Real Time People Detection Combining Appearance and Depth Image Spaces Using Boosted Random Ferns

Victor Vaquero, Michael Villamizar and Alberto Sanfeliu

Abstract This paper presents a robust and real-time method for people detection in urban and crowed environments. Unlike other conventional methods which either focus on single features or compute multiple and independent classifiers specialized in a particular feature space, the proposed approach creates a synergic combination of appearance and depth cues in a unique classifier. The core of our method is a Boosted Random Ferns classifier that selects automatically the most discriminative local binary features for both the appearance and depth image spaces. Based on this classifier, a fast and robust people detector which maintains high detection rates in spite of environmental changes is created.

The proposed method has been validated in a challenging RGB-D database of people in urban scenarios and has shown that outperforms state-of-the-art approaches in spite of the difficult environment conditions. As a result, this method is of special interest for real-time robotic applications where people detection is a key matter, such as human-robot interaction or safe navigation of mobile robots for example.

Keywords People detection · RGBD · Learning · Boosted Random Ferns

1 Introduction

From social robotics aiming to help people in different ways, to autonomous vehicles that needs to detect pedestrians and obstacles in order to avoid them and provide a safe navigation, robots these days are designed to share spaces with humans. Hardware and Software have evolved rapidly incorporating better sensors and algorithms on perception systems. However, robust algorithms are required for robots

V. Vaquero(✉) · M. Villamizar · A. Sanfeliu
Institut de Robotica i Informatica Industrial - CSIC-UPC, Barcelona, Spain
e-mail: {vvaquero,mvillami,sanfeliu}@iri.upc.edu
http://www.iri.upc.edu

L.P. Reis et al. (eds.), *Robot 2015: Second Iberian Robotics Conference*,
Advances in Intelligent Systems and Computing 418,
DOI: 10.1007/978-3-319-27149-1_45

coexisting with humans in populated environments and people detection algorithms are therefore fundamental.

Typical computer vision approaches on people detection use monocular vision systems and analyse individual appearance (RGB) images looking for a set of pre-learned features, as in [3]. However, other imaging spaces exist, as for example Histograms of Oriented Gradients (HOG) [1]. It has been proved that image HOG space is more robust to illumination and object appearance changes, obtaining substantial gains over features based on the appearance RGB domain. For a deeply analysis of the state-of-the-art in vison-based pedestrian detectors, the reader is referred to recent surveys and benchmarks, i.e [2], [6], [11], [5], [7].

We here present a real-time, robust and reliable method for detecting people in RGB-D images (Figure 6). This is done by computing a classifier that is able to learn how to combine in an effective way cues from both the RGB and Depth image spaces. In our approach we use Random Ferns (RFs), to compute simple and fast Local Binary Features (LBFs). Other works as [15], make use of RFs over the image appearance space (RGB intensities or HOG computed from greyscale images). However, our novelty resides in creating a synergistic combination of RGB and Depth spaces that allows the classifier to keep detecting people when one space is badly degraded, as it will be backed up by the other.

To summarize, the main contributions of the proposed approach are:

- A robust people detector able to learn by means of boosting algorithms the best combination of cues from different image spaces, creating a synergic environment where one space is still able to maintain high detection rates in case that others are spoiled. Our method outperforms the state of the art detectors achieving around an 89% of EER in the people database [9] which has challenging variability of poses, shapes and illumination.
- A fast and simple classifier that uses Random Ferns to compute and evaluate features over the defined spaces. This allows real-time applications because are based on Local Binary Comparisons and therefore does not require computationally expensive calculus. We have obtained performance results of 15-20 fps with a C++ version of the algorithm running in a standard computer (64 bits Intel Core-i7-3770 with 8Gb RAM, running Ubuntu 14.04). There is no need for using GPUs computation as in [13], [8].
- An open approach for a single classifier independently of the sources of information. Unlike other approaches, we create one unique classifier that seeks the best discriminative features combining any input spaces. This means that other imaging data (i.e, thermal information), could be easily used.

2 Related Work

This section presents the related work on people detection over RGB-D images as well as some other works where Random Ferns where used for similar purposes.

