Evaluation Metric for Image understanding

Baptiste Hemery, Hélène Laurent, Christophe Rosenberger

EVALUATION METRIC FOR IMAGE UNDERSTANDING

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

[1] GREYC laboratory
ENSICAEN - Université de Caen - CNRS
6 boulevard Maréchal Juin
14000 Caen, France

[2] PRISME Institute
ENSI de Bourges
88 boulevard Lahitolle
18020 Bourges Cedex, France

ABSTRACT

We propose in this paper a new evaluation metric that enables to quantify the quality of an image interpretation result. This metric takes into account the *a priori* knowledge used by the interpretation algorithm and the ground truth associated with the original image. We combine two metrics that evaluate the localization and recognition results of each detected object. We show that the proposed metric fulfills some theoretical properties and has a correct behavior face to empirical experiments on an image benchmark database. We think that this metric could be a reliable reference for image and video understanding competitions.

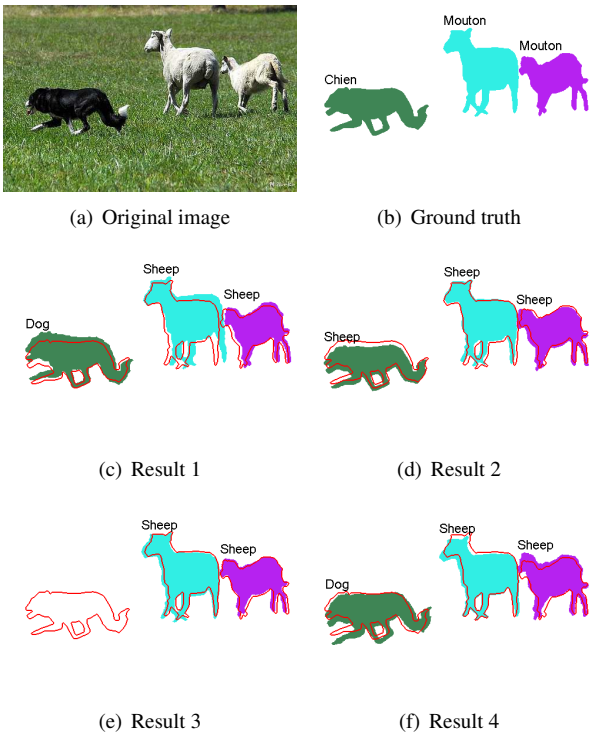
***Index Terms*—** image understanding, object localization, object recognition, evaluation metric.

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, 6] or the French Robin Project [7]. 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. It could be 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 segmentation evaluation or image retrieval evaluation can be found in the literature and should reveal themselves relevant for the evaluation of image understanding results. In our work, we intend to define a reliable quality score of an interpretation result. As for example in Figure 1, we would like to distinguish automatically the best result.



(a) Original image (b) Ground truth

(c) Result 1 (d) Result 2

(e) Result 3 (f) Result 4

Fig. 1. Examples of interpretation results on a single image

We studied 33 different localization metrics from the state

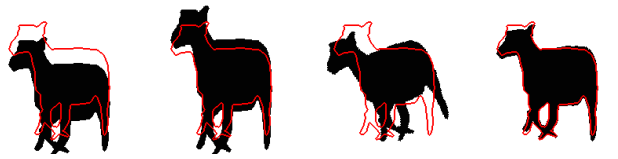
This work was made possible with the financial support of the French Higher Education and Research Ministry.

of the art by considering some theoretical properties and an experimental study. The results of this comparative study are presented in section 2. Section 3 presents the proposed quality score of an image understanding result. We present some experimental results in section 4. Finally, conclusions and perspectives of this work are given in section 5.

2. BACKGROUND

In a previous work [8], we aimed to evaluate the quality of metrics for the evaluation of localization results. We first referenced up to 33 different metrics allowing to evaluate a localization result. Some of these metrics were not created with the specific purpose of localization evaluation, but for segmentation or image quality evaluation.

We then evaluated each of these metrics. To do so, 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 2, some examples of alterations. We finally obtained a total of **118.080** synthetic localization result images.



(a) Translation (b) Scale change (c) Rotation (d) Perspective

Fig. 2. Four examples of alterations

We computed the metric between the ground truth and each synthetic result and obtained a 3D curve for each triplet {metric, alteration, ground truth}. From these curves, we verified if all the metrics fulfill some properties. The chosen properties that a metric should fulfill to correctly evaluate localization results:

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 most properties are fulfilled, the better is the metric. The conclusion of [8] was to use a region based metric, and more particularly PAS [5, 6], MAR_{lce}, MAR_{gce} [9] or VIN [10] metrics.

3. DEVELOPED METHOD

The method 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.

