SUBJECTIVE EVALUATION OF IMAGE UNDERSTANDING RESULTS

Baptiste Hemery[1], *Hélène Laurent*[2], *and Christophe Rosenberger*[1]

[1] GREYC Laboratory
ENSICAEN - University of CAEN - CNRS
6 Boulevard Maréchal Juin, 14000 CAEN, FRANCE
baptiste.hemery@greyc.ensicaen.fr
christophe.rosenberger@greyc.ensicaen.fr

[2] PRISME Institute
ENSI de Bourges - University of Orléans
88 Boulevard Lahitolle, 18020 BOURGES, FRANCE
helene.laurent@ensi-bourges.fr

ABSTRACT

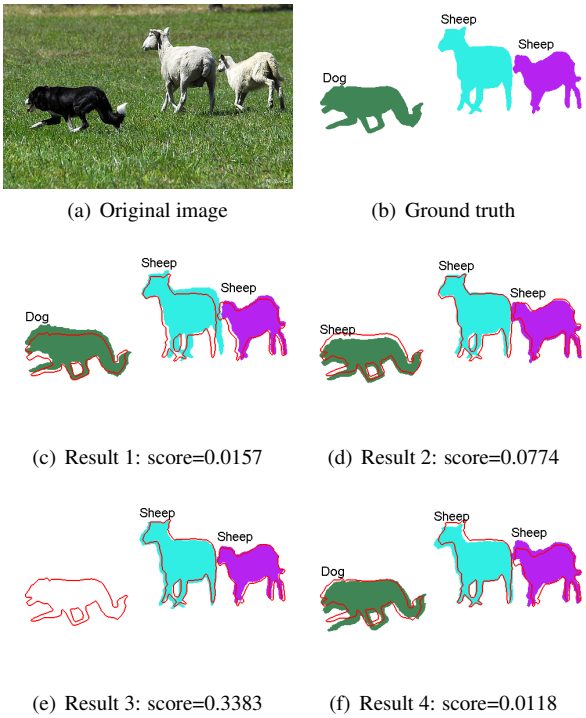
Image understanding has many applications. Given an image and a ground truth, it is possible to measure the quality of understanding results provided by different algorithms or parameters. In this paper, we ask some users to make a subjective evaluation of image understanding results by sorting them from the best to the worst. We compared the results with some provided by a metric we defined recently. Experimental results show the good behavior of this metric compared to the human judgment.

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, 10], 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 [6] or the French Robin Project [4]. Given a manually made ground truth, these competitions use metrics to evaluate and compare the results obtained by different localization 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. We think that it is then necessary to have a reliable quality score of an understanding result given the associated ground truth.

In a previous work [8], we proposed a metric that enables the evaluation of such results. The metric range from 0 to 1, the lower the score is, the better it is. As an example, figure 1 presents the evaluation obtained with this metric for four different 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. The aim of this work is to check if this metric corresponds to what can be obtained by an evaluation done by humans. In order to reach this objective, 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.



(a) Original image (b) Ground truth

(c) Result 1: score=0.0157 (d) Result 2: score=0.0774

(e) Result 3: score=0.3383 (f) Result 4: score=0.0118

Figure 1: Examples of understanding results on a single image and their associated score

This paper is organized as follows: the first part briefly presents our previous work on the evaluation metric for image understanding results. We then present the principles of the conducted subjective evaluation. The results are presented as well as the conclusions of this study.

2. PREVIOUS WORKS

In a previous work [9], we studied the different existing localization metrics in the literature. In order to compare these metrics, we defined an evaluation protocol (see figure 2): we alter the ground truth and check if results given by a metric fulfill some properties. A correct metric should fulfill most of

 ISSN 2076-1465

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.

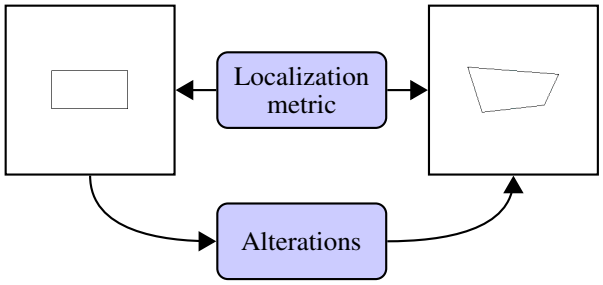


Figure 2: Evaluation protocol of localization metric

Based on these properties and results from [9], we defined a metric that enables the evaluation of understanding results in [8]. As far as we know, there is no other metric that can evaluate such a result. The metric is composed of four stages, as we can see on figure 3: (i) Matching objects, (ii) Local evaluation, (iii) Over- and Under- detection compensation and finally (iv) Global evaluation score computation.

