Online Human-Assisted Learning using Random Ferns

M. Villamizar, A. Garrell, A. Sanfeliu and F. Moreno-Noguer
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
{*mvillami,agarrell,sanfeliu,fmoreno*}*@iri.upc.edu*

Abstract

We present an Online Random Ferns (ORFs) classifier that progressively learns and builds enhanced models of object appearances. During the learning process, we allow the human intervention to assist the classifier and discard false positive training samples. The amount of human intervention is minimized and integrated within the online learning, such that in a few seconds, complex object appearances can be learned.

After the assisted learning stage, the classifier is able to detect the object under severe changing conditions. The system runs at a few frames per second, and has been validated for face and object detection tasks on a mobile robot platform. We show that with minimal human assistance we are able to build a detector robust to viewpoint changes, partial occlusions, varying lighting and cluttered backgrounds.

1 Introduction

The standard approach in object recognition is to build a classifier offline using large amounts of training data, and then use this classifier at run-time. This approach has shown impressive results in a wide variety of challenging scenarios corrupted by noisy backgrounds, occlusions, viewpoint and scale changes and variations of object appearances [3, 9, 12, 14, 15, 13].

However, there are situations in which offline learning is not feasible, either because the training data is obtained continuously, or because the size of the training set is very cumbersome, and a batch processing becomes impractical. In these cases, novel online learning methods that use their own predictions to train and update a classifier have been proposed [6, 7, 11, 10]. Yet, although these approaches have shown great adaptation capabilities, they are prone to suffer from drifting when updating the classifier with wrong predictions. This has been recently addressed by combining offline and online strategies [4, 8].

Work supported by the Spanish Ministry of Science and Innovation under projects RobTaskCoop(DPI2010-17112), PAU+(DPI2011-27510) and MIPRCV(Consolider-Ingenio 2010)(CSD2007-00018), and the EU ARCAS Project FP7-ICT-2011-287617.

Figure 1. The TIBI robot interacting with people.

In this paper, we advocate for a completely online approach, in which the human assistance will be integrated within the learning loop in an active and efficient manner. For this purpose, we will take advantage of recent human-computer/robot interaction strategies, that have been shown effective for tasks such as people guidance [5] or robot teaching [1].

At the core of our approach there is an Online Random Fern classifier [8], which can be progressively learned using its own hypotheses as new training samples. Yet, to avoid feeding the classifier with false positive samples, the robot will ask for the human assistance when dealing with uncertain hypotheses (Fig. 1). The main issue to resolve will be to minimize the amount of human intervention in order to make the whole learning process as efficient as possible. This will be handled by appropriately defining the range of confidence scores for which the human is required. As we will show in the results section, the resulting *online human assisted* classier significantly improves a completely offline Random Ferns [12], both in terms of recognition rate and number of false positives. Fig 2 shows a few sample frames of the detection results, once the classifier learning is saturated (i.e., when no further human intervention is required). We are able to handle large occlusions, scalings and rotations, at about 5 fps.

2 Building the Human-Assisted Classifier

We next describe the ingredients of our approach: (1) the human-robot interaction; (2) the classifier; and (3) the criterion to demand for the human intervention.

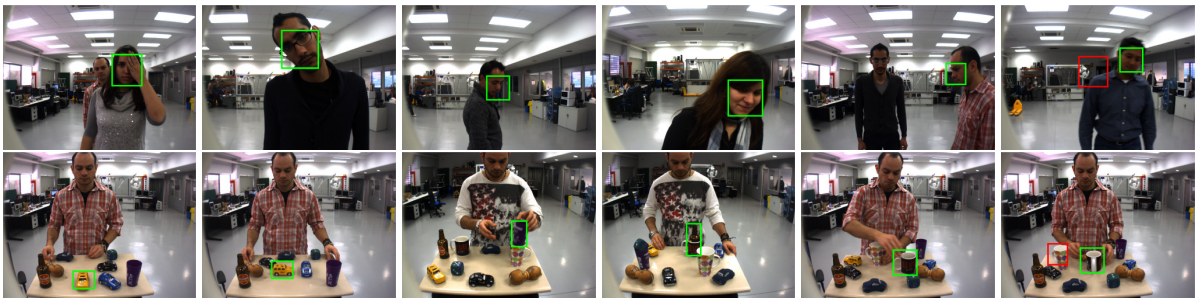

Figure 2. **Detection examples showing the output of the human-assisted classifier for face and object recognition. Green rectangles indicate correct detections and red ones are false positives.**

2.1 Human-Robot Interaction

We consider a scenario in which the classifier is learned using a computer onboard a mobile robot, equipped with devices such as a keyboard and a screen that enable the interaction with the human. In order to efficiently perform the interaction with the person, the robot will formulate a set of concise questions, that expect for a 'yes' or 'not' answer. Table 1 shows a few examples of such questions.