2.1 People Detection on RGB-D Images

Recent technologies that allows the easy capture of RGB-D images of a scene have upraised the limits of standard vision-based detectors. Affordable technologies such as stereo-vision systems (*Point Grey's - Bumblebee*), structured-light (*Microsof's - Kinect*) and time-of-flight cameras (*Asus' Xtion*), are nowadays being used for building new approaches on people detection.

In this way, Spinello *et al* used in [13] RGB-D data from a *Kinect* camera and trained two different classifiers separately, one for the RGB (appearance space) while the other for the Depth image. In a second step both classifiers were combined in their *Combo* classifier. A double effort is done in this work, firstly creating two separate classifiers, and secondly merging their outputs.

In contrast, we have created a single unique classifier devoted to both sources of information at the same time, which is able to learn and take the most discriminative features from each of the input image spaces (Appearance and Depth).

We have tested our method over the same database than [13] and [8] and under similar conditions, obtaining better and remarkable results with an average Equal Error Rate around 89%. Furthermore, as our Boosted Random Ferns performs simple binary comparisons, lower computational effort is required allowing real time performance without the need of GPUs.

Further works have also studied the way to leverage the RGB-D data, adapting the information that the classifier used from each of the sources depending on its arriving quality [4], [14]. On the contrary our Shared RFs and Boosting approach allows to developed a single strong classifier that learn and take the best discriminative RFs over each of the sources of information, which allows to keep detecting people when one of the input sources is distorted, as the other source is still able to produce good results.

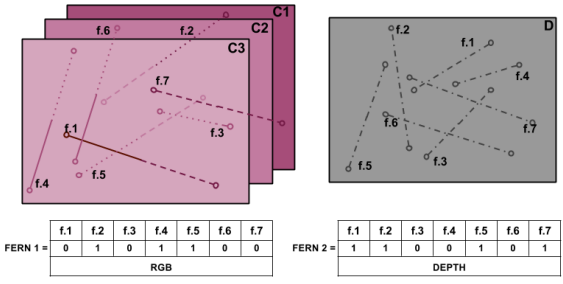

Fig. 1 Example of Ferns devoted to different spaces. Images presents three and one images spaces (i.e: RGB and Depth). Ferns compute Local Binary Features (LBFs, named as f_i) from random positions and channels over the image space creating weak classifiers as presented at the bottom of the image.

2.2 Boosted Random Ferns (BRFs)

Random Ferns (RFs), are presented in [10] and use hundreds of simple binary comparisons as features for modelling a class posterior probability. Extensions on this work, have been done in [17], where a *pool* of shared RFs was created obtaining a faster and efficient method for detecting even multiple classes. On top of that, in [16], a boosting algorithm was used for training the classifier with the best samples over the image domain.

In our method, we keep the pool of shared ferns, but we distinct ferns depending on the input space, being able to work in different spaces simultaneously. Ferns for RGB access to three different channels of information (Figure 1 - left), whereas others exist in only one dimension (Figure 1 - right). The boosting algorithm decides which ferns and where should be used to build the strongest possible classifier, as explained in Section 3. With this strategy, we keep the simplicity, speed and efficiency of the algorithm allowing its real time execution.

3 Developed Approach

3.1 Random Ferns on RGB-D Domains

In contrast to the original formulation of the Random Ferns of [10], proposed for keypoint classification, we write the Ferns expression in terms of likelihood ratios between classes. The aim is then to find by means of the boosting algorithm the combinations of features and its positions that maximize this ratio.

A posterior object class probability given a set of n Local Binary Features (LBFs) can be expressed by means of the Bayes rule as,

$$P(C|f_1, f_2, ..f_n) = \frac{P(f_1, f_2, ..f_n|C)P(C)}{P(f_1, f_2, ..f_n)}, \tag{1}$$

where C refers to class *People* (C^p) or *Background* (C^b), and f_i are the LBFs computed by the Ferns as shown in Figure 1. Therefore, we aim to maximize the posterior probability ratio of class *People*, with respect to the *Background* class.

