Evaluation of Human Detection Algorithms in Image Sequences

Yannick Benezeth, Baptiste Hemery, Hélène Laurent, Bruno Emile, Christophe Rosenberger

Evaluation of Human Detection Algorithms in Image Sequences

Yannick Benezeth[1], Baptiste Hemery[2], Hélène Laurent[3a],
Bruno Emile[3b], and Christophe Rosenberger[2]

[1] Orange Labs
4 rue du Clos Courtel
35510 Cesson-Sévigné
France

[2] Laboratoire GREYC
ENSICAEN - Université de Caen
- CNRS
6 bd du Maréchal Juin
14000 Caen, France

[3a] ENSI de Bourges, Institut PRISME
88 bd Lahitolle, 18020 Bourges Cedex
[3b] Institut PRISME, Université d'Orléans
2 av F. Mitterrand, 36000 Châteauroux
France

Résumé This paper deals with the general evaluation of human detection algorithms. We first present the algorithms implemented within the $CAPTHOM$ project dedicated to the development of a vision-based system for human detection and tracking in an indoor environment using a static camera. We then show how a global evaluation metric we developped for the evaluation of understanding algorithms taking into account both localization and recognition precision of each single interpretation result, can be a useful tool for industrials to guide them in the elaboration of suitable and optimized algorithms.

1 Introduction

Face to the huge development of image interpretation algorithm dedicated to various applications [1,2,3], such as target detection and recognition or video surveillance to name a few, the need of adapted evaluation metrics, which could help in a development of well thought-out algorithm or in the quantification of the relative performances of different algorithms, has become crucial. Wide annotated databases and metrics have been defined within several research competitions such as the Pascal VOC Challenge [4] or the French Robin Project [5] in order to evaluate object detection and recognition algorithms. Whatever these metrics either focus on the localization aspect or the recognition one, but not both together. Moreover, concerning the recognition objective, most of the competitions use Precision/Recall and ROC curves [4,6,7], evaluating the algorithms on the whole database. An interpretation evaluation metric, taking into account both aspects and working on a single interpretation result, is then needed.

This article presents our works concerning the development of vision-based systems for human detection and tracking in a known environment using a static camera and the definition of an adaptable performance measure able to simultaneously evaluate the localization, the recognition and the detection of interpreted objects in a real scene using a manually made ground truth. If in a general way, the localization and the recognition have to be as precise as possible, the relative importance of these two aspects can change depending of the foreseen application. We describe in section 2 the successive algorithms implemented for the $CAPTHOM$ project which more particularly focused

on indoor environments. The proposed evaluation metric of a general image interpretation result is presented in section 3. Its potential interest is illustrated in section 4 on the $CAPTHOM$ project. Section 5 presents conclusions and perspectives of this study.

2 Visual-based system developments for human detection in image sequences

Within the $CAPTHOM$ project, we attempt to develop a human detection system to limit power consumption of buildings and to monitor low mobility persons. This project belongs to the numerous applications of human detection systems for home automation, video surveillance, etc. The foreseen system must be easily tunable and embeddable, providing an optimal compromise between false detection rate and algorithmic complexity.

The development of a reliable human detection system in videos deals with general object detection difficulties (background complexity, illumination conditions etc.) and with other specific constraints involved with human detection (high variability in skin color, weight and clothes, presence of partial occlusions, highly articulated body resulting in various appearances etc.). Despite of these difficulties, some very promising systems have already been proposed in the literature. It is especially the case of the method proposed by Viola and Jones [8] which attempts to detect humans in still images using a well-suited representation of human shapes and a classification method. We first of all implemented this method in a sliding window framework analyzing every image and using several classifiers. This method is based on Haar-like filters and adaboost. In an indoor environment, partial occlusions are actually frequent. The upper part of the body (head and shoulders) is often the only visible part. As it is clearly insufficient to seek in the image only forms similar to the human body in its whole, we implemented four classifiers: the whole body, the upper-body (front/back view), the upper-body (left view) and the upper-body (right view). In a practical way, the classifier analyzes the image with a constant shift in the horizontal and vertical direction. As the size of the person potentially present is not known a priori and the classifier has a fixed size, the image is analyzed several times by modifying the scale. The size of the image is divided by a scale factor (sf) between two scales. This method is called $Viola[8]$ in the following paragraphs.