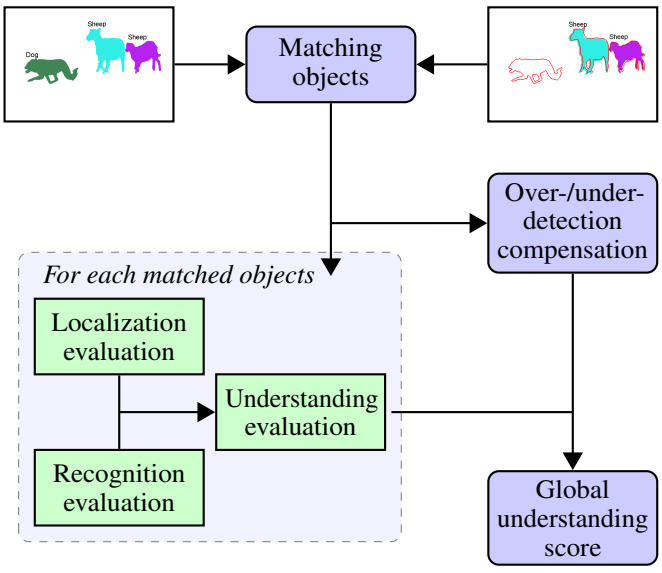


Figure 3: Image understanding metric

The first stage is necessary to match objects from the ground truth and from the understanding result. The local evaluation stage corresponds to the evaluation of each matched object. We first evaluate the localization of the object and then its recognition. Given these two scores, we compute the local score as the combination of the localization and the recognition scores. Then, 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. Several parameters enable to tune the metrics. We can, for example, provide a distance matrix between each class present in the databases, which will enable to better evaluate recognition mistakes. We can also use a parameter α to balance the weight of localization and recognition evaluation in the local score. Results, presented in [8], show that the proposed metric enables the evaluation of image understanding results. However, we would like to know its relative behavior compared to the evaluation done by humans. That is why we asked many individuals to evaluate image understanding results. This subjective evaluation of the metric is presented in the next section.

3. SUBJECTIVE EVALUATION

This subjective evaluation of image understanding results has two goals. The first one is to compare results obtained by our evaluation metric and those obtained from humans. This will enable us to check if our metric gives a human like evaluation of image understanding results. The second goal of this study is to check if the properties defined in [9] are naturally fulfilled by the judgments of humans.

3.1 Data acquisition

In order to acquire feedbacks from individuals, we created a web site where a user can create an account and then answer to questions. Questions in this questionnaire present the original image, the ground truth and four image understanding results. An example of question can be seen in figure 4. The user is asked to order image understanding results from the most to the less similar to the ground truth.

3.2 Questions

The study is composed of 12 questions. The 12 original images used in this study come from the Pascal database [6], where the original image and the ground truth is provided. The corresponding taxonomy, according to the one used in Calltech256 [7], is presented in figure 5. For the first goal of this study, which is to compare our metric to human evaluation, all questions will be used.

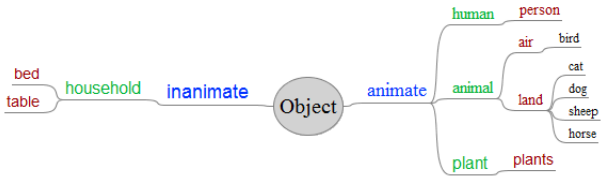


Figure 5: Taxonomy of objects present in the study

However, concerning the properties, questions were specifically designed to answer 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. The second property is the symmetry and 4 questions are

