}, MAR_{gce} [13] or VIN [18] metrics.

2.2 Evaluation of a recognition algorithm result

The evaluation of a recognition algorithm is also a challenging task. However, it is not possible to directly calculate a distance on the labels returned by the algorithms. Indeed, these identifiers are non ordered categorical variables and calculate a distance between these variables would be meaningless. Naively, the only method for calculating a distance on these variables is to see if they are equal, and thus that the distance is 0 if the identifiers are the same and 1 otherwise. This quantification is imprecise and does not weigh an error between classes.

It is interesting to note that it is possible to calculate *a priori*, i.e. before having the recognition results to evaluate, the set of distances between all object classes present in the used database or application. Indeed, the number of classes is limited and known in advance for a particular application. For example, we would like the distance between classes "cat" and "dog" to be smaller than the distance between classes "car" and "dog". All these distances can then be stored in a distance matrix DM, the problem is how to calculate this matrix.

It is possible to overcome this problem by calculating the distance between

object descriptions. The distance is then dependent on the representation of the object. In the case of an object represented by a graph, the edition distance can be used. It can be seen as the edition cost to turn the first graph G_1 into a graph isomorphic to the second graph G_2. This edition is done by a succession of elementary editions, each with an individual cost. The edition distance is the sum of the costs necessary for this transformation. This distance is complicated to calculate, however, the algorithm from[19, 20] can be used to approximate that distance.

In a previous article [21], we were interested in the representation of an object by a cloud of descriptor points. This cloud of points is a set of automatically detected points on the image. Each point is characterized by a vector calculated in its neighborhood. To build this cloud of points, we used the SIFT descriptors proposed in [22] to detect and characterize the keypoints of the object. With two images of objects, we had two clouds of points, each associated with an object. We matched points of the first object with those of the second one when possible (considering the value of the descriptor), then we defined a similarity measure based on the number of successful matches. It is so possible to obtain a measure of similarity between different classes of objects by taking the similarity scores averaged over a large training base. In some cases, SIFT descriptors fail to characterize images (in presence of coarse

texture as for example), other techniques such as a multi-scale analysis could be used [23] to achieve this goal.

Finally, we can construct the DM matrix or by manually filling subjectively or by using a priori knowledge as a taxonomy or hierarchy.

2.3 Evaluation of an interpretation result

All the evaluation metrics proposed in the literature are concerned with aspects of localization and recognition of objects of interest. A metric for evaluating an interpretation result, in other words taking into account at the same time, these two aspects do not exist. This is the main contribution of this paper.

3 Developed metric

We propose in this paper a metric which is able to quantify the quality of an interpretation result given a ground truth image. The quality score takes into account the localization and recognition (associated to a confidence index) of detected objects as well as missing or over-detected objects in the interpretation result. If we re-examine the four different interpretation results presented in figure 1, the goal we want to achieve is to automatically

determine which result is the best one. The objective is also to have a metric that can be tuned to a specific application, for example by balancing the importance of localization and recognition in the final evaluation result. The proposed metric is composed of four stages, as we can see on figure 4: (i) Objects matching, (ii) Local evaluation, (iii) Over- and Under- detection compensation and at last (iv) Global evaluation score computation.

3.1 Matching

The first stage is necessary to match objects from the ground truth and the interpretation result, so that we can compute a local score for matched objects. Moreover, this enables us to put forward missed objects (under-detection) and multiple detections (over-detection). In order to achieve this goal, we compute a matching score matrix as in [24]. The number of rows corresponds to the number of objects in the ground truth, and the number of columns corresponds to the number of objects in the interpretation result. In each cell of this matrix, we indicate the overlapping of objects. The recovery is computed with the PAS metric [5, 6] we defined in equation 6.

From this matrix, we perform the object matching as an image can contain

multiple ones and with not necessary the same number in the two results. After this step, we obtain a correspondence matrix, where value 1 in cell (u, v) indicates that object u from the ground truth is associated with object v in the interpretation result. At this step, two different methods are possible. The first is a "one-to-one" method computed with the Hungarian algorithm [19]. Such a method is used in the Pascal VOC Challenge [6] for example, and associates one object in the ground truth with one object in the interpretation result. The second one is a "one-to-many" method and enables to associate one object in the ground truth with several objects in the interpretation result and vice-versa. For this method, we use a threshold on the overlapping matrix. This method is used in [25] in the context of document segmentation. Default setting of the metric is the "one-to-many" method with a threshold empirically set to 0.2. This thresold correponds to 20% of overlapping between the the ground truth object and the detected one. A higher value would lead to a higher confidence, but we want to be able to make the difference between bad localization and no detection at all.

3.2 Local evaluation

The local evaluation stage corresponds to the evaluation of each matched object k, that is to say a cell (u, v) in the correspondence matrix having a value

1. We first evaluate the localization of the object and then its recognition. The evaluation metric for the localization is the Martin's one [13] adapted to one object:

$$S_{loc}(I_{gt}, I_i, u, v) = \frac{1}{\text{card}(I)} \min\Big(\frac{\text{card}(I_{gt\setminus i}^{r(u,v)})}{\text{card}(I_{gt}^{r(u,v)})}, \frac{\text{card}(I_{i\setminus gt}^{r(u,v)})}{\text{card}(I_{i}^{r(u,v)})}\Big) \tag{10}$$

with $\text{card}(I)$ the number of pixels in the image and $\text{card}(I_{gt\setminus i}^{r(u,v)})$ the number of pixels present in the ground truth object u and not present in the detected object v. This score ranges from 0 to 1, 0 corresponds to a perfect localization result. We then compute a recognition score. For this, we use the DM distance matrix as described in section 2. If no matrix is used, a default distance matrix is used, filled with 1 except on the diagonal that contains only 0. The metric also enables the use of a confidence index provided by the image understanding algorithm for the recognition result for each detected object. In the following, the confidence index provided by the image understanding algorithm is set to 1 for each detected object (high confidence). The obtained score is the following:

$$S_{rec}(u, v, \mu) = DM(cl(u), cl(v)) * ind(cl(u), cl(v), \mu) \tag{11}$$

with

$$ind(cl(u), cl(v), \mu) = \begin{cases} \frac{1-\mu}{2} & \text{if } cl(u) = cl(v) \\ \frac{1+\mu}{2} & \text{otherwise} \end{cases} \quad (12)$$

μ being the confidence index related to the recognition result and $cl(u)$ the class of object u. Then, we calculate an interpretation score which combines the two previous scores:

$$S(u, v) = \alpha * S_{loc}(I_{gt}, I_i, u, v) + (1 - \alpha) * S_{rec}(I_{gt}, I_i, u, v) \quad (13)$$

The default value of the α parameter is 0.8. However, it can be modified by the user according to the wished preponderance between recognition and localization. We have chosen the 0.8 value in order to penalize more localization errors. After calculating localization and recognition scores, we obtain the local score matrix.