Prior probabilities, $P(f_1, f_2, ..f_n)$, are common for all the classes, so it can be removed. Moreover, assuming uniform prior probabilities for both classes, $P(C_p) = P(C_b)$, the posterior probability is written by the likelihood ratio as,

$$\log \frac{P(C^p|f_1, f_2, ..f_n)}{P(C^b|f_1, f_2, ..f_n)} = \log \frac{P(f_1, f_2, ..f_n|C^p)}{P(f_1, f_2, ..f_n|C^b)} \tag{2}$$

Computing the complete joint probability for a large feature set is not feasible. A solution is to split the previous equation into m subsets ($F_i = \{f_1, f_2, ..f_r\}$), with $r = n/m$. These feature subsets will be our Ferns, and assuming they are

independent, their joint log-probability can be computed as,

$$\log \frac{\prod_{i=1}^{m} P(F_i|C^p, g_i, d_i)}{\prod_{i=1}^{m} P(F_i|C^b, g_i, d_i)} = \sum_{i=1}^{m} \log \frac{P(F_i|C^p, g_i, d_i)}{P(F_i|C^b, g_i, d_i)} , \quad (3)$$

where the parameter g_i ($g \in \mathbb{R}^2$) corresponds to the image coordinates location where the Fern F_i is evaluated, and d_i belongs to any of the image spaces evaluated (RGB, HOG, Depth or HOD, as will be explained in Section 4.1).

In this way, each Fern captures the co-occurrence of r binary features computed locally on the working spaces of the image, and therefore encodes people local features. Its response is represented by a combination of boolean outputs as seen at the bottom of Figure 1, where for instance the Fern F_1 (which applies to the RGB space - left of the Figure) is made of $r = 7$ features, results in 0, 1, 0, 1, 1, 0, 0, that outputs a value of $(0101100)_2 = 44$.

The Fern probability may then be written using the class conditional probability (for people, C^p and for background C^b), the Fern location g, the image space where the Fern works d, and the feature set of observations z_i as:

$$\sum_{i=1}^{m} \log \frac{P(F_i|C^p, g_i, d_i, z_i = k)}{P(F_i|C^b, g_i, z_i = k)} , \quad k = 1, 2, ...K , \quad (4)$$

with k, the observation index.

3.2 Combining RGB-D Spaces in a Single People Classifier

As have been seen, a weak classifier is created when any of the Shared RFs of our initial pool is taken and computed over its corresponding image domain at a certain position, having a score which corresponds to the Fern observation. Our aim is then to learn which are the most discriminative weak classifiers and build a single and robust classifier, which is done by means of boosting. Algorithm 1 summarizes the following described steps in order to build our final classifier.

More formally, we want to build a single people classifier $H(x)$, using the most discriminative Shared RFs F_i from our pool while maximizing Eq. 4. In this way, a Real Adaboost algorithm [12] is used for learning the best combination of these RFs over the image space locations g_i by iteratively assembling weak classifiers and adapting their weighting values.

The final people classifier is therefore defined as a sum of T weak classifiers,

$$E(x) = \sum_{t=1}^{T} h_t(x) > \beta_e , \quad (5)$$

where β_e is a threshold with a zero default value and $h_t(x)$ is the value of weak classifier, defined by

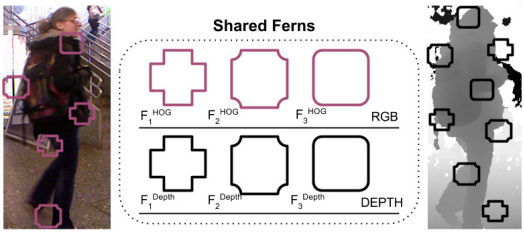

Fig. 2 Example of a pool of Shared Random Ferns (center) applied to an image of the database. In this case, three ferns are devoted to RGB space and other 3 to depth. The boosting algorithm applies the RFs over different positions generating weak classifiers, and choose the most discriminative ones. Here, five weak classifiers are chosen over the RGB space and eight over depth.

$$h_t(x) = \frac{1}{2} \log \frac{P(F_t|C^p, g_t, d_t, z_t = k) + \epsilon}{P(F_t|C^b, g_t, d_t, z_t = k) + \epsilon} , \quad k = 1, .., K , \tag{6}$$

being ϵ a smoothing factor.

Figure 2 shows a naive example of our people classifier, in which a pool of six shared random Ferns is created for a $d = 2$ image space, three RFs for the RGB space (d_1) and other three for Depth one (d_2).