In order to reduce the search space of classifiers localizing regions of interest in the image, we added a change detection step based on background subtraction. We chose to model each pixel in the background by a single Gaussian distribution. The detection process is then achieved through a simple probability density function thresholding. This simple model presents a good compromise between detection quality, computation time and memory requirements [9,10]. The background model is updated at three different levels: the pixel level updating each pixel with a temporal filter allowing to consider long time variations of the background, the image level to deal with global and sudden variations and the object level to deal with the entrance or the removal of static

objects. This method is called $Viola[8] + BS$ afterwards.

We finally developped a method using additionally temporal information. We propose a method using advantages of tools classically dedicated to object detection in still images in a video analysis framework. We use video analysis to interpret the content of a scene without any assumption while objects nature is determined by statistical tools derived from object detection in images. We first use background subtraction to detect objects of interest. As each connected component detected potentially corresponds to one person, each blob is independently tracked. Each tracked object is characterized by a set of points of interest. These points are tracked, frame by frame. The position of these points, regarding connected components, enables to match tracked objects with detected blobs. The tracking of points of interest is carried out with the pyramidal implementation of the Lucas and Kanade tracker [11,12]. The nature of these tracked objects is then determined using the previously described object recognition method in the video analysis framework. Figure 1 presents an example of tracking result with partial occlusion. This method is called $CAPTHOM$ in the following.

FIGURE 1. Illustration of tracking result with a partial occlusion. First row: input images with interest points associated with each object (one color per object), second row: tracking result

For more information about the three considered methods, the interested reader can refer to [13].

3 Evaluation metric

The developed evaluation metric [14] is based on four steps corresponding to: (i) Objects matching, (ii) Local evaluation of each matched object in terms of localization and recognition, (iii) Over- and under-detection compensation and (iv) Global evalua-

tion score computation of the considered interpretation result.

Figure 2 illustrates the different stages on an original image extracted from the 2007 Pascal VOC challenge. For this image, the ground truth is composed of 4 objects which all belong to the human class. The interpretation result contains as for it two detected persons. We can note that the first person of the ground truth is well localized and recognized. The last three persons are well recognized but poorly localized. Indeed, only one object has been detected instead of three.

The first step, consisting in matching the objects of the ground truth and of the interpretation result, is done using the PAS metric [4]:

$$PAS(I_{gt}, I_i, u, v) = \frac{\mathrm{Card}(I_{gt}^{r(u)} \cap I_i^{r(v)})}{\mathrm{Card}(I_{gt}^{r(u)} \cup I_i^{r(v)})} \tag{1}$$

with $\mathrm{card}(I_{gt}^{r(u)})$ the number of pixels from the object u in the ground truth, and $\mathrm{card}(I_i^{r(v)})$ the number of pixels from the detected object v in the interpretation result. The number of rows of the resulting matching score matrix 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. This matrix is computed, as in [15]. The values range from 0 to 1, 1 corresponding to a perfect localization. From the matching score matrix, we can match objects by two methods: the first one consists in using an Hungarian algorithm, which implies one-to-one matching as in [4] ; the second one consists in simply applying a threshold, which enables multiple detections as in [16]. We use the threshold method, with a threshold set to 0.2 by default, as it allows that each object of the interpretation result can be assigned to several objects from the ground truth or vice-versa. The first person of the ground truth (object 1) is well localized in the interpretation result (object 2). Their recovery score exceeding the threshold, they are matched resulting in value 1 in the corresponding cell of the assignment matrix. Concerning the persons group, only two objects of the ground truth (objects 3 and 4) are matched with the one object of the interpretation result (object 1).