3.3 Over and under-detection compensation

After calculating a score for each matched object, we examine the over or under-detected objects. Under-detection corresponds to rows of the correspondence matrix which have not been associated to any object of the interpretation result, that is to say which have no 1. Over-detected objects correspond to the columns of the correspondence matrix which have not been associated to

any object in the ground truth. First, we take into account under-detection. For this, we look for the first row u without any association. Then, for this row, we look for the column v without any association. We then associate the object u to the object v in the correspondence matrix. In the local score matrix, we insert the score 1 for this association. This is repeated until all the rows are associated. For over-detection, the same tasks are done except that rows and columns are exchanged.

3.4 Global score

The global score is calculated from the local score matrix with compensation. It corresponds to the average value of local scores. In order to take into account the size of objects in the evaluation metric, it is possible to use a weighted sum of local scores (with a weight depending on the size of each object).

3.5 Illustration

We illustrate the proposed evaluation metric with the default parameters on an image interpretation result related to the original image presented in figure 6.

In figure 7, we illustrate each step of the evaluation process, with all default values: i.e the "one-to-many" method with a threshold of 0.2 for the matching step, no confidence matrix for the local recognition score step, no DM distance matrix and α is 0.8 for the local evaluation step, and no weight is given to objects for the computation of the global score. The ground truth is made of seven objects: the first four objects are from the "person" class, followed by an object from the "bus" class, then an object from the "plane" class and finally an object of the "car" class which is not much visible. The interpretation result presents four objects: the first one of the "truck" class, followed by an object from the "plane" class and then by two objects from the "person" class.

We can notice that the plane and the bus have been well localized although the bus has been recognized as a truck. The four persons have been well recognized but not well localized. Indeed, only one object was detected instead of three. Finally, the car has not been detected at all.

The first step consists in matching objects from the ground truth to those of the interpretation result. The overlapping matrix presents the result of the PAS metric for each object couple (u, v). Thus, for the bus which is the fifth object in the ground truth and the first of the interpretation result, masks overlap well. This leads to an overlapping score equal to 0.941. As this result

is superior to the threshold, these two masks are matched as we can see in the correspondence matrix. This is the same for the plane as well as for a person which are both clearly matched. The group of three persons corresponds to the objects 2, 3 and 4 in the ground truth and to the object 3 in the interpretation result. The overlapping score is therefore lower with 0.235 and 0.242 for objects 3 and 4 and 0.186 for object 2. As the threshold is 0.2, only two objects from the ground truth are matched with this group of the interpretation result.

The second step computes the local scores for each matched object. Therefore, a localization score matrix is computed as well as a recognition score matrix. For localization, we can observe that the group of persons is well localized with localization scores lower than 0.01 while the other person is not as well localized with a localization score equal to 0.065, which nevertheless is a good localization score. The plane and the bus are well localized with scores of 0.017 and 0.024. For recognition, we can notice that, except the bus, objects are well recognized. This explains why their scores are 0s. As the bus is recognized as a truck, its score is 1. Using a confidence matrix would enable to reduce this score considering the fact that both objects are quite similar. A local score matrix is then calculated as the combination of the two previous matrices. We can notice that the score of the bus is strongly impacted by the recognition error.

The third step is compensation. We check the correspondence matrix to identify the rows or columns which have not been associated. We can notice that all columns carry at least one 1 which means that the objects of the interpretation result have been matched to at least one object from the ground truth, and that there is no over detection. However, rows 2 and 7 are empty, the corresponding objects in the ground truth have not been well detected. This under-detection is then compensated by adding the score 1 to the corresponding rows. Columns are added to avoid to match these objects to existing ones from the interpretation result which have already been correctly detected.

The last step consists in computing the local score matrix including the compensation step. We replace in the compensation matrix the values corresponding to each detected object by their local scores. Other values (corresponding to under and over-detected objects) are unchanged. From this matrix, we compute the average score. This finally gives the global score. In our case, we compute the average of 7 scores: 5 of them come from the correspondence and 2 of them come from compensation. The final score obtained is 0.328. This score is quite high because the non detection of two objects is highly penalized: the missing objects contribution is of 0.285 and the present objects contribution is of 0.043.

4 Validation

We validate the proposed metric through experiments.

4.1 Experimental protocol

We have tested the proposed evaluation metric on a large database extracted from the Pascal VOC challenge 2008 database[1] [6]. This database provides a set of interesting ground truths displaying a localization mask as well as a class associated to each object. We have applied various alterations to each object present in the ground truths. We then studied the metric behavior according to the various alterations. This section presents the database as well as the alterations we have applied to the ground truths.

4.1.1 Database

Several sets in the database of Pascal VOC Challenge 2008 are available, each of them corresponds to a type of algorithm. We have used the set "Segmentation Taster Set". This group presents ground truths whose localization representation is a mask, which suits the proposed evaluation metric. In table 1, we reference the number of images having N objects. Among the 1022 images, 1002 contain between 1 and 8 objects. The 20 other images have 9 to 21 objects. As there were not at least 10 images including N (N

[1] `http://www.pascal-network.org/challenges/VOC/databases.html`

≥ 9) objects, these images have been rejected. We can notice that most images contain only one or two objects. We finally consider *2.134* objects to alter.