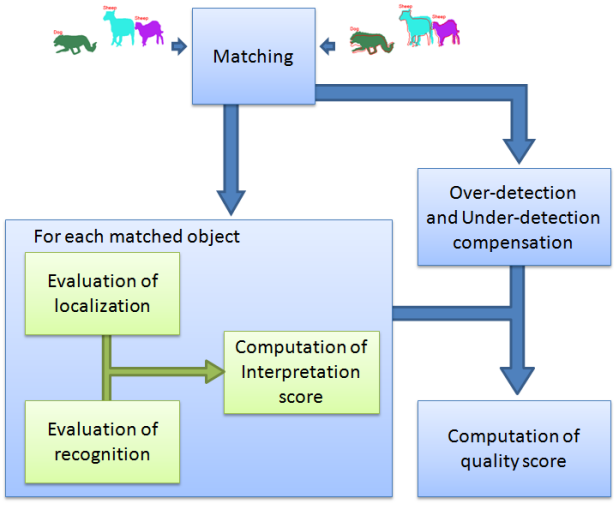


Fig. 3. Principle of the evaluation process

The first stage is necessary to match objects from the ground truth and from the interpretation result. Moreover, this enables the detection of missed objects (over-detection) and the detection of multiple detection (under-detection). To match objects, we compute a matching score matrix as in [11]. 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 recovery of objects. The recovery is computed with the PAS metric:

$$PAS(I_{gt}, I_i) = \frac{\text{Card}(I_{gt}^r \cap I_i^r)}{\text{Card}(I_{gt}^r \cup I_i^r)} \quad (1)$$

with $\text{card}(I_{gt}^r)$ the number of pixels from the object in the ground truth, and $\text{card}(I_i^r)$ the number of pixels from the detected object in the interpretation result. The matching scores range from 0 to 1, 1 corresponds to a perfect localization. We perform then the assignment with the Hungarian algorithm [12]. An example of such a matrix can be found in Figure 4. It shows 4 objects in the ground truth, 5 objects in the interpretation result and 3 matched objects. The fourth object in the ground truth is not detected (under-detection) and two objects in the interpretation result do not correspond to an object in the ground truth (over-detection). We decided to associate only one object per object in the ground truth as in [6] and not to look after multiple detection of the same as in [11, 13].

The local evaluation stage corresponds to the evaluation of each matched object k. We first evaluate the localization

0,65				
			0,76	
0,09		0,72		

(a) Matching score matrix

1				
			1	
		1		

(b) Assignment matrix

Fig. 4. Example of matching score and assignment matrix

of the object and then its recognition. The evaluation of the localization is the Martin's one [9] adapted to one object:

$$S_{loc}(I_{gt}, I_i, k) = \frac{1}{\text{card}(I)} \min\left(\frac{\text{card}(I_{gt\setminus i}^{r(k)})}{\text{card}(I_g^{r(k)}t)}, \frac{\text{card}(I_{i\setminus gt}^{r(k)})}{\text{card}(I_i^{r(k)})}\right) \quad (2)$$

with $\text{card}(I)$ the number of pixels in the image and $\text{card}(I_{gt\setminus i}^{r(k)})$ the number of pixels present in the ground truth object k and not present in the detected object. The evaluation of the recognition part aims to compare the class of the object in the ground truth and in the interpretation result. We give the score $S_{rec}(I_{gt}, I_i, k) = 0$, if classes are the same and 1 otherwise.

Given these two scores, we compute the interpretation score $S(k)$ as the combination of the localization and the recognition scores. We use a parameter α, set at 0.8, to balance these two scores. A lower value of this parameter gives too much impact to the recognition score.

$$S(k) = \alpha * S_{loc}(I_{gt}, I_i, k) + (1 - \alpha) * S_{rec}(I_{gt}, I_i, k) \quad (3)$$

After the computation of the local score for each matched object, we obtain an interpretation score matrix as presented in Figure 5. Then, the third stage aims at compensating the under-detection and over-detection. This stage fills rows and columns without score with 1, giving the final score matrix. Finally, the global score is computed as the mean of local scores.

4. EXPERIMENTAL RESULTS

In this section, we study the evolution of the global score face to different alterations of interpretation results. There are three kinds of possible alteration: wrong localization, wrong recognition, forgotten or wrongly added.

To study the evolution of the global score face to localization, we use images from the Pascal VOC Challenge 2008 [6]. Among the database, we randomly choose 6 images within 2 to 6 objects per image. We then randomly choose two objects per image and alter them with the four alterations presented in