This human-robot interaction can be performed in a very dynamic and efficient manner. In addition, the robot has been programmed with behaviors that avoid having large latency times, specially when the human does not know exactly how to proceed. Strategies for approaching the person in a safe and social manner, or attracting people's attention have been designed for this purpose [2, 17].

2.2 The Online Classifier

The main difference between our ORFs classifier and the Random Ferns (RFs) detector of Ozuysal et al. [12], is that we learn and update the classifier on the fly. Our classifier is, therefore, continuously refined and adapted to the changing conditions of both the target and background. We next briefly describe the RFs, and the extension we do to train them online.

Random Ferns. RFs are a semi-naïve classifier which has demonstrated successful recognition rates and high efficiency for keypoint matching. Recently, they have also been applied to object tracking and categorization tasks [8, 15, 16].

Random Ferns consist of random and simple binary features computed from pixel intensities [12]. More formally, each Fern F_t is a set of m binary features $\{f_1^t, f_2^t, \ldots, f_m^t\}$, whose outputs are Boolean values comparing two pixel intensities over an image I. Each feature can be expressed as:

$$f(x) = \begin{cases} 1 & I(\mathbf{x_a}) > I(\mathbf{x_b}) \\ 0 & I(\mathbf{x_a}) \leq I(\mathbf{x_b}) \end{cases}, \qquad (1)$$

where $\mathbf{x_a}$ and $\mathbf{x_b}$ are the pixel coordinates. These coordinates are defined at random during the learning stage.

Type of utter	Example
Assistance	Is your face inside the rectangle? I'm not sure if I see you, am I?
No detection	I can't see you, move a little bit. Can you stand in front of me?

Table 1. **Sample phrases uttered by the robot.**

Fig. 3(Left) shows two Ferns, each one having three binary features –red, green and blue pairs of points–. The Fern output is represented by the combination of their Boolean feature outputs. For instance, the output z_t of a Fern F_t made of $m = 3$ features, with outputs $\{0, 1, 0\}$, is $(010)_2 = 2$.

Online Random Ferns. ORFs are Random Ferns which are continuously updated and refined using their own detection hypotheses. Initially, the parameters of the classifier are set using the first frame in which a bounding box around the target has been manually selected by the human, using the keyboard, mouse or touchscreen. Several random affine deformations are applied to this training sample in order to enlarge the initial training set, and initialize the RFs. In addition, we increase efficiency (both for the training and detection stages) by sharing RFs, as proposed in [16].

As shown in Fig. 3(Center), during the online training, the number of positive p_z and negative n_z samples falling within each output of each Fern is accumulated. Then, given a sample bounding box centered at x and a Fern F_t, we approximate the probability that x belongs to the positive class by $P(F_t = z|x) = p_z/(p_z + n_z)$, where z is the Fern output [8]. The average of all Fern probabilities gives the response of the online classifier:

$$H(x) = \frac{1}{k} \sum_{t=1}^{k} P(F_t|x), \qquad (2)$$

where $\frac{1}{k}$ is a normalization factor. If the classifier confidence $H(x)$ is above 0.5, the sample x will be assigned to the positive class. Otherwise, it will be assigned to the negative class.

To continuously update the classifier, upon the arrival of a new image we perform a bootstrapping step. That is, we run the classifier within the image and retain

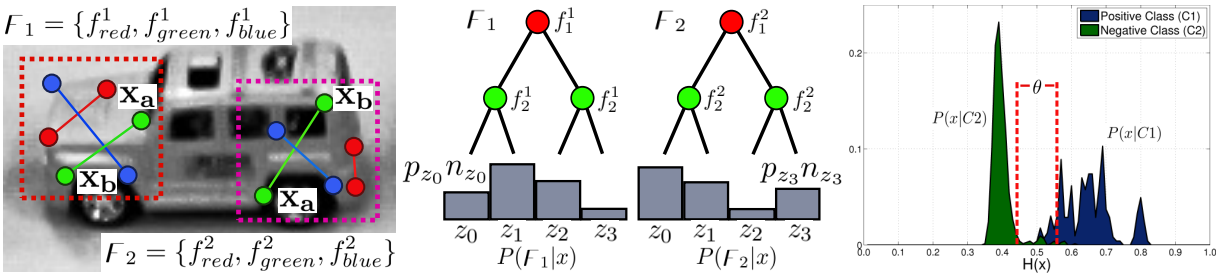

Figure 3. ***Left:*** **Randon Ferns (RFs).** ***Center:*** **Ferns probabilities.** ***Right:*** **Human-assistance criterion.**

the bounding box x_i with maximal confidence (only if $H(x_i) > 0.5$). This bounding box, and other nearby hypotheses are considered as new positive samples, while hypotheses which are far away are considered as new false positive samples. These positive and false positive samples are then evaluated for all the Ferns to update the aforementioned p_z and n_z parameters.