The database contains 20 different classes among "aeroplane", "bicycle", "bird"... We classified these classes following the categories present in the Caltech256 database [26], as we can see in figure 8. An extract of the obtained distance matrix DM is given in figure 2. This DM matrix was computed by taking into at which depth is the common parent node between two classes divided by the maximum depth. For example, for the score between "car" and "bicycle" objects, the common node is "ground" which 2 node above the objects, whereas the maximum is 5 node above (from "'motorized" to "objects"). Thus, the score is 2/5, or 0.4 in the DM matrix presented in table 2.

4.1.2 Alterations

First of all, we have considered the same alterations as in this study [27]: translation, scale change, rotation and perspective change. For each of them, we have used an alteration parameter whose value varies between 1 and 20 according to two different directions: horizontal and vertical for the translation, scaling and perspective, clockwise and counter-clockwise for rotation. Concerning the translation, scaling and perspective alterations, the

parameter corresponds to the distance, in pixels, a pixel from the object is moved at most. For example, if we scale a rectangle with a parameter value of 20, it means that the pixel at the corner of this rectangle will be 20 pixels farther from the center of that rectangle. Concerning the rotation alteration, the parameter is the degree of the rotation. This leads to 160 alterations per object, that is to say a total of *341,440* considered alterations. Afterwards, we considered recognition alterations by modifying objects class. For this, we have replaced the class of one to all objects in the image by the class "*Other* ". Likewise, we also observed the effect of under-detection and over-detection on the proposed evaluation metric. Thus, we have deleted or added 1 to 8 objects in a way that it does not match any other object in the image.

4.1.3 Parameters

We also considered the metric evolution according to its various parameters. We have first observed the effect of the matching process as well as effect of the threshold of the "one-to-many" matching method. Thus, we have compared the "one-to-one " method and "one-to-many" method with threshold values of 0.2, 0.3, 0.4 and 0.5. We also observed the effects of using a distance matrix between classes from the database.

4.2 Experimental results

4.2.1 Localization

The curves on figure 9 show the average evolution of the metric according to various localization alterations and according to the number of objects in the ground truth. Each curve represents the metric evolution according to the power of alteration. Here, the metric is presented with the default values. We first observe that the more objects the ground truth contains, the less penalizing is the metric. The global score corresponding to the average of local scores, this result is correct and appropriate to the proposed specifications. Whatever the alteration or number of objects, the curves are uniformly regular and strictly monotonic. Moreover, the metric also fulfills the separability property and is symmetric although this last property is not visible on the curves.

As we can see on figure 3, which shows images all altered with the same alteration parameter, translation and rotation (figure 3 (a) and (c)) alterations are the more alterating ones. Moreover, we can see on figure 9 that the metric penalizes translation and rotation the most (with a similar power of alteration), followed by scaling and perspective change. Hence, the metric evolves accordingly to the power of the altertion, and behaves correctly

for localization alterations.

4.2.2 Recognition

As for localization, we studied the metric evolution according to the number of objects whose class was altered. Given the fact that default parameters are used, the class associated to an altered object does not have any effect on the result. That is why the class "Other" has been assigned to every altered object. Figure 10 shows the metric evolution according to the number of altered objects, the various curves representing different numbers of objects in the ground truth. We notice that the metric has a good behavior as the maximum score, which is 0.2 ($1 - \alpha$, the α parameter default value is 0.8), is reached when all objects from the ground truth are altered.

4.2.3 Over- and under-detection

The effect of over-detection and under-detection on the global metric evolution has also been studied. We observed the evaluation metric evolution with the default parameters according to the number of objects deleted or added to the ground truth. These results are presented in figure 10. Negative numbers correspond to deleted objects whereas positive numbers correspond to added objects. Each curve is related to a different number of objects in the ground truth.

All these situations are correctly processed and the metric is always more penalizing when there are few objects in the ground truth. We can notice that under-detection is slightly more penalized than over-detection. This can be easily be explained by the fact the global score is computed as the mean local score: i.e the sum of local score divided by the number of object. When an object is over-detected or under-detected, the numerator of that division is changed in the same way, but the denominator is changed, and increased, only in case of over-detection, which leads to a lower increase of the global score.

In conclusion, when the metric is used with the default parameters, it gives good results. However, it could be interesting to adapt the parameters specifically for a certain application. The rest of this section presents the effects of the parameters setting on the metric behavior.

4.2.4 Parameters

Matching The first parameter to be considered is the matching method choice including the influence of the "one-to-many" matching method threshold. In order to achieve this study, we observed the evolution of the proposed evaluation metric according to the four localization alterations, exclusively on

ground truths containing only one object. Each curve in figure 11 represents a specific definition of parameters of the evaluation metric.

This figure highlights the fact that the results obtained with the "one-to-many" setting is more penalizing than the "one-to-one" setting. Furthermore, it underlines that the higher the threshold is, the more penalized the alterations are. This can be explained by the fact that there is only one object in the ground truth. Thus, the "one-to-one" method always associates the altered detection to the object in the ground truth, whereas the "one-to-many" method does not associate it anymore from a certain threshold value. The greater this value is, the earlier will this phenomenon appear. It should also be noticed that it appears earlier for rotation and translation than for the scale change alteration. Lastly, we can notice that the perspective change alteration is not important enough to induce this behavior.

Recognition results weighting Finally, we observed the effect of the distance matrix on the recognition results. In order to do this, we computed the distance matrix from the taxonomy created at figure 8. Figure 12 shows the evolution of the recognition score according to whether the distance matrix is used or not. Figure 12 (a) presents the recognition score evolution for the

(10th) “potted plant” class according to the recognized class. We can notice that the use of a distance matrix enables a more precise evaluation score. Figure 12 (b) presents the average recognition score evolution according to the recognized class. The results have been sorted to present the score evolution according to the affected class from the closest to the furthest. This confirms the fact that the recognition result is better evaluated with a distance matrix than without.