At each iteration t of the boosting step, the classifier tries to find the most discriminative weak classifiers according to a sample weight distribution $D(i)$ by calling a weak learner. At iteration t, the probability $P(F_t|C^p, g_t, d_t, z_t)$ is computed under the $D(i)$ distributions as,

$$P(F_t|C^p, g_t, d_t, z_t = k) = \sum_{i:z_t(x_i)=k} D_t(i) , \quad k = 1, .., K \tag{7}$$

The classification power of each weak classifier is measured by means of the Bhattachryya distance between people and background distributions as,

$$Q_t = 2 \sum_{k=1}^{K} \sqrt{P(F_t|C^p, g_t, d_t, z_t = k) P(F_t|C^b, g_t, d_t, z_t = k)} \tag{8}$$

Figure 3 shows a density map of weak classifiers applied over the HOG and Depth image spaces, resulting after identifying in the boosting step, which areas of each space are the most discriminative in the training classifier set.

The final classifier will therefore be composed by the combination of the selected weak classifiers from the image spaces. Figure 4 shows an example of a resulting distribution from a pool of twelve shared RFs and 600 weak classifiers. In this example, around 350 weak classifiers are applied over the HOG space, whereas 250 are in the Depth space. Although the Depth space have less weak classifiers, its distribution is more focused on certain areas (as seen in Figure 3), so it still manages to have a high accuracy so that being able to correctly driving the detector even if the HOG space is corrupted at certain point.

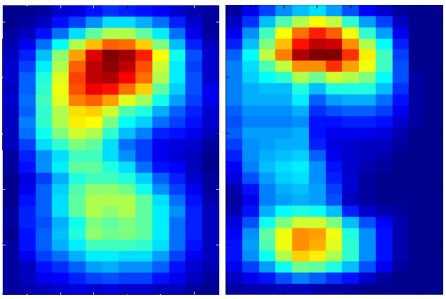

Fig. 3 Resulting weak classifiers density map over the HOG space (left) and the Depth space (right) of a real database image. The boosting algorithm find the best position to apply the Random Ferns from the pool which would produce the most discriminative results. The weak classifiers are mainly focused in the head area and uniformly distributed over the rest of the body for the HOG space whereas for the Depth space main discriminative areas are the head and feet zones. High density areas are represented in red colors.

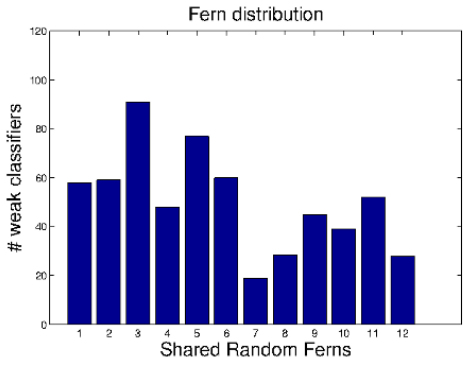

Fig. 4 Example of the resulting distribution of the most discriminative selected Ferns. In this case, a pool of 12 Shared Random Ferns where created, being the 6 first devoted to the HOG domain, whereas the rest to the Depth space. As it can be seen, the boosting algorithm has given more weight to the HOG image.

4 Experiments

In this Section, the experiments that validate our people detector will be presented. In order to show how our contributions outperforms other similar algorithms for people detection in RGB-D, we compare the results with the ones obtained in [13] using the same RGB-D database and under the same circumstances.

This public database for people detection [9] was collected indoor by three different *Kinect* sensors in the lobby of a university canteen. We have noticed some missing tracks of annotated people in the public website, and therefore no full groundtruth

Algorithm 1. Detector computation.