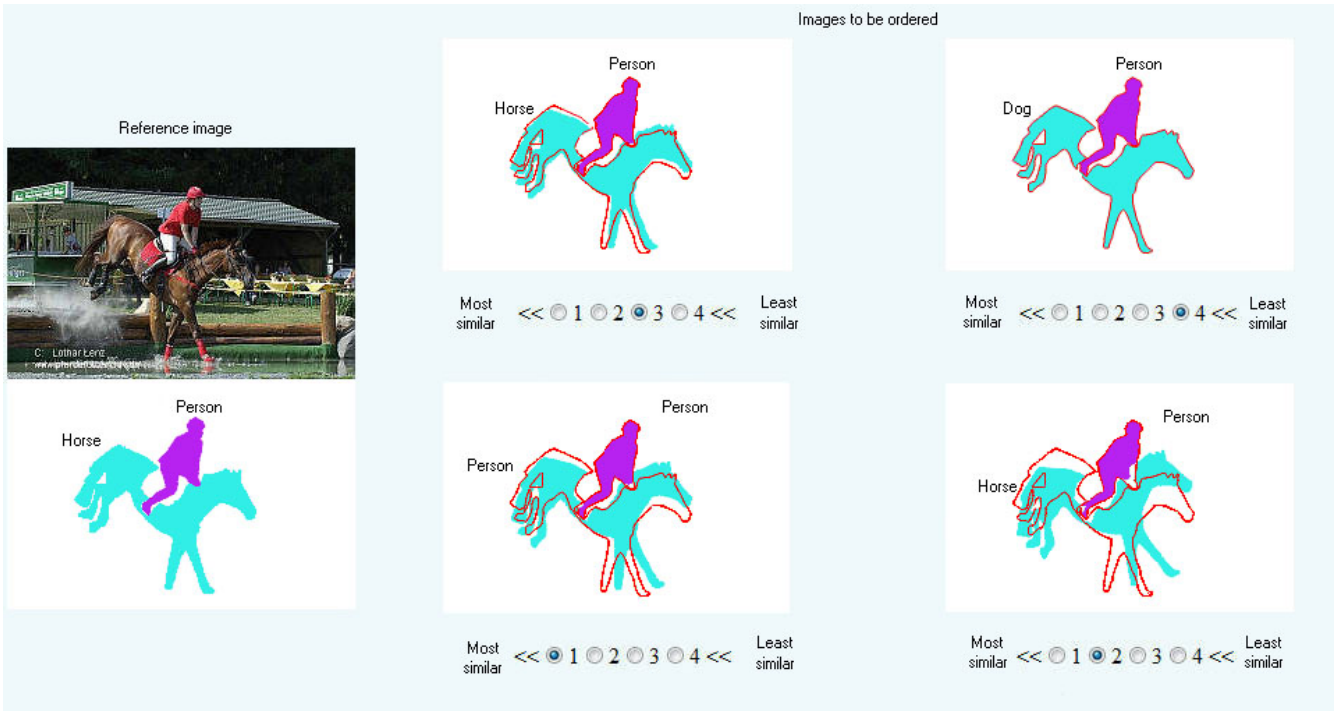


Figure 4: One question from the questionnaire

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.

4. DEVELOPED METHOD

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 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 (ordering of understanding results).

4.1 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.

$$Pearson(X_i, E[X]) = \frac{Cov(X_i, E[X])}{\sqrt{Cov(E[X], E[X]).Cov(X_i, X_i)}} \quad (1)$$

where X_i represents the answers of the user i, $E[X]$ represents the average value of answers given by users and $Cov(.,.)$ 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.

$$\begin{cases} Pearson(X_i, E[X]) \geq \theta & acceptX_i \\ otherwise & rejectX_i \end{cases} \quad (2)$$

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

4.2 Evaluation of global performances of the metric

As we have relative measures, we can compare the quality of different image understanding results and sort them as in [5]. 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 results given by individuals and by using our 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):

$$SCC = \sum_{k=1}^{12} \sum_{i=1}^{6} |I(i,k) - M(i,k)| \quad (3)$$

where $I(i,k)$ and $M(i,k)$ are respectively the individuals and the metric results for the ith 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:

$$SRCC = (1 - \frac{SCC}{SCC_{max}}) * 100 \quad (4)$$

where SCC_{max} corresponds to the biggest difference of the 72 comparison results. In our case, $SCC_{max} = \binom{4}{2} * 12 * 2 = 144$. 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.