In terms of the specific parameters that define our classifier, we would like to point that in all the experiments described in the following section, the ORFs have been initialized using 100 positive and negative (background) samples. Each classifier uses 10 Random Ferns computed at $k = 200$ random image locations. The number of binary features has been set to $m = 9$ and the fern size is set to 12×12 pixels. By default, RFs are computed on gray-level images.

2.3 Human Assistance

ORFs are continuously updated using their own detection hypotheses. However, in difficult situations in which the classifier is not confident about its response, the human assistance will be required. The degree of confidence is determined by the response $H(x)$. Ideally, if $H(x) > 0.5$ the sample should be classified as a positive. Yet, as shown in Fig. 3(Right), we define a range of values θ (centered on $H(x) = 0.5$) for which we are not truly confident about the classifier response. Note that the width of θ represents a trade off between the frequency of required human interventions, and the recognition rates. A concise evaluation of this parameter will be performed in the experimental section.

3 Experimental Validation

We have evaluated our classifier on face and object datasets acquired using a mobile robot platform. The face dataset has 12 sequences of 6 different persons (2 sequences per person). Each face classifier is trained using one image sequence and tested in the second one. Similarly, the object dataset has 8 image sequences of 4 objects: a yellow toy car, an elvis mug, a beer bottle and a purple vase. Each sequence has approximately 200 images. Note (from Fig. 2), that these datasets are quite challenging as faces and objects appear under partial occlusions, 3D rotations and at different scales. In addition, fast motions and similar objects disturb the recognition method.

Method	θ	PR-EER	Assistance
RFs	–	55.81	–
ORFs	–	74.79	–
A-ORFs	0.05	76.31	$4.66\% \pm 0.46$
A-ORFs	0.1	76.51	$9.54\% \pm 0.87$
A-ORFs	0.2	79.44	$16.25\% \pm 1.09$
A-ORFs	0.3	82.06	$25.72\% \pm 1.65$

Table 2. **Face recognition rates.**

Face Recognition. In order to validate our approach we have used three different strategies for building the classifier. First, we considered an offline Random Ferns approach (RFs) which is learned using just the first frame of the training sequence and is not updated. The second approach considers an ORFs methodology without human intervention. Finally, our assisted approach which we denote by A-ORFs. Remind that the human assistance is only required during the training stage. During the test, all classifiers remain constant, with no further updating or assistance.

Fig. 4(Left) shows the Precision-Recall curves of the three methodologies, and Table 2 depicts the Equal Error Rates (EER). Both graphs show that the A-ORFs consistently outperform the other two approaches. This was in fact expected, as the A-ORFs significantly reduce the risk of drifting, for which both the RFs and ORFs are very sensitive, especially when dealing with large variations of the training sequence.

What is remarkable about our approach is that its higher performance can be achieved with very little human effort. This is shown both in the last 4 rows of Table 2 and in Fig. 4(Center), where we illustrate how the amount of human assistance influences the detection rates. Observe that with just assisting in a 4% of the training frames, the detection rate with respect to ORFs increases a 2%. This improvement grows to an 8% when the human assists on a 25% of the frames.

Object Recognition. The previous strategies have also been used to learn the appearance of specific objects. The recognition rates are plotted in Fig. 4(Right), and show similar patterns as those obtained for the face detection, in which the A-ORFs clearly outperformed RFs

