Parametrization of an image understanding quality metric with a subjective evaluation

B. Hemery\textsuperscript{1}, H. Laurent\textsuperscript{2}, B. Emile\textsuperscript{2}, C. Rosenberger\textsuperscript{1}\*  

\textsuperscript{1} GREYC Laboratory  
ENSICAEN - Université de Caen Basse-Normandie - CNRS  
6 boulevard du Maréchal Juin, 14000 Caen - France  
baptiste.hemery@greyc.ensicaen.fr  
christophe.rosenberger@greyc.ensicaen.fr  
phone: (+33) 2 31 53 81 35, fax: (+33) 2 31 53 81 10  

\textsuperscript{2} PRISME Laboratory  
ENSI de Bourges - 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 understanding has many real industrial applications (video-monitoring, image retrieval . . . ). Given an image and an associated ground truth, it is possible to quantify the quality of understanding results provided by different algorithms or parameters. For that, it is necessary to take into account many factors for each object in the image: localization and recognition errors and under or over-detection of objects. In order to define an evaluation metric for quantifying the quality of an image understanding result, we have to set, as for example, the weights of each kind of error in the global score. For a correct parameters setting of an evaluation metric we defined previously, we conducted a subjective evaluation of image understanding results involving many experts in image processing. We present in this paper the developed method and analyze the obtained results to ponderate the various errors in an appropriate way. We show the benefit of this kind of study to define the correct parameters of the metric in order to have a judgment as reliable as the one provided by experts. Experimental results on many images from the PASCAL VOC Challenge show the good behavior of this metric compared to the human judgment.

Keywords: Image understanding, Subjective evaluation, Object localization, Object recognition, Evaluation metrics.
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, the extracted information conditions the performances of the resulting process. It is required for this 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] or the French Robin Project [6]. 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. The main objective of these competitions is to compare different image understanding algorithms by evaluating their global behavior for different scenarios and parameters. This a good way to quantify the efficiency of such algorithms if the benchmark database is significant and contains all the cases in a real industrial use. Nevertheless, there exist very few evaluation metrics that are able to measure the quality score of a single understanding result.

Such kind of metric could have many benefits. First of all, it could allow the comparison of multiple algorithms on a single test image before computing the performance on a large database. Second, it could be useful to set some parameters of an algorithm on a representative test image. This metric could be useful to analyze precisely the behavior of an image understanding algorithm. Indeed, when computing the performance of a new
algorithm on a large benchmark, it could be interesting to identify the contexts where the algorithm is not sufficiently efficient.

In a previous work [7], we proposed a metric that enables the evaluation of such results given a ground truth. The metric ranges from 0 to 1, value 0 being obtained when the understanding result perfectly fit the ground truth. As an example, figure 1 presents the evaluations obtained with this metric for four understanding results. Taking into account and penalizing all possible errors: localization, recognition, under-detection, over-detection, it enables an objective comparison of these results.

As an example, figure 1 presents the evaluation obtained with this metric for four understanding results. It enables an objective comparison of these results. Result 1 and 4 have better scores since all objects are correctly recognized even if the localization is less precise than the result 2. Result 2 has as bad score because the dog is recognized as a sheep, and result 3 has a bad score since one object is missing. In order to define this metric, we had to penalize all possible errors (localization, recognition, under-detection, over-detection).

The aim of this work is to check if this metric corresponds to what can be obtained by an evaluation done by humans. Are the different errors penalized in the same way as a human? Are we able to predict the human judgement with this metric? How good are we to achieve that? In order to answer these questions, we defined a subjective evaluation protocol. We asked many individuals to compare several image understanding results. We then compare the obtained subjective comparison with the objective one given by this metric.

This paper is organized as follows: the first part briefly presents the principle of the evaluation metric of image understanding results. We make then a state of the art of
methods for conducting a subjective evaluation process. The developed method is then explained in our context. Experimental results of the subjective evaluation are presented, we show then how to exploit these results in the definition of the quality metric. Finally, we give some conclusions and some perspectives of this study.
Quality score of an image understanding result

In order to quantify the quality of an image understanding result given a ground truth, we have to consider three aspects: localization, recognition and under/over detection. We present some metrics from the literature used in this context in the following.