4.3 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.

$$\begin{cases} H0: \mu_1 = \mu_2 = ... = \mu k \\ H1: \exists i, j, \mu_i \neq \mu_j \end{cases} \quad (5)$$

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.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \, \bar{r}_i 2 - 3\,(N+1) \quad (6)$$

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.

$$\bar{r}_i 2 = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i} \; and \; \bar{r} = \frac{1}{2}(N+1)$$

$$\begin{cases} p-value \geq 0.05 & accept\ H_0 \\ otherwise & reject\ H_0 \end{cases} \quad (7)$$

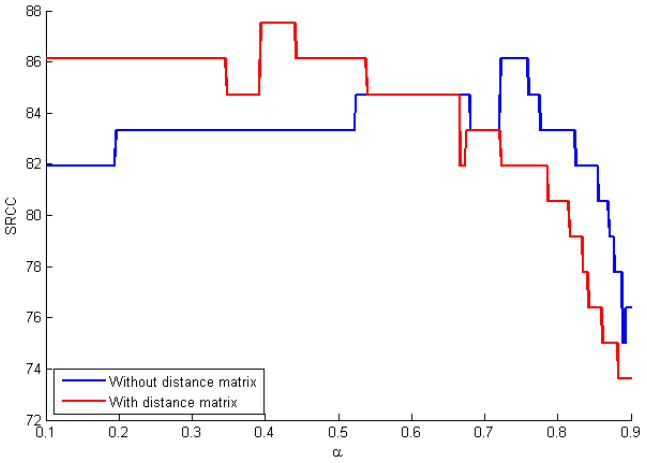


Figure 6: SRCC

5. EXPERIMENTAL RESULTS

5.1 Global behavior of the evaluation metrics

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

We then present in figure 6 the evolution of the *SRCC* as a function of the parameter α, and with or without using a distance matrix. The distance matrix used for the evaluation is computed from the taxonomy presented in figure 5: the distance between two classes depends on their distance on the graph. It permits us to balance a recognition results considering the similarity of the affected class and the real one. We can see 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 α 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 α is correctly set.

5.2 Study on properties

5.2.1 *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.

5.2.2 *Symmetry*

For this property, we expect that two images will be ordered in the same way, so 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

Table 1: Monotony: p-value and mean ordering of image understanding results

Question	p-value	Order of answer		
Q3	0	1.0000	2.1125	3.9250
Q4	0	1.0189	2.3208	
Q6	0	1.3673	3.1020	
Q9	0	1.0000	2.2639	3.9028
Q12	0	3.0244	3.8659	

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 2: Symmetry: p-value of questions

Question	Q3	Q6	Q9	Q12
p-value	0.0003	0.5379	0.8944	0.0000

5.2.3 Shape and size

Questions 3 and 9 present one object with exactly the same alteration. We can see if the size and shape of the original object affect the ordering. For both questions, images are correctly ordered, which shows that individuals order images independently of the size or shape of the original object in the image.

5.2.4 Most penalizing alteration

By analyzing the order of answer from 8 questions, we can conclude that the less penalizing alterations are the localization ones, in order: perspective changes, translation, scale change and rotation. The recognition alteration errors are less penalized. We notice than 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. Then comes the combination of localization and recognition alteration before detection alteration. Among detection alteration, the fusion of several objects in the ground truth detected as one object is the less penalized, then over-detection and finally the under-detection.

5.2.5 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

Question	Order of answers			
Q2	1.2105	1.8772	3.0526	3.7368
Q10	1.1034	1.9483	3.0862	3.5690

6. 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 our metric presented in [8]. 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 default parameter is quite good, but could be improved, by choosing a default value of 0.75 for the α 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.

Perspectives concern the definition of optimal weighting coefficients of alterations in the metric we defined to maximize the adequacy to the human judgment.

REFERENCES

[1] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[2] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2003.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition (CVPR)*, 1:886–893, 2005.

[4] E. D'Angelo, S. Herbin, and M. Ratiéville. Robin challenge evaluation principles and metrics, Nov. 2006. http://robin.inrialpes.fr.

[5] C. Delgorge, C. Rosenberger, G. Poisson, and P. Vieyres. Evaluation of the quality of ultrasound image compression by fusion of criteria with a genetic algorithm. In *International Conference on Advances in Pattern Recognition (ICAPR)*, volume 3687, pages 464–472. LNCS, 2005.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop.

[7] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, http://authors.library.caltech.edu/7694, 2007.

[8] B. Hemery, H. Laurent, and C. Rosenberger. Evaluation metric for image understanding. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 4381–4384, Nov. 2009.

[9] B. Hemery, H. Laurent, C. Rosenberger, and B. Emile. Evaluation protocol for localization metrics - application to a comparative study. In *Proceedings of the 3rd international conference on Image and Signal Processing*, pages 273–280, 2008.

[10] F. Jurie, C. Schmid, I. Gravir, and F. Montbonnot. Scale-invariant shape features for recognition of object categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 90–96, 2004.