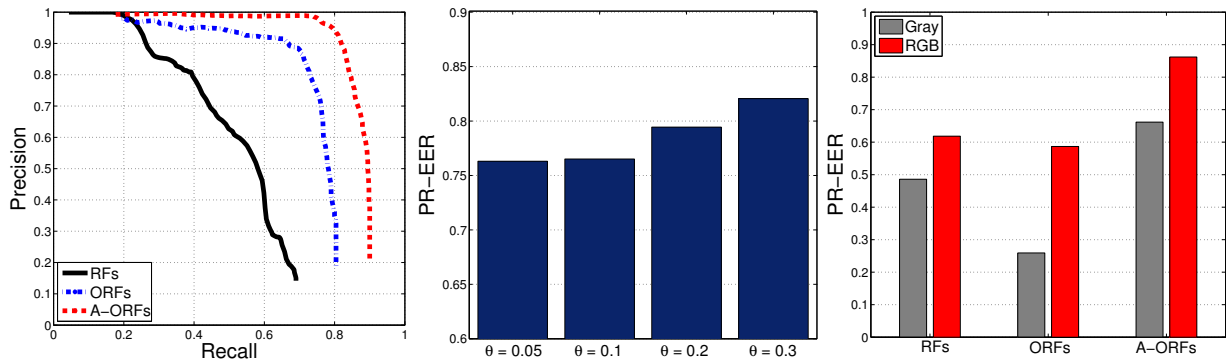

Figure 4. Face and object recognition results. ***Left:*** **Precision-Recall curves for different detection approaches.** ***Center:*** **Recognition rates in terms of human assistance.** ***Right:*** **Equal Error Rates using different color spaces.**

and ORFs. Surprisingly, the ORFs perform worse than the RFs. This is because the ORFs suffered from large drifting errors which corrupted the classifier.

In addition, since the Random Ferns can be extended to learn input images with multiple channels, we have evaluated the impact of using gray scale or RGB images. As shown in Fig. 4(Right), using the full color information consistently improves the detection rates, and allows discriminating between similar objects which appear in the scene.

Finally, a few frames of the detection results are shown in Fig. 2. These results were obtained at 5 fps. Yet, we could easily speed up our algorithm by applying simple temporal consistency criteria limiting the region of the image in which to search for object candidates.

4 Conclusions

We have presented an approach in which an online learning algorithm is assisted by a human to build a robust classifier. The whole process is performed very efficiently and with minimal human effort. We have used our approach to teach a mobile robot platform to detect faces and specific objects. The results show that our classifier is very discriminative under challenging scenarios, such as when the objects appear under partial occlusions or large changes of appearance. In the future, we will explore additional applications of our system. These include training with very large databases, and efficiently annotating complex databases.

References

[1] A. Andreopoulos, S. Hasler, H. Wersing, H. Janssen, J. Tsotsos, and E. Korner. Active 3d object localization using a humanoid robot. *in Transactions on Robotics*, 2011.

[2] D. Feil-Seifer and M. Mataric. Defining socially assistive robotics. *in ICRA*, 2005.

[3] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. *in CVPR*, 2010.

[4] J. Gall, N. Razavi, and L. V. Gool. On-line adaption of class-specific codebooks for instance tracking. *in BMVC*, 2010.

[5] A. Garrell, A. Sanfeliu, and F. Moreno-Noguer. Discrete time motion model for guiding people in urban areas using multiple robots. *in IROS*, 2009.

[6] M. Godec, C. Leistner, A. Saffari, and H. Bischof. On-line random naive bayes for tracking. *in ICPR*, 2010.

[7] H. Grabner and H. Bischof. On-line boosting and vision. *in CVPR*, 2006.

[8] Z. Kalal, J. Matas, and K. Mikolajczyk. P–n learning: Bootstrapping binary classifiers by structural constraints. *in CVPR*, 2010.

[9] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. *in CVPR*, 2008.

[10] F. Moreno-Noguer, A. Sanfeliu, and D. Samaras. Dependent multiple cue integration for robust tracking. *in PAMI*, 2008.

[11] F. Moreno-Noguer, A. Sanfeliu, and D. Samaras. Integration of conditionally dependent object features for robust figure/background segmentation. *in ICCV*, 2005.

[12] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *in PAMI*, 2010.

[13] M. Villamizar, J. Andrade-Cetto, A. Sanfeliu, and F. Moreno-Noguer. Bootstrapping boosted random ferns for discriminative and efficient object classification. *in Pattern Recognition*, 2012.

[14] M. Villamizar, H. Grabner, J. Andrade-Cetto, A. Sanfeliu, L. V. Gool, and F. Moreno-Noguer. Efficient 3d object detection using multiple pose-specific classifiers. *in BMVC*, 2011.

[15] M. Villamizar, F. Moreno-Noguer, J. Andrade-Cetto, and A. Sanfeliu. Efficient rotation invariant object detection using boosted random ferns. *in CVPR*, 2010.

[16] M. Villamizar, F. Moreno-Noguer, J. Andrade-Cetto, and A. Sanfeliu. Shared random ferns for efficient detection of multiple categories. *in ICPR*, 2010.

[17] D. Wilkes, R. Pack, A. Alford, and K. Kawamura. Hudl, a design philosophy for socially intelligent service robots. *in AAAI*, 1997.