4.2.5 Illustrations

If we go back to the interpretation results given in figure 1 and evaluate them with the proposed evaluation metric, we obtain the following results: result (c), score = 0.1242; result (d), score = 0.1574; result (e), score = 0.3716; and result (f), score = 0.1935. For this evaluation, we use the default parameters of the evaluation metric. We can see that result (e) obtain the worst score, as there is a missing object. Result (f) is following: there are only two objects, but one is overlapping two objects in the ground truth, which is less penalizing than for result (e). Results (c) and (d) both have a correct detection of objects. However, the first one is better as it does not have any recognition mistake.

We give another example of evaluation results in figure 13 with the default parameters. We can see that result (d) is given as the best one, even if two potted plants are not correctly localized. We could change this behavior by switching the matching process from "one-to-many" to "one-to-one", and one of the potted plant would be a miss. Result (c) is considered worse than result (d) due to the cat misclassification. The result (f) seems to be correct, but there is a potted plant in the upper left corner that is not present in the ground truth. This clearly shows the importance of the ground truth. Result (e) is clearly the worst one as none potted plant is detected.

We present some illustrations on real image understanding results. The results have be obtained from two recent papers. The firt image understanding algorithm has been proposed by Gonzales-Diaz and Díaz-de-María [28]. The second one concerns the method proposed by Shotton *et al.* [29]. Results are given on images from the MSRC 21-class database on figures 14 and 15. Of course, these two methods provide good results but the proposed metric is able to distinguish some small differences between them.

4.3 Discussions

The results given by the proposed metric are satisfying. We have shown that the metric takes into account: (i) a bad localization, (ii) a bad recognition and (iii) a bad detection. The alterations are penalized according to their importance, from the greatest to the least: first, bad detection, second, bad recognition and at last bad localization results. Similarly, concerning the penalization of possible alterations of localization, rotation and translation problems overcome the scaling and perspective problems. The method can also be customized thanks to several parameters. The various configurations modify the evaluation results. In particular, the "one-to-many" matching method enables a more severe evaluation by increasing the value of the matching threshold. Moreover, the metric can take into account a confidence index given by the interpretation algorithm, leading to a more subtle evaluation result.

5 Conclusions

We propose in this paper an original metric that enables the evaluation of an interpretation result given a ground truth. This metric can take into account, at the same time, information from the localization, recognition and detection of objects. It is based on four step: matching, local score

computation, compensation for misdetection and global score computation. The metric was created in order to be customizable for a specific application. The matching method can be adjusted and the matching threshold as well. We can also balance the importance of recognition and localization in the local score computation. Finally, the use of a distance matrix between classes enables a subtle evaluation result. The results obtained with this evaluation metric correspond to our objectives. Moreover, we have seen that it manages well the different possible alterations, enabling to automatically compare several interpretation results and emphasizing the best one.

Some perspectives are clearly visible. First, we could study the way to automatically create a distance matrix between classes from a database. Second, we would like to study how the proposed method behave face to a subjective evaluation of interpretation results. The objective would be to check if the proposed metric achieve an appropriate behavior compared to what can be obtain by an evaluation done by humans. Finally, we could study the performance of this metric in practical applications.

References

[1] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, "Detecting moving objects, ghosts, and shadows in video streams", IEEE Transactions on Pattern Analysis and Machine Intelligence, 25, pp. 1337-1342, 2003.

[2] F. Jurie, C. Schmid, "Scale-invariant shape features for recognition of object categories", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2, pp. 90-96, 2004.

[3] F. Li, J. Carreira, C. Sminchisescu, "Object recognition as ranking holistic figure-ground hypotheses", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1712-1719, 2010.

[4] P. Arbelez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, J. Malik, "Semantic segmentation using regions and parts", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3378-3385, 2012.

[5] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko and others, "The 2005 PASCAL visual object classes challenge", 2005.

[6] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results", 2008.

[7] E. D'Angelo, S. Herbin, M. Rativille, "ROBIN Challenge Evaluation principles and metrics", 2006.

[8] J. Pont-Tuset, F. Marques, "Measures and meta-measures for the supervised evaluation of image segmentation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2131-2138, 2013.

[9] Y. Yang, S. Hallman, D. Ramanan, C. Fowlkes, "Layered object models for image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 34, pp. 1731-1743, 2012.

[10] B. Hemery, H. Laurent, C. Rosenberger, "Evaluation Metric for Image understanding", IEEE International Conference on Image Processing (ICIP), 2009.

[11] M. Basseville, "Distance measures for signal processing and pattern recognition", Elsevier Signal Processing, 18, pp. 349-369, 1998.

[12] D.L. Wilson, A.J. Baddeley, R.A. Owens, "A new metric for grey-scale image comparison ", International Journal of Computer Vision, 24, pp. 5-17, 1997.

[13] D. Martin, C. Fowlkes, D. Tal, J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation

Algorithms and Measuring Ecological Statistics", 8th International Conference on Computer Vision, 2, pp. 416-423, 2001.

[14] M. Everingham, H. Muller, B. Thomas, "Evaluating image segmentation algorithms using the pareto front", International Conference on Computer Vision (ECCV), Springer, pp. 34-48, 2002.

[15] A. Hafiane, S. Chabrier, C. Rosenberger, H. Laurent, "A new supervised evaluation criterion for region based segmentation methods", LNCS4678 International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS), pp. 439-448, 2007.

[16] W. Pratt, O. Faugeras, A. Gagalowicz, "Visual discrimination of stochastic texture fields", IEEE Transactions on Systems, Man, and Cybernetics, pp. 796-804, (8), 1978.

[17] B. Hemery, H. Laurent, B. Emile, C. Rosenberger, "Comparative Study of Localization Metrics for the Evaluation of Image Interpretation Systems", Journal of Electronic Imaging, (19), 2010.

[18] L. Vinet, "Segmentation et mise en correspondance de rgions de paires d'images stroscopiques" Universit de Paris IX Dauphine, 1991.

[19] J. Munkres, "Algorithms for the Assignment and Transportation Problems", Journal of the Society for Industrial and Applied Mathematics, SIAM, vol. 5, (32), 1957.

[20] K. Riesen, M. Neuhaus, H. Bunke, "Bipartite graph matching for computing the edit distance of graphs", Lecture Notes in Computer Science : Graph-Based Representations in Pattern Recognition, Springer, 4538, pp. 1-12, 2007.