The second step consists in the local interpretation evaluation of each matched object. The localization is first evaluated using the $Martin$ metric [17] adapted to one object:

$$S_{loc}(I_{gt}, I_i, u, v) = \min\left(\frac{\mathrm{card}(I_{gt\setminus i}^{r(u)})}{\mathrm{card}(I_{gt}^{r(u)})}, \frac{\mathrm{card}(I_{i\setminus gt}^{r(v)})}{\mathrm{card}(I_i^{r(v)})}\right) \tag{2}$$

with $\mathrm{card}(I_{gt}^{r(u)})$ the number of pixels of object u present in the ground truth and $\mathrm{card}(I_{gt\setminus i}^{r(u)})$ the number of pixels of object u present in the ground truth but not present in the interpretation result. This metric has been chosen according to the comparative study conducted in [18] on the performances of 33 localization metrics face to different alterations like translation, scale change, rotation... The obtained localization score ranges from 0 to 1, 0 corresponding to a perfect recovery between the two objects

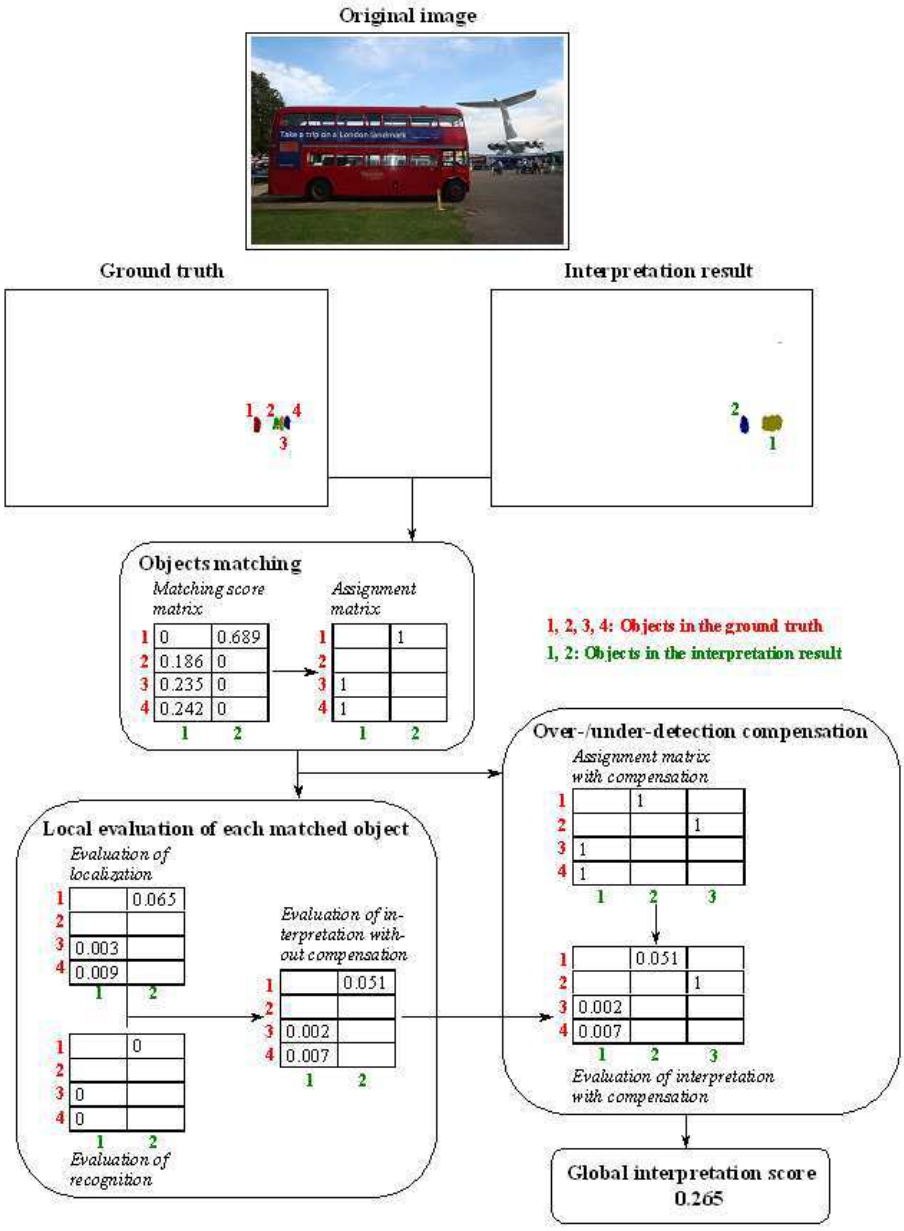

FIGURE 2. Illustration of the global evaluation of an interpretation result

and consequently to a perfect localization. We can note that all the matched objects are quite well localized obtaining low scores, the poorest score 0.065 corresponding to the second object of the interpretation result, namely the lonely person. The evaluation of