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 purposes such as segmentation evaluation or image retrieval evaluation can be found in the literature [8, 9, 10, 11] and should reveal themselves relevant for localization evaluation. The existence of all these metrics simply expresses the lack about a well-known localization algorithms evaluation. Concerning the recognition step, very few methods exist. Moreover, only statistical approaches exist. The two most used methods are the ROC and the PR curves. Receiver operating characteristic (ROC) curves [12, 13] are a graphical representation of the true positive rate, also known as the recall, versus the false positive rate. The confusion matrix [6] is a good alternative to ROC or PR curves for the evaluation of recognition. Nevertheless, these metrics for evaluating the recognition efficiency is not well fitted for a single result for many ones given a significant benchmark database. Finally, some other statistical methods can be used, as the $\chi^2$ method [14] or the Kappa measure [15] for example. In a previous paper [7], we defined an evaluation metric that enables the evaluation of an understanding result given a ground truth. As far as we know, none other metric having the same ability exists in the literature.

2.1 General principle

The computation of the proposed metric is composed of four stages, as we can see in figure 2: (i) Matching objects, (ii) Local evaluation, (iii) Over- and Under- detection compensation and finally (iv) Global evaluation score computation.
The first stage consists in matching objects between the ground truth and the understanding result. We use the matching metric proposed by the Pascal VOC challenge [5]. The local evaluation stage corresponds to the evaluation of each matched object. This step first takes into account the localization of the associated object positions. We use the localization metric proposed by Martin et al. [10]. The comparative study we made on localization metrics [16] showed that this metric respects many desired properties and can be considered as the one of the most reliable in the state of the art. Second, the object recognition is evaluated. We verify if the object had been correctly identified and if it is not the case, we penalize the error based on the similarity between the correct object class and the detected one. We use for that a similarity measure based on local features [17]. The local evaluation of one object depends of these two factors: localization and recognition. By default, the weight of localization is set to $\alpha$ while the recognition one is set to $1 - \alpha$. 

**Figure 2: Principle of the evaluation metric of an image understanding result**
The third stage aims at compensating the under- and over-detection. This stage affects the local score of under- and over-detection objects with 1, which is the worst score. Finally, the global score is computed as the mean of local scores.

In order to choose the localization metric used in local evaluation score, we realized in a previous paper [16] a comparative study of the existing metrics in the literature. In order to compare these metrics, we defined an evaluation protocol: we alter the ground truth and check if results given by a metric fulfill some properties. A correct metric should fulfill most of the following properties:

- **Strict Monotony**: a metric should penalize the results the more they are altered,
- **Symmetry**: a metric should equally penalize two results with the same alteration, but in opposite directions,
- **Uniform Continuity**: a metric should not have an important gap between two close results,
- **Topological dependence**: a metric result should depend on the size or the shape of the localized object.

Several parameters enable to tune the evaluation metric. We use a distance matrix between each class present in the databases, which will enable to better evaluate recognition mistakes. We can also to set the parameter $\alpha$ to balance the weight of localization and recognition evaluation in the local score. Results, presented in [7], show that the proposed metric enables the correct evaluation of image understanding results if considering some properties we defined.

However, we would like to know its relative behavior compared to a subjective evaluation (expert judgments). Do we obtain a similar judgment with the metric than humans?
How can we set the $\alpha$ value? We propose to use a subjective evaluation. In the next section, we propose a general methodology based on a subjective evaluation methodology for the setting of the evaluation metric.

3 Developed method

The general objective of the proposed methodology is to answer the previous questions. We define a subjective evaluation protocol associated with a statistical analysis. First of all, we need to acquire some data from users.

3.1 Data acquisition

In order to acquire feedbacks from individuals, we created a website where a user can create an account and then answer to questions. The questionnaire presents the original image, the ground truth and four image understanding results. The user is asked to order image understanding results from the most to the less similar to the ground truth. The ground truth is given for the user in order to help him to compare the different understanding results. This relative comparison of different results in the same screen is more adequate for users than presenting alternatively the different results to compare. Two examples of question can be seen in figure 3.