[21] B. Hemery, H. Laurent, B. Emile, C. Rosenberger, "Comparative Study Of Local Descriptors For Measuring Object Taxonomy", IEEE International Conference on Image and Graphics (ICIG), 2009.

[22] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 60, pp. 91-110, 2004.

[23] J. Kim, C. Liu, F. Sha, K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2307-2314, 2013.

[24] I.T. Phillips, A.K. Chhabra, "Empirical Performance Evaluation of Graphics Recognition Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 21, pp. 849-870, 1999.

[25] C. Wolf, J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms", International Journal on Document Analysis and Recognition, Springer, 8, pp. 280-296, 2006.

[26] G. Griffin, A. Holub, P. Perona, "Caltech-256 Object Category Dataset", California Institute of Technology, 2007.

[27] B. Hemery, H. Laurent, C. Rosenberger, B. Emile, "Evaluation Protocol for Localization Metrics - Application to a Comparative Study", International Conference on Image and Signal Processing (ICISP), pp. 273-280, 2008.

[28] I. Gonzlez-Daz, F. Daz-de-Mara, "A region-centered topic model for object discovery and category-based image segmentation", Pattern Recognition, 46, pp. 2437-2499, 2013.

[29] J. Shotton, C.R. Winn, A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context", International Journal of Computer Vision, 81, pp. 2-23, 2009.

Baptiste Hemery is a research engineer at ENSICAEN (France). He obtained his Phd from the Uni- versity of Caen Basse-Normandie in 2009. He belongs to the GREYC laboratory in the E-payment & Biometrics research unit. His research interests concern image interpretation evaluation, biometric systems and machine learning for fraud detection.

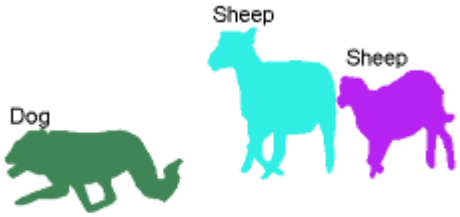
Helene Laurent is an associate professor at ENSI of Bourges (France). She obtained his Phd from the university of Nantes in 1998. She belongs to the PRISME research laboratory in the Image and Signal system processing unit. Her research interests concern segmentation evaluation.

Bruno Emile is an associate professor at IUT of Châteauroux (France). He obtained his Phd from the university of Nice in 1996. He belongs to the PRISME research laboratory in the Image and Signal system processing unit. His research interests concern object detection and computer vision.

Christophe Rosenberger is full professor at ENSICAEN (France). He obtained his Phd from the university of Rennes I in 1999. Since 2007, He belongs to the GREYC laboratory in the E-payment & Biometrics research unit.. His research interests concern biometric systems.

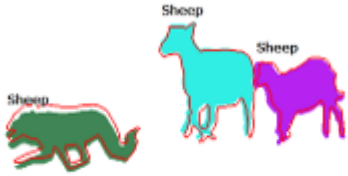
List of Figures

(*a*) *orignal image*



(*b*) *ground truth*

(*c*) *Interpretation result* 1



(*d*) *Interpretation result* 2

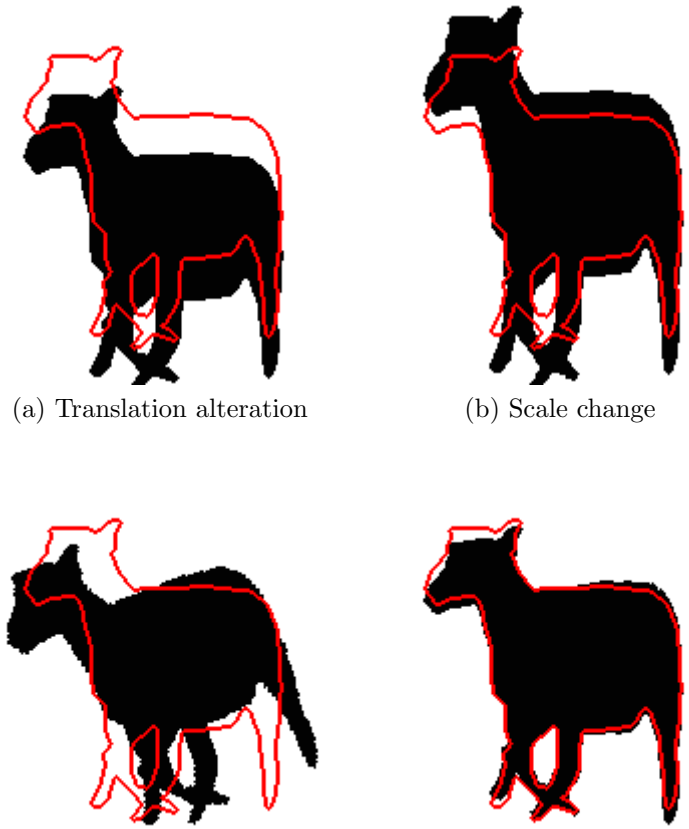
(*e*) *Interpretation result* 3

(*f*) *Interpretation result* 4

Figure 1: Examples of synthetic image understanding results on a single image (original image and ground truth extracted from [6]): red lines show the localization ground truth.