Given a number of weak classifiers T, and N RGB-D image samples labelled $(x_1, y_1)...(x_n, y_n)$, where $y_i \in \{+1, -1\}$ is the label for people category (C^p) and background classes (C^b), respectively:
Construct a shared feature pool ϑ consisting of M Random Ferns divided uniformly to be devoted to appearance or depth spaces $d_i, i \in 1, 2$.
Initialize sample weights $D_1(i) = \frac{1}{N}$.
for $t = 1$ to T **do**
  for $m = 1$ to M **do**
    Under current distribution D_t, calculate h_m and its Bhattachryya distance Q_m.
  end for
  Select the h_t that minimizes Q_m.
  Update sample weights.
  $D_{t+1}(i) = \frac{D_t(i) \exp[-y_i h_t(x_i)]}{\sum_{i=1}^{N} D_t(i) \exp[-y_i h_t(x_i)]}$
end for
Final strong classifier.
$H(x) = sign\left(\sum_{t=1}^{T} h_t(x) - \beta_e\right)$

information in available. This make the database a challenging one, but even under these circumstances our people detector outperforms previous cited works obtaining remarkable results of almost 89% of EER. In the fourth column of Figure 6, some of the missing annotations that our detector has positively find are shown.

The dataset in its actual version contains 3399 images (for each RGB and depth source) from which, up to 2351 are annotated providing 4498 2-D boxes of people. In total, 1748 images include full people, and from all the annotations, up to 2248 of the boxes corresponds to people that is fully visible, whereas the rest is considered as occluded.

In each implementation of our experiments, 1096 random crops from the 2248 of fully visible people annotations have been extracted to be used as positive training samples. The remaining 1152 annotated people on each implementation where used for testing purposes. Unlike [13] did using 5000 negative samples of background for training, we just also use 1096, in order to keep the equity between people and background classes on training.

Alike the authors did and aiming to reproduce the same circumstances of their experiments for a fair comparison, a *no-reward-no-penalty* policy was used as also introduced in [3]. Therefore, partially occluded annotated people do not count for true positives nor false positives. Also, in the same way a detection is considered as true positive if its intersections with the corresponding annotation is at least the 40%.

4.1 Approaches for Combining RGB-D Data

In our people detecton approach, we have mainly work with four different image spaces. Apart from standard RGB and Depth spaces, Histograms of Oriented Gradients (HOG) and Histograms of Oriented Depths (HOD) has been used for creating

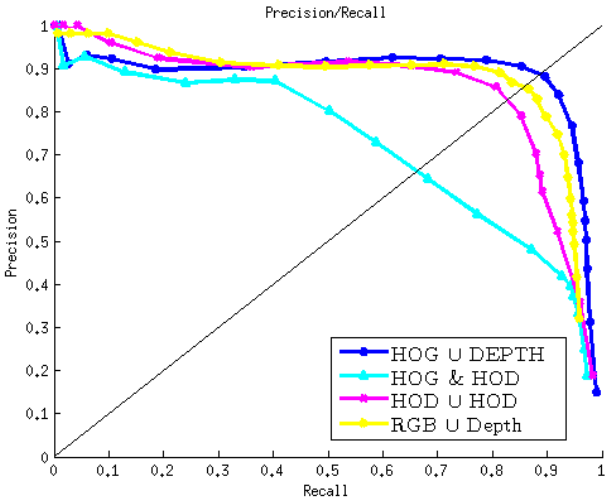

Fig. 5 Detection performance evaluation in terms of the Precision-Recall plots for different depth- and appearance-based feature configurations.

the combined classifier. HOG image domain was introduced in [1] and its features showed to obtain substantial gains over standard RGB features, even though it is calculated over the greyscale transformed image. As the Depth of a scene can be observed as a kind of a greyscale image, the same process of computing HOG can be extrapolated to it, obtaining in this way the areas where there exists variations in depth. In [13], authors called this technique Histogram of Oriented Depths (HOD), so in order to keep notation simply we have adopted the same name.

At this point, to validate our people detector method for combining Appearance and Depth information in any of the described spaces, we have implemented 4 different combination approaches. Each combination have been tested up to 10 times,

Table 1 Qualitative results on the performance of the different classifier combinations approaches tested. Columns present the Equal Error Value (EER), along with the recall and precision values achieved at this point. In addition the mean number of True Positives (tp), False Positives (fp) and False Negatives (fn) detections over the set of 1152 test images are included. However, some of these fp, are existing people correctly detected, but without background associated.