3.2 Structure of the questionnaire

The subjective evaluation is composed of 12 questions. The 12 original images used in this study come from the Pascal database [5], where the original image and the associated ground truth is provided. The corresponding taxonomy, according to the one used in Caltech256 [18]. This taxonomy is used to penalize more or less a recognition error. As for example, if an object belongs to the cat class, the error of affecting it to the dog one is less important than affecting it to the bird one. For the first goal of this study, which is to
However, concerning the desired properties of a metric, questions were specifically designed to verify them. Some questions also aim to verify several properties. The first property is the strict monotony and 5 questions are dedicated to this property: questions 3
and 9 for the translation alteration, question 4 for the rotation alteration, question 6 for the scale change alteration and question 12 for the recognition alteration. We present to the expert some results with different levels of alterations. As for example, the localization of an object in the result is more and more altered with a translation or other alterations. The second property is the symmetry and 4 questions are dedicated to: questions 3 and 9 for the translation alteration, question 6 for the scale change alteration and question 12 for the perspective alteration. The third property is the continuity, but it cannot be evaluated through this subjective evaluation. Finally, the fourth property corresponds to the effect of the size and shape of the object and two questions are dedicated to this property: questions 3 and 9 have one object with the same alteration for the four image understanding results.

Moreover, we would like to answer some other questions. The first one is to define which alteration is the most penalizing one among translation, scale change, perspective change and rotation for the localization, and also recognition errors and over- or under-detection errors: 8 questions are dedicated to this purpose. Finally, we also verify if humans are able to reproduce their evaluation: 2 questions present exactly the same original image, ground truth and image understanding results. The questionnaire has been designed to answer all these questions an experts ignore the real intent of questions.

The web site was available for one week. 88 individuals participated, and 83 completed the study for the 12 questions. These individuals are researchers in computer science, but not specifically in the image processing field. Acquired data consist in 12 matrices, one for each question, with 88 lines corresponding to each individual which started the study, and 4 columns corresponding to its answer (sorting of understanding results). We think this number of answers is really significant. To our knowledge, none equivalent study has been one in this field implying some many users.
3.3 Filtering data

First of all, we suppress data corresponding to questions not completed by the 5 individuals who did not complete the study. Then, we filter the remaining data. This step consists in collecting the relevant information by suppressing of this study the answers too dissimilar compared to the mean answer for each question. This technique enhances the reliability of the extracted knowledge. We have used the linear Pearson correlation factor as defined in equation 1 for the answer selection.

$$\text{Pearson}(X_i, E[X]) = \frac{\text{Cov}(X_i, E[X])}{\sqrt{\text{Cov}(E[X], E[X]) \cdot \text{Cov}(X_i, X_i)}}$$

(1)

where $X_i$ represents the answers of the user $i$, $E[X]$ represents the average value of answers given by users and $\text{Cov}(.,.)$ is the covariance function. The Pearson correlation factor between two variables gives a value between $[-1, 1]$ and denotes the linear relationship between them. The decision criterion given by equation 2, with $\theta = 0.7$ empirically chosen, permits to select the answers that will be considered for the further analysis.

$$\begin{cases} 
\text{Pearson}(X_i, E[X]) \geq \theta & \text{accept} X_i \\
\text{otherwise} & \text{reject} X_i 
\end{cases}$$

(2)

Among the 1014 answers collected, 232 are rejected by this filtering process.

3.4 Evaluation of the global performance

As we have relative measures, we can compare the quality of different image understanding results and sort them as in [19]. For each question of the subjective study, the 4 image understanding results are sorted according to the average score given by the individuals. Given this sorting, we can extract 6 comparisons results for each pair of image understanding result given by individuals and by using the image understanding metric.
In order to define the similarity between the criterion and our reference given by the individuals’ scores, an absolute difference is measured between the criterion comparison and the individuals’ one. We define the cumulative similarity of correct comparison (SCC):

\[
SCC = \sum_{k=1}^{12} \sum_{i=1}^{6} |I(i, k) - M(i, k)|
\]  

where \(I(i, k)\) and \(M(i, k)\) are respectively the individuals and the metric results for the \(i\)th comparison of question \(k\). A comparison result is a value in \([-1, 1]\). If an image understanding result is better than another one, the comparison value is set to 1 otherwise it equals -1. In order to more easily compare this error measure, we also define the similarity rate of correct comparison (SRCC), which represents the absolute similarity of comparison referenced to the maximal value:

\[
SRCC = (1 - \frac{SCC}{SCC_{max}}) \times 100
\]

where \(SCC_{max}\) corresponds to the most important difference of the \(6 \times 12 = 72\) comparison results. In our case, \(SCC_{max} = \binom{4}{2} \times 12 \times 2 = 144\). The binomial coefficient \(\binom{4}{2}\) corresponds to the number of possibilities to compare 2 answers among 4, 12 is the number of questions in the study and 2 corresponds to the fact that a comparison is set to be between -1 and 1. The SRCC value allows us to compare the judgment done by users and an image understanding metric.