(a) Bounding box (red) (b) Contour (red) (c) Mask

Figure 2: Different representations of a localization result



(a) Translation alteration (b) Scale change

(c) Rotation alteration (d) Perspective alteration

Figure 3: Four examples of alterations

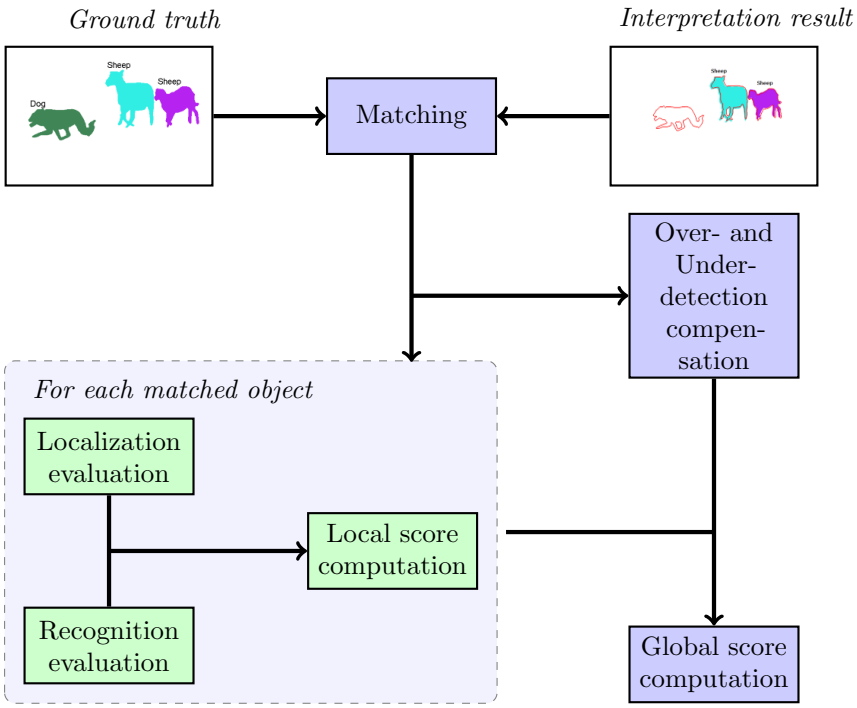


Figure 4: Principle of the evaluation process

(a) Ground truth (b) Localization result (c) Overlapping of both regions

Figure 5: Illustration of the PAS metric

Figure 6: Original image extracted from [6]