Method	EER Value	Recall	Precis	tp	fp	fn
HOG & HOD	0.6534	0.6516	0.6552	747.333	394.000	399.667
HOG ∪ HOD	0.8199	0.7995	0.8403	917.000	174.333	230.000
RGB ∪ DEPTH	0.8758	0.8767	0.873	1017.667	148.333	130.333
HOG ∪ DEPTH	0.8861	0.8881	0.8841	1018.667	133.667	128.333

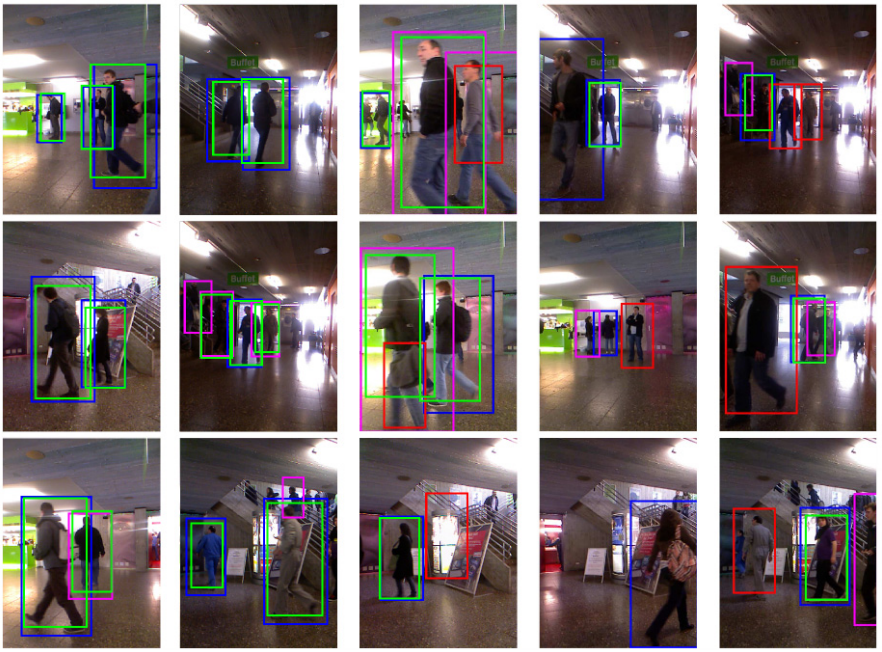

Fig. 6 Sample images showing the output of the proposed approach for people detection in urban and crowded scenarios. Full people annotations (bounding boxes) are indicated by blue boxes in images, whereas magenta rectangles correspond to occluded people. Correctly detections (true positives) provided by our classifier are represented in green, whereas false positives are shown by red boxes. Columns 1 and 2 depicts some true positive detections, while column 3 shows some false positives. Column 4 shows some examples of not detected people, as well as a *false* false positive in the middle (a correct detection but not annotated in the dataset). Finally, column 5 presents some other *false* false positives, which really penalises our detector. However, despite the fact of the presence of this *false* false positives, our experiments on people detection combining appearance and depth images spaces achieves very good results around a 89% of EER, which surpass other methods.

calculating afterwards the mean of values. In each experiment, a shared pool of 12 RFs was created (6 for each of the 2 dimensions), with 9 LBFs each. These values were chosen after previous experiments showed to be the ones with best computational efficiency vs robustness ratio. Final qualitative results of the experiments are shown in Table 1, and Precision-Recall curves for the image spaces combinations can be consulted in Figure 5.

– HOG ∪ Depth
  This combination has obtained the best results for people detection over the database. The union of an already proved accurate image feature domain as HOG, along with the depth information of the scene, is very robust and even under the presence of *false* false positives due to the lack of annotations. In our experiments

it has achieved remarkable results outperforming other methods such as the one presented in [13], with an EER of almost 89%.

– HOG ∪ HOD
  In the same way as the authors of the used database did in [13], we have build a classifier based also in the union of both HOG and HOD image domains in order to do a full and fair comparison of the results. The results for this approach can been observed in magenta color at Figure 5, obtaining a EER of around the 82%, which is 7 points less than our best combination.
– HOG & HOD
  In this approach both the HOG and HOD information is fused, instead of combining it. For this, two options have been evaluated. In one hand we added the results of both histograms, which would favor the areas where both gradients exists. On the other hand the multiplication has been done, which would penalize more the areas where there are no gradients while at the same time would exalts the places where both domains have gradients. However, both of these approaches have thrown a really bad performance of around an 65% of EER when applied in the current database.
– RGB ∪ Depth
  A final combination of both RGB and depth raw information has been done. Results for this approach over the database can be observed in yellow in Figure 5. Quite good results are appreciated, of around an EER of 87%, and as here only raw data is used, less computational time is employed so better speeds of around 15 frames per second are obtained.