the recognition part consists in comparing the class of the object in the ground truth and in the interpretation result. This comparison can be done in different ways. A distance matrix between each class present in the database can be for example provided, which would enable to precisely evaluate recognition mistakes. On an other way, numerous real systems track one specific class of objects and do not tolerate some approximation in the recognition step. They work in an all or nothing scheme. $S_{rec}(I_{gt}, I_i, u, v) = 0$ if classes are the same and 1 otherwise. It is the case in the developped human detection system where all detections correspond *de facto* to the right class, namely a human. The recognition evaluation matrix containing only ones, the misclassification is then indirectly highly penalized through the over and under-detection compensation. As we have to maintain an important weight for the penalization of bad localization, we choose a high value of the α parameter ($\alpha = 0.8$). We finally compute the local interpretation score $S(u, v)$ between two matched objects as a combination of the localization and the recognition scores:

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

The third step is the compensation one. Working on the assignment matrix, empty rows or columns are tracked and completed. In our example, there is no empty column meaning that all objects of the interpretation result have been matched with at least one object of the ground truth. There is consequently no over-detection. On the other hand, one row (2) is empty ; one object of the ground truth has not been detected. This under-detection is compensated adding one column with score 1 at the corresponding line.

Finally, the global interpretation score is computed, taking into account the compensation stage and averaging the local interpretation scores.

4 Evaluation of human detection algorithms

In order to evaluate the detection methods presented in section 2, we realized a set of reference scenarios corresponding to the specific needs expressed by the industrial partners involved in the CAPTHOM project. An extract of a scenario example is presented in figure 3. At each location, a set of characteristics (temperature, speed, posture, activity...) is associated with the formalism defined within the CAPTHOM project [19].

The three classes of scenarios from which we have built the evaluation dataset are:

- Set 1: scenarios involving a normal use of a room. In these scenarios, we need to detect humans that are static or moving, sitting or standing in offices, meeting rooms, corridors and dining rooms.
- Set 2: scenarios of unusual activities (slow or fast falls, abnormal agitation).
- Set 3: scenarios gathering all false detections stimuli (illumination variation, moving objects etc).

In the following, Set 4 is defined as the union of these 3 sets. In total, we used 29 images sequences in 10 different places. Images have a resolution of 320 x 240 and have an "average" quality. Each images sequence lasts from 2 to 10 minutes.

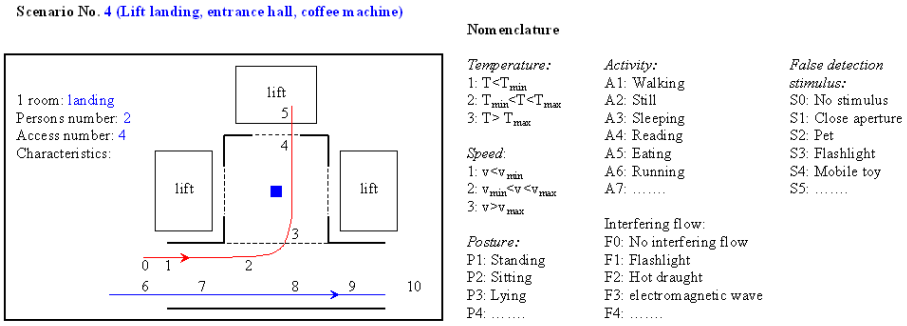

FIGURE 3. Extract of a scenario example defined by the industrial partners involved in the *CAPTHOM* project.

Figures 4 and 5 present results obtained with the $CAPTHOM$ algorithm on videos extracted from our test dataset.

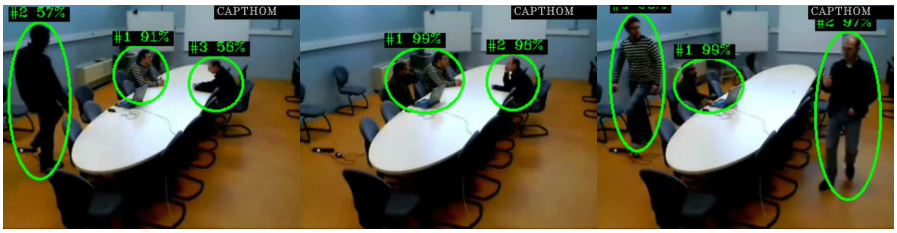

FIGURE 4. Example of results obtained with the $CAPTHOM$ method on a video presenting partial occlusion