0,13				
			0,01	
		0,09		

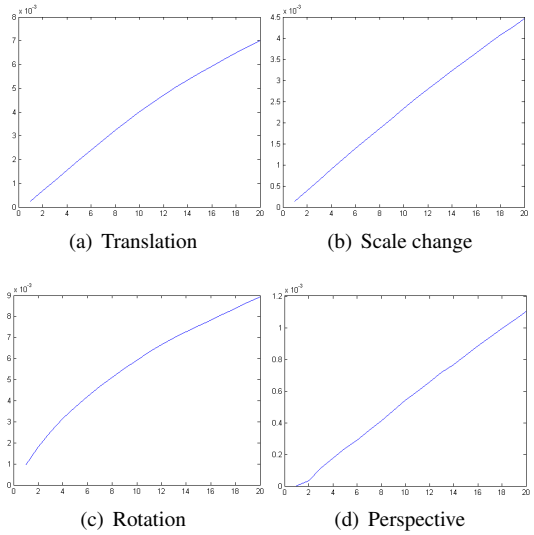
(a) Interpretation score matrix

0,13				
			0,01	
		0,09		
	1			
				1

(b) Interpretation score matrix with compensation

Fig. 5. Example of local score matrix and local score one with compensation

Figure 2: translation, scale change, rotation and perspective alteration. The mean result of the global score face to each alteration is presented in Figure 6. Curves show the global score versus the power of the alteration. We can see that the more we alter the interpretation result, the more the global score penalizes the interpretation result. Moreover, we can see that the translation and rotation alterations are more penalized than the scale change alteration and even more than the perspective one. This seems correct regarding Figure 2, where all images are altered with an alteration power of 20.



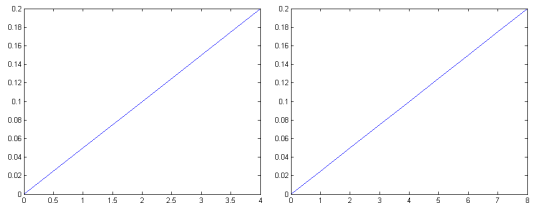
(a) Translation (b) Scale change

(c) Rotation (d) Perspective

Fig. 6. Results for the localization part

Concerning the recognition part, we study how evolves the global score when the class of the localized object is not correct, as for example, when a dog is recognized as a sheep. To do this, we alter the class of objects and study the evolution of the global score. Figure 7 shows the evolution of the global score face to the number of altered objects, for 2 ground truths: one with 4 objects and the other with 8 ob-

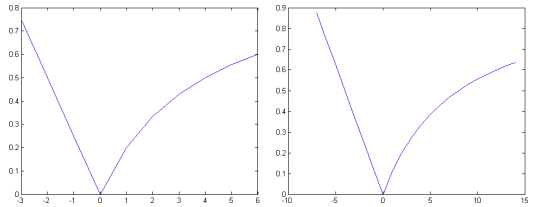
jects. We can see that the more there are altered objects, the more the global score penalizes the interpretation result.



(a) 4 objects (b) 8 objects

Fig. 7. Results for the recognition part

Finally, we study the impact of missing objects and over-detected ones. Figure 8 shows the evolution of the global score face to the number of missing objects or over-detected for 2 ground truths with 4 and 8 objects. We can see that these situations are correctly penalized. Moreover, missing an object is more penalized than over-detecting some objects.



(a) 4 objects (b) 8 objects

Fig. 8. Results for over- and under-detection

If we compute the global interpretation score for images in Figure 1, we obtain the following results: (result 1: 0.0157), (result 2: 0.0774), (result 3: 0.3383) and (result 4: 0.0118). Result 1 and 4 have better scores since all objects are correctly recognized even if the localization is less precise than for 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 an object is missing. Moreover, if we compute the score of the ground truth, the obtained score is 0 as expected.

5. CONCLUSIONS AND PERSPECTIVES

Results show that the proposed quality score enables the evaluation of image understanding results. We penalize the different errors from the most important to the less one: forgotten object, wrong recognition and bad localization. The proposed metric fulfills many desired properties and has a correct behavior face to different situations. Note that the importance of recognition face to localization can be adjusted for a particular application.

Perspectives are to improve the balance between the localization score, the recognition score and the compensation for missed and over-detected objects. Moreover, we want to improve the recognition score by introducing a distance matrix between classes.

6. REFERENCES

[1] 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*, vol. 25, no. 10, pp. 1337–1342, 2003.

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

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

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

[5] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M.Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G.Dorko, et al., "The 2005 pascal visual object classes challenge," 2005, http://www.pascal-network.org/challenges/VOC/.

[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/index.html.

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

[8] 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*, 2008, pp. 273–280.

[9] D. Martin, C. Fowlkes, D. Tal, and 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*, vol. 2, pp. 416–423, July 2001.

[10] L. Vinet, *Segmentation et mise en correspondance de régions de paires d'images stéréoscopiques*, Ph.D. thesis, Université de Paris IX Dauphine, Juillet 1991.

[11] I.T. Phillips and A.K. Chhabra, "Empirical performance evaluation of graphics recognition systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, vol. 21, no. 9, pp. 849–870, 1999.

[12] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, pp. 32, 1957.

[13] C. Wolf and J.M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *International Journal on Document Analysis and Recognition*, vol. 8, no. 4, pp. 280–296, 2006.