### 3.5 Validation of properties

In order to determine whether there is a significant relationship between answers from a question, we use the Kruskall-Wallis test (KW). It is a non-parametric (distribution free) test, which is used to decide whether K answers are dependent. In other words, it is used to test two hypothesis given by equation 5: the null hypothesis \(H_0\) assumes answers given by individuals are identical (i.e., there is no difference between the answers) against the
alternative hypothesis $H_1$ which assumes that there is a statistically significant difference between answers from a question.

$$\begin{align*}
H_0 & : \mu_1 = \mu_2 = \ldots = \mu_k \\
H_1 & : \exists i, j, \mu_i \neq \mu_j
\end{align*}$$

The Kruskal-Wallis test statistic is given by equation 6 and the p-value is approximated, using chi-square probability distribution, by $Pr(\chi^2_{g-1} \geq K)$. The decision criterion used to choose the appropriate hypothesis is defined in equation 7.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \bar{r}_i - 3(N+1)$$

with $n_i$ is the number of answers in result $i$, $r_{ij}$ is the rank of answer $j$ from result $i$, $N$ is the total number of answers across all results.

$$\bar{r}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$$

$$\begin{align*}
\text{p-value} & \geq 0.05 \quad \text{accept } H_0 \\
\text{otherwise} & \quad \text{reject } H_0
\end{align*}$$

This statistical test will allow us to verify if the properties we defined are significant based on users’ answers. We have now all the material to answers all the questions of this study.

4 Experimental results

We present in this section the analysis of the subjective evaluation and we exploit it for parameterizing the evaluation metric we defined.
4.1 Global behavior of the evaluation metric

We first computed the SRCC value with default parameters: we do not use a distance matrix to balance misclassification and the $\alpha$ parameter, which is used to balance localization and recognition scores, is set to 0.8. The obtained $SRCC$ is 83.33%, which shows that the proposed metric is able to order image understanding results correctly in most of cases.

We then present in figure 4 the evolution of the $SRCC$ as a function of the parameter $\alpha$, and with or without using a distance matrix (weighting of a recognition error). The distance matrix used for the evaluation is computed from the taxonomy presented in figure ??: the distance between two classes depends on their distance on the graph. It permits us to balance a recognition result considering the similarity of the affected class and the real one. We can see on figure 6 that the metric performs globally correctly as the minimum value of the $SRCC$ is 73.61%, and can be up to 87.50% with a parameter $\alpha$ equals to 0.40 and a distance matrix. We can also remark that the use of a distance matrix enables better performance of the evaluation metric once the parameter $\alpha$ is correctly set.

![Figure 4: SRCC values by considering similarity of object classes (red) and without (blue)](image)
4.2 Study on properties

We defined some properties to compare evaluation metrics in the state of the art. The aim of this analysis is to validate these properties taking into account the subjective evaluation results.

4.2.1 Monotony

In order to verify if individuals order image understanding results the more they are altered, we have to check the p-values given by the Kruskal-Wallis test to be sure that responses are independent (p-values lower than 0.05), and we can also check that responses are correctly ordered. Five questions in the study present 2 or 3 images to be ordered with regards of monotony, and obtained results are presented in table 1. The p-values are 0 for all 5 questions, which clearly shows that these results are independent. Moreover, images are correctly ordered. We can conclude that the monotony property is expected by individuals.

<table>
<thead>
<tr>
<th>Question</th>
<th>p-value</th>
<th>Order of answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3</td>
<td>0</td>
<td>1.0000 2.1125 3.9250</td>
</tr>
<tr>
<td>Q4</td>
<td>0</td>
<td>1.0189 2.3208</td>
</tr>
<tr>
<td>Q6</td>
<td>0</td>
<td>1.3673 3.1020</td>
</tr>
<tr>
<td>Q9</td>
<td>0</td>
<td>1.0000 2.2639 3.9028</td>
</tr>
<tr>
<td>Q12</td>
<td>0</td>
<td>3.0244 3.8659</td>
</tr>
</tbody>
</table>

4.2.2 Symmetry

For this property, we expect that two images will be ordered in the same way for two opposite alterations (rotation of an angle $\theta$ or $-\theta$). We check if the p-values are higher than 0.05. As we can see in table 2, 2 out of 4 questions have a p-values higher than 0.05. The symmetry of images on question 3 is not correctly handled by individuals but is correct for question 9, where the alteration is the translation for both question. The symmetry of scale change alteration of question 6 is correctly managed by individuals, but not the perspective
alteration in question 12. The symmetry property is not as clear as the monotony property for individuals.