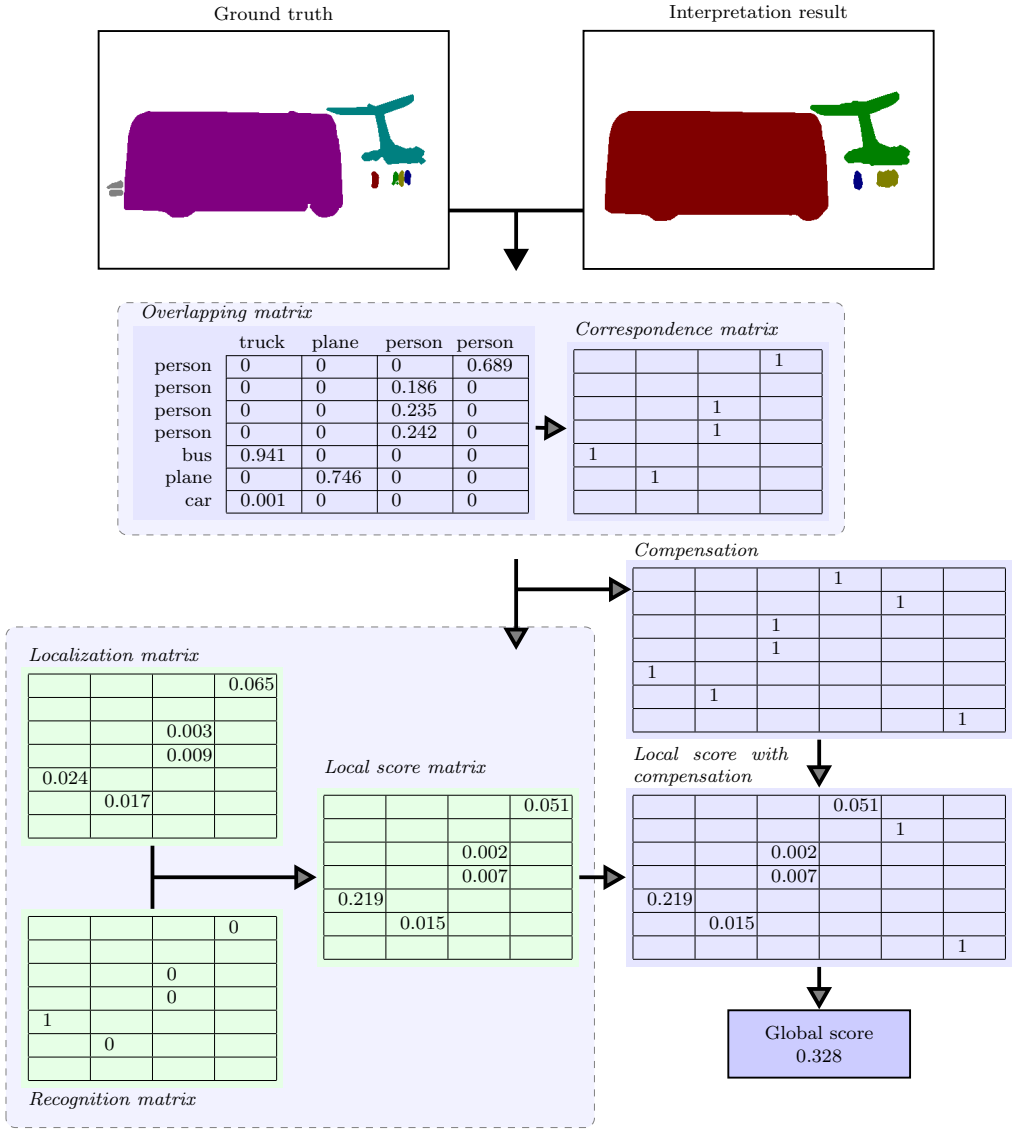


Figure 7: Example of global evaluation of an interpretation result

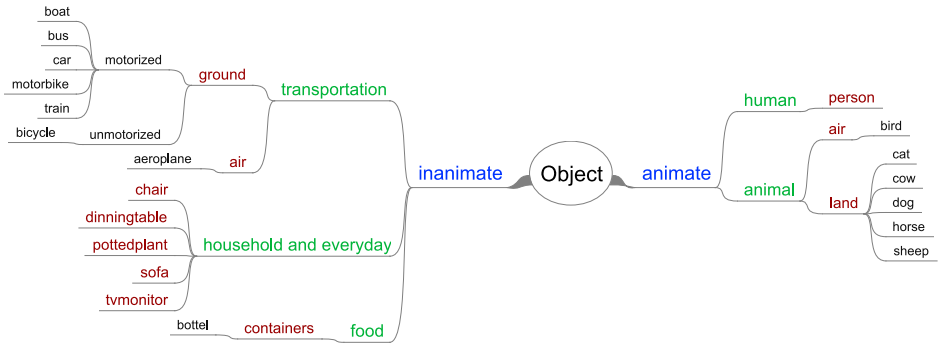
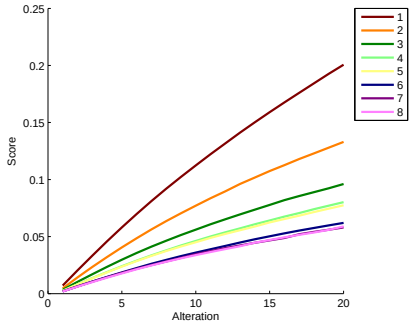
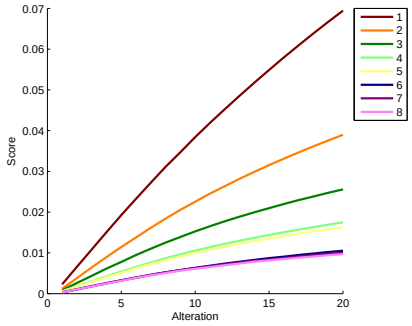


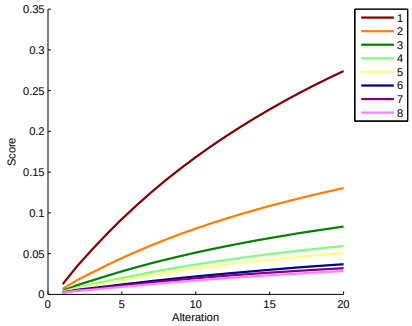
Figure 8: Classes of Pascal 2008 organized with Caltech256 database taxonomy



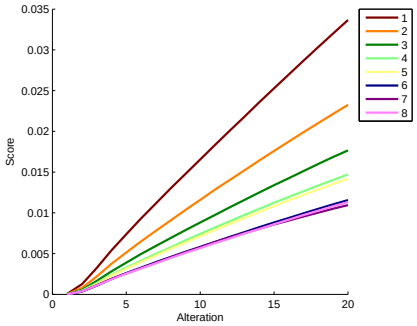
(a) Translation alteration



(b) Scale change

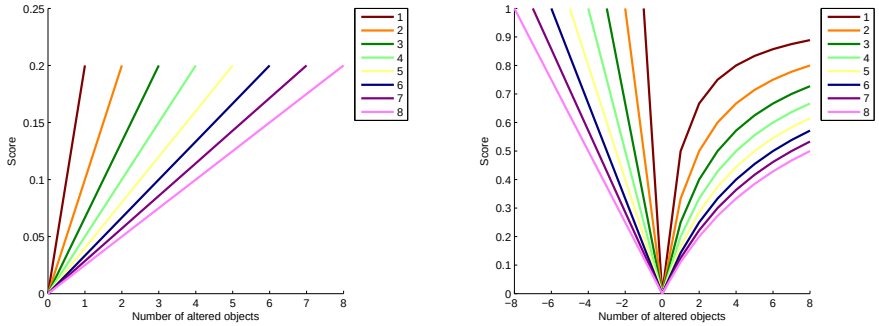


(c) Rotation alteration



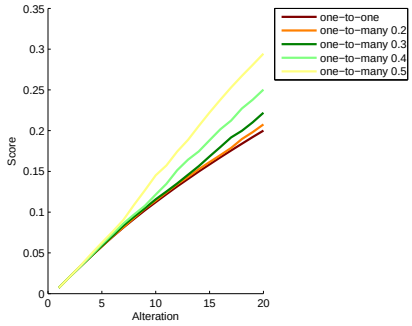
(d) Perspective alteration

Figure 9: Localization results

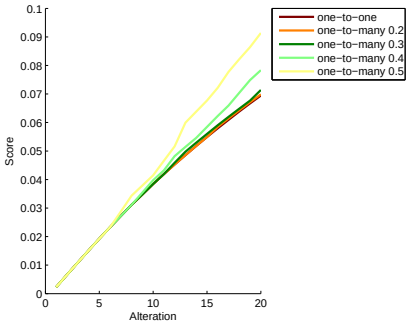


(a) Recognition (b) Over- and under-detection

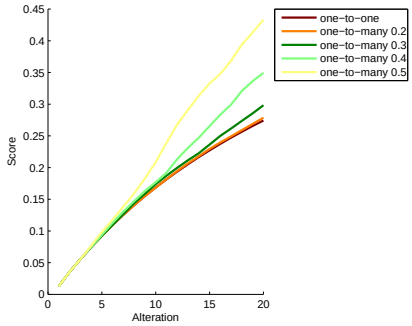
Figure 10: Recognition and detection results