The choice of the evaluation metric parameters, done for this study, corresponds to an expected interpretation compromise which can be encountered in many real applications. We use a parameter α, set at 0.8, to balance the localization and the recognition scores. This high value has been chosen to maintain an important weight for the penalization of bad localization. It results from a wide subjective evaluation of interpretation results we conducted, involving researchers of the French community, to better understand when a bad localization is more penalizing than a misclassification [20]. One objective of this study was to be able to guide the users in the metric parameters choice and more specifically in the α ponderation parameter choice. 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 the proposed metric. With $\alpha = 0.8$, the obtained similarity rate of correct comparison was 83.33%, which shows that our metric is able to order image understanding results correctly in most of cases. Preserving good performances concerning the localization aspect will allow our system to achieve higher level information such as path or activity

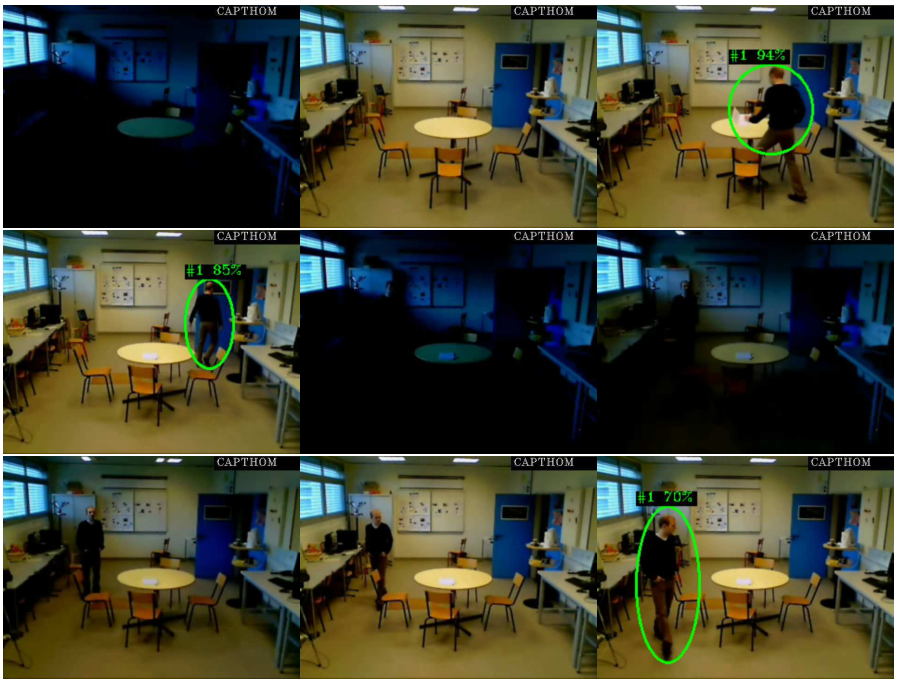

FIGURE 5. Example of results obtained with the $CAPTHOM$ method on a video presenting illumination changes

estimation.

Table 1 presents the mean evaluation results obtained for the three methods on the various sets of the test database using the designed interpretation evaluation metric. sf corresponds to the scale factor used from the sliding window framework analysis. We can note that the introduction of background subtraction results in algorithms that are less sensitive to the choice of this parameter. Combining properly defined test databases and an tunable evaluation metric allow the industrials to obtain a deep insight into their research developments.They can indeed quantify the performances gap between different algorithms and motivate their further technological choices. The proposed evaluation metric is also suitable for the choice of the algorithms parameters.