<table>
<thead>
<tr>
<th>Question</th>
<th>Q3</th>
<th>Q6</th>
<th>Q9</th>
<th>Q12</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.0003</td>
<td>0.5379</td>
<td>0.8944</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

4.2.3 Shape and size

Questions 3 and 9 present one object at different size with exactly the same alterations. We can quantify in this test if the size and shape of the original object affects the sorting results. For both questions, image understanding results are correctly ordered. That shows that individuals order images independently of the size or shape of the original object in the image.

4.2.4 Relative importance of alterations

One important question concerns the setting of the evaluation metric considering the relative importance of the different alterations. By analyzing the answers from 8 questions, we can conclude that the less penalizing alterations are the localization ones. That means that users considered other errors more important during their choices of sorting. Among the localization errors, we can identify in increasing order the importance of localization errors: perspective changes, translation, scale change and rotation.

Secondly, recognition errors are more penalized than localization ones. We notice that the class has an effect on evaluation: in question 12, the table recognized as a bed is better evaluated than if it is recognized as a horse. The combination of localization and recognition alterations are then more penalizing.

Last, the most penalizing alteration is the detection one. We can notice that the fusion...
of several objects in the ground truth detected as one object is the less penalized. We can conclude that under detection is less penalized than over-detection.

4.2.5 Reproducibility of evaluation

In order to verify if an individual can reproduce the evaluation, questions 2 and 10 contain exactly the same images. We can see in table 3 that image understanding results are ordered in the same way. We can conclude that individuals are able to reproduce their evaluation.

Table 3: Reproducibility: mean ordering of image understanding results of the same question

<table>
<thead>
<tr>
<th>Question</th>
<th>Order of answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2</td>
<td>1.2105 1.8772 3.0526 3.7368</td>
</tr>
<tr>
<td>Q10</td>
<td>1.1034 1.9483 3.0862 3.5690</td>
</tr>
</tbody>
</table>

5 Conclusion and perspectives

In this study, we present a subjective evaluation of image understanding results. We compare results from this evaluation to the evaluation performed by an evaluation metric we presented in [7]. Results show that the metric we defined is able to perform a correct judgment up to 87.50% of comparisons between understanding results similarly to individuals. Moreover, it shows that the default parameters provide a reliable judgment, but could be improved, by choosing a default value of 0.75 for the $\alpha$ parameter, or by using a matrix distance.

The second conclusion of this study is that properties chosen to evaluate metrics were correct. It also shows that individuals are able to reproduce their evaluation. Moreover, we show that alterations are not managed in the same way: localization alterations are the less penalizing, then comes recognition alteration and finally detection alteration. This is an interesting information for researchers in this field. Indeed, to increase the performance of
image understanding results, it is better to focus on penalizing alterations such as detection. Perspectives of this study concern the study of other alterations by analyzing some image understanding results.

References


List of Figures

1 Examples of understanding results on a single image and their associated score ........................................... 5
2 Principle of the evaluation metric of an image understanding result .......................................................... 7
3 Some questions from the questionnaire ........................................................................................................ 10
4 SRCC values by considering similarity of object classes (red) and without (blue) ........................................ 15
Baptiste Hemery is an assistant professor at ENSICAEN (France). He obtained his Phd from the University of Caen Basse-Normandie in 2009. He belongs to the GREYC laboratory in the image and computer security research units. His research interests concern image interpretation evaluation and biometric systems.

Helene Laurent is an assistant professor at ENSI of Bourges (France). She obtained her Phd from the university of Nantes in 1998. She belongs to the Prisme Institute of the Orleans university in the ISS research unit. Her research interests concern image processing evaluation.

Bruno Emile is an assistant professor at IUT of Châteauroux (France). He obtained his Phd from the university of Nice in 1996. He belongs to the Prisme Institute of the Orleans university in the ISS research unit. His research interests concern object detection and recognition.

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. His research interests concern evaluation of image processing and biometric systems.