5 Conclusions

We have presented a robust, fast and accurate method for people detection in RGB-D data that obtains remarkable results compared to other classical people and pedestrian detectors. The presented approach is based on the combination of RGB and Depth image spaces, as well as its derived spaces (Histograms of Oriented Gradients and Depths) by means of extracting features using Shared Random Ferns. A learning approach making use of a boosting algorithm selects in the training phase the most discriminative weak classifiers between all the possible permutations of Random Ferns over the combined image domains. This combination of Random Ferns and boosting, allows to create a single classifier where its components act synergistically providing accurate detections even when one of the spaces are spoiled or distorted.

The proposed method has been validated in a recent and challenging RGB-D database of people with four different ways of combining the image spaces, and shows remarkable results of around an 89% of EER in spite of the difficult environment conditions.

Acknowledgements This work has been partially funded by the EU project CargoANTs FP7-SST-2013-605598 and the Spanish CICYT project DPI2013-42458-P.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. I, pp. 886–893 (2005)
2. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(4), 743–761 (2012)
3. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**, 2179–2195 (2009)
4. Enzweiler, M., Gavrila, D.M.: A multilevel mixture-of-experts framework for pedestrian classification. IEEE Transactions on Image Processing **20**, 2967–2979 (2011)
5. Gandhi, T., Trivedi, M.M.: Pedestrian protection systems: Issues, survey, and challenges. IEEE Transactions on Intelligent Transportation Systems **8**, 413–430 (2007)
6. Gerónimo, D., López, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**, 1239–1258 (2010)
7. Guo, L., Wang, R.B., Jin, L.S., Li, L.H., Yang, L.: Algorithm study for pedestrian detection based on monocular vision. In: 2006 IEEE International Conference on Vehicular Electronics and Safety, ICVES, pp. 83–87 (2006)
8. Luber, M., Spinello, L., Arras, K.O.: People tracking in RGB-D data with on-line boosted target models. In: IEEE International Conference on Intelligent Robots and Systems, pp. 3844–3849 (2011)
9. Luber, M., Spinello, L., Arras, K.O.: RGB-D People Dataset. annotated people and tracks in rgb-d kinect data (2011). http://www2.informatik.uni-freiburg.de/spinello/RGBD-dataset.html (accessed September 30, 2014)
10. Özuysal, M., Fua, P., Lepetit, V.: Fast keypoint recognition in ten lines of code. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
11. Porikli, F., Davis, L., Hussein, M.: A Comprehensive Evaluation Framework and a Comparative Study for Human Detectors. IEEE Transactions on Intelligent Transportation Systems **10**, 417–427 (2009)
12. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning **37**, 297–336 (1999)
13. Spinello, L., Arras, K.O.: People detection in RGB-D data. In: IEEE International Conference on Intelligent Robots and Systems, pp. 3838–3843 (2011)
14. Spinello, L., Arras, K.O.: Leveraging RGB-D data: adaptive fusion and domain adaptation for object detection. In: IEEE International Conference on Robotics & Automation, pp. 4469–4474, May 2012
15. Villamizar, M., Andrade-Cetto, J., Sanfeliu, A., Moreno-Noguer, F.: Bootstrapping Boosted Random Ferns for discriminative and efficient object classification. Pattern Recognition **45**(9), 3141–3153 (2012)
16. Villamizar, M., Moreno-Noguer, F., Andrade-Cetto, J., Sanfeliu, A.: Efficient rotation invariant object detection using boosted random ferns. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1038–1045 (2010)
17. Villamizar, M., Moreno-Noguer, F., Andrade-Cetto, J., Sanfeliu, A.: Shared random ferns for efficient detection of multiple categories. In: International Conference on Pattern Recognition (ICPR), pp. 2–5 (2010)