5 Conclusion and perspectives

We presented in this paper the potential interest of a global evaluation metric for the development of industrial understanding algorithms. The originality of the proposed measure lies in its ability to simultaneously take into account localization and recognition aspects together with the presence of over- or under-detection. Concerning the foreseen application, industrial partners involved in the project also have in mind to

	Set 1	Set 2	Set 3	Set 4
Viola [8], sf=1.1	**0.614**	**0.672**	0.565	0.597
Viola [8], sf=1.4	0.719	0.707	0.105	**0.436**
Viola [8], sf=1.5	0.758	0.739	**0.092**	0.451
Viola [8]+BS, sf=1.1	**0.429**	**0.642**	**0.050**	**0.276**
Viola [8]+BS, sf=1.4	0.618	0.747	0.071	0.380
Viola [8]+BS, sf=1.5	0.663	0.745	0.082	0.405
CAPTHOM	**0.338**	**0.089**	**0.043**	**0.176**

TABLE 1. Performances evaluation of the different interpretation algorithms developped within the CAPTHOM project.

extend the system for car park video surveillance. In that case, the detection and distinction between different classes could be interesting and give even more sense to the misclassifcation error introduced in the evaluation metric. We are actually working on the use of taxonomy information for ponderating the misclassification error. The introduction of a distance matrix between classes taking into account their more or less important similarity could improve the adaptability of the proposed metric. For some applications, some misclassifications could have less repercussions than others. As an example, it could be suitable to less penalize an interpretation result where a bus is recognized as a truck, as these two objects are very similar, than an interpretation result where a bus is recognized as a building.

Références

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 (PAMI)*, vol. 25(10), pp. 1337–1342, 2003.
2. T. Deselaers, D. Keysers and H. Ney, "Improving a discriminative approach to object recognition using image patches," *Lecture Notes In Computer Science (LNCS)*, vol. 3663, pp. 326–333, 2005.
3. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection, " in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.
4. 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.pascalnetwork.org/challenges/VOC/voc2008/workshop/index.html.
5. E. D'Angelo, S. Herbin, and M. Ratiéville, *Robin challenge evaluation principles and metrics*, Nov. 2006, http://robin.inrialpes.fr.
6. H. Muller, W. Muller, D.M. Squire, S. Marchand- Maillet and T.Pun, "Performance evaluation in content-based image retrieval: Overview and proposals," *Pattern Recognition Letters*, vol. 22(5), pp. 593–601, 2001.
7. N.A. Thacker, A.F. Clark, J.L. Barron, J. Ross Beveridge, P. Courtney, W.R. Crum, V. Ramesh and C. Clark, "Performance characterization in computer vision: A guide to best practices," *Computer Vision and Image Understanding*, vol. 109(3), pp. 305–334, 2008.

8. P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features, ” in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 511–518.
9. C. Wren, A. Azarbayejani, T. Darrell and A. Pentland, “Pfinder : Real-time tracking of the human body,” *Transaction on Pattern Analysis and Machine Intelligence*, 1997.
10. Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent and C. Rosenberger, “Review and Evaluation of Commonly-Implemented Background Subtraction Algorithms, ” in *Proc. International Conference on Pattern Recognition (ICPR)*, 2008.
11. B.D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision, ” in *Proc. International joint conference on artificial intelligence*, 1981, pp. 674–679.
12. J.Y. Bouguet, *Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm*. Technical report, Intel Corporation, Microprocessor Research Labs, 1999.
13. Y. Benezeth, B. Emile, H. Laurent and C. Rosenberger, “Vision-based system for human detection and tracking in indoor environment,” *Special Issue on People Detection and Tracking of the International Journal of Social Robotics (IJSR)*, 2009.
14. B. Hemery, H. Laurent and C. Rosenberger, “Evaluation metric for image understanding,” in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2009.
15. I. T. Phillips and A. K. Chhabra, “Empirical performance evaluation of graphics recognition systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 21(9), pp. 849–870, 1999.
16. 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 (IJDAR)*, vol. 8(4), pp. 280–296, 2006.
17. 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,” in *Proc. International Conference on Computer Vision (ICCV)*, 2001, pp. 416–423.
18. B. Hemery, H.Laurent, C. Rosenberger and B. Emile, “Evaluation Protocol for Localization Metrics - Application to a Comparative Study,” in *Proc. International Conference on Image and Signal Processing (ICISP)*, 2008.
19. P. David, V. Idasiak and F. Kratz, “A Sensor Placement Approach for the Monitoring of Indoor Scenes, ” in *Proc. European Conference on Smart Sensing and Context (EuroSSC)*, LNCS, Vol. 4793, 2007, pp. 110–125.
20. B. Hemery, H. Laurent and C. Rosenberger, “Subjective Evaluation of Image Understanding Results,” in *Proc. European Signal Processing Conference (EUSIPCO)*, 2010.