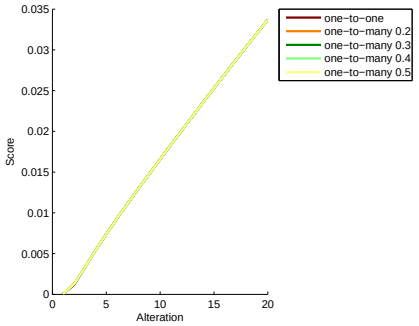
(a) Translation alteration



(b) Scale change

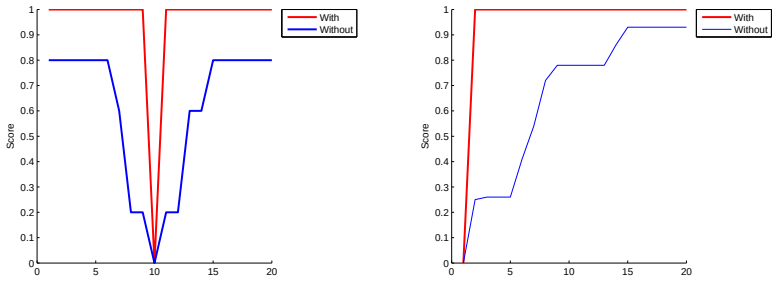


(c) Rotation alteration



(d) Perspective alteration

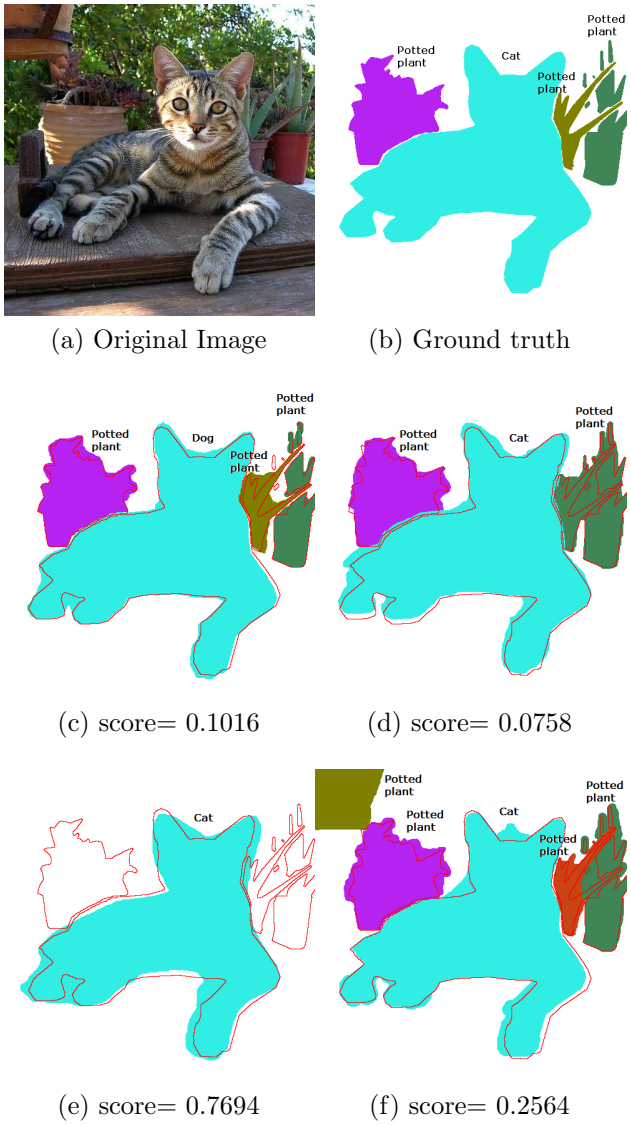
Figure 11: Influence of parameters setting



(a) Result for "potted plant" class (class number 10)

(b) Mean result

Figure 12: Recognition score with and without the use of a distance matrix

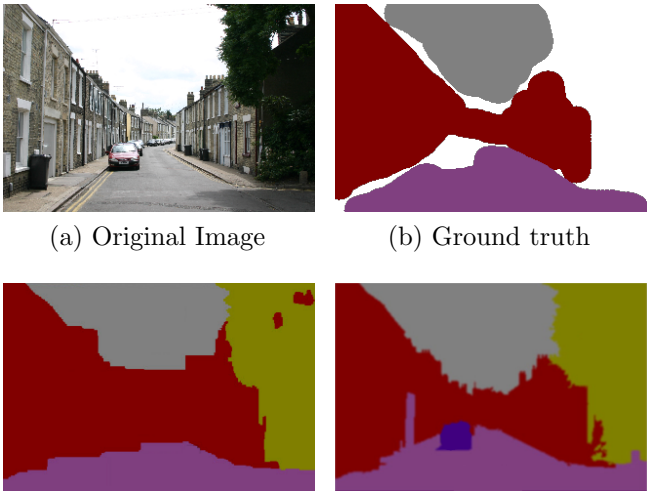


(a) Original Image (b) Ground truth

(c) score= 0.1016 (d) score= 0.0758

(e) score= 0.7694 (f) score= 0.2564

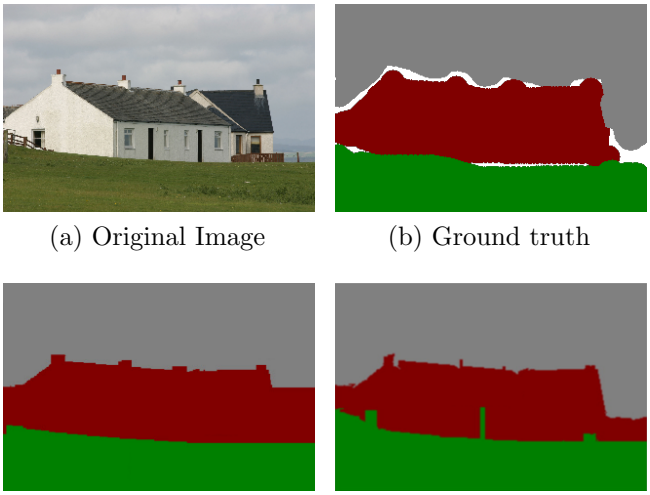
Figure 13: Examples of evaluation results (original image and ground truth extracted from [6]). Result (d) is given as the best one by the proposed evaluation metric.



(a) Original Image (b) Ground truth

(c) score= 0.3161 (d) score= 0.4470

Figure 14: Examples of evaluation results: Result (c) obtained by the method proposed by Shotton et al. [29] is given as the best one by the proposed evaluation metric. The result (d) is penalized because the over-detection of the car (not present in the ground truth).



(a) Original Image (b) Ground truth

(c) score= 0.3969 (d) score= 0.3814

Figure 15: Examples of evaluation results: Result (d) proposed by Gonzales-Diaz and Díaz-de-María [28] is given as the best one by the proposed evaluation metric. The right of the house is better segmented on this result.

List of Tables

Table 1: Number of images containing N objects

N objects	1	2	3	4	5	6	7	8
N images	498	237	106	61	40	32	16	12

Table 2: Extract of the distance matrix DM obtained from taxonomy

	"aeroplane"	"bicycle"	"car"	"bird"
"aeroplane"	0	0.6	0.6	1
"bicycle"	0.6	0	0.4	1
"car"	0.6	0.4	0	1
"bird"	1	1	1	0