Evaluation of Boosting-SVM and SRM-SVM Cascade Classifiers in Laser and Vision-based Pedestrian Detection

Oswaldo Ludwig, Cristiano Premebida, Urbano Nunes, and Rui Araújo

Abstract— **Pedestrian detection systems constitute an important field of research and development in computer vision, specially when applied in protection/safety systems in urban scenarios due to their direct impact in the society, specifically in terms of traffic casualties. In order to face such challenge, this work exploits some developments on statistical machine learning theory, in particular structural risk minimization (SRM) in a cascade ensemble. Namely, the ensemble applies the principle of SRM on a set of linear support vector machines (SVM). The linear SVM complexity, in the Vapnik sense, is controlled by choosing the dimension of the feature space in each cascade stage. To support experimental analysis, a multi-sensor dataset constituted by data from a LIDAR, a monocular camera, an IMU, encoder and a DGPS is introduced in this paper. The dataset, named Laser and Image Pedestrian Detection (LIPD) dataset, was collected in an urban environment, at day light conditions, using an electrical vehicle driven at low speed. Labeled pedestrians and non-pedestrians samples are also available for benchmarking purpose. The cascade of SVMs, trained with image-based features (HOG and COV descriptors), is used to detect pedestrian evidences on regions of interest (ROI) generated by a LIDAR-based processing system. Finally, the paper presents experimental results comparing the performance of a Boosting-SVM cascade and the proposed SRM-SVM cascade classifiers, in terms of detection errors.**

I. Introduction

Protection systems for pedestrian safety, in urban environment, is an emerging scientific research area of Advanced Driver Assistance Systems (ADAS) which achieved a notable development in the last decade. It is a still growing research field, evidenced by recent projects, challenges [1], [2], and recent publications [3], [4]. For instance, in the last years, three significant surveys in pedestrian detection and protection systems, in the context of Intelligent Vehicles (IV) and Intelligent Transportation Systems (ITS), were published in [4], [5], [6]. The main reasons behind the interest in this scientific domain are basically:

- society concerns: it is an appealing topic of research due to its direct impact in the society, in terms of traffic casualties and the large economic and societal cost implied;
- industry involvement: there is a strong interest of the automotive industry, notorious by the continuous enhancements on safety features in the vehicles;
- research interest: international project consortia, international challenging competitions and awards, and the scientific community in several fields have demonstrated, and still do, large interest in innovations and developments associated with this topic.

Pedestrian protection systems can be divided, in general words, in two fields of research: passive and active safety systems [5]. Active safety systems, which is of interest here, are based on pedestrian detection using sensors onboard the vehicle, and/or on the infrastructure, with the role of predicting and anticipating possible risks of collision. In particular, active pedestrian detection systems using onboard LIDAR, or laserscanner, and monocular camera will be emphasized in this work. More specifically, this paper is focused on two cascade ensembles of SVMs designed to detect pedestrian evidences inside ROIs generated by a LIDAR-based processing module. The proposed cascades, involving a series of SVMs, perform direct negative rejection in each stage, with the purpose of reducing the number of negatives and the computational time in the subsequent stages, which is of particular importance in pedestrian detection since the number of negatives is much larger than the positives. The first cascade introduced in this paper (inspired in [7] and [8]), designated by Boosting-SVM cascade, is trained using a boosting process where the number of features, in a given stage, increases *wrt* to the preceding stage; thus, the complexity of the cascade and its classification capability increase as more stages are added to the structure. The second ensemble, named SRM-SVM cascade, follows a learning strategy which selects a suitable number of features for each ensemble stage, in such a way as to control its complexity, in order to minimize the structural risk of each classification stage.

The laser and image pedestrian detection (LIPD) dataset and its statistics are introduced in Section II. Our LIDAR-based ROI generation approach is presented in Section III. Section IV briefly presents some background material. The proposed cascades of SVMs are detailed in Section V, while experimental results in pedestrian detection using the LIPD dataset are presented in Section VI. Finally, Section VII presents some conclusions.

II. The LIPD dataset

The LIPD dataset contains, besides monocular images and LIDAR scans, data from two proprioceptive sensors, an IMU and an incremental encoder, in conjunction with DGPS and battery-bank state data (terminal voltage, current and temperature). The dataset was recorded using the sensor

This work has been supported by National Funds through Fundação para a Ciência e a Tecnologia de Portugal (FCT), under project grants PTDC/SEN-TRA/099413/2008 and PTDC/EEA-AUT/113818/2009. O.Ludwig is supported by FCT under grant SFRH/BD/44163/2008. The authors are with the Department of Electrical and Computer Engineering, Institute of Systems and Robotics, University of Coimbra, Portugal. e-mails:{oludwig,cpremebida,urbano,rui}@isr.uc.pt

TABLE I

LIPD DATASET: SENSORS, INTERFACES AND USED ACQUISITION FREQUENCY

Sensor	Manufacturer	Interface	Acquisition rate
LIDAR (Alasca-XT)	Ibeo	Ethernet	12.5Hz
Camera (Guppy)	Allied	FireWire	30fps
IMU	XSens	USB	120Hz
DGPS	TopCon	USB	5Hz
Encoder and batteries	-	USB	10Hz

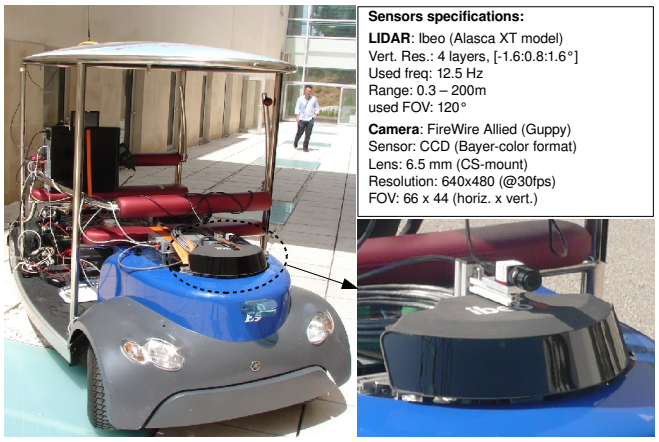


Fig. 1. ISRobotCar and vehicle-mounted sensor setup, enlarged at the bottom-right part, used in the dataset collection. A short specification of the sensors is presented at the top-right side of the figure.

system mounted on the ISRobotCar (autonomous electric vehicle with a chassis from Yamaha Europe and control systems developed in the Institute for Systems and Robotics of Coimbra University) as shown in Fig. 1. For that purpose, ISRobotCar was driven through areas of the engineering *campus* of the University of Coimbra and neighboring areas[1]. Table I outlines the sensors and their manufacturers, the data communication interface protocols, and the frequency of data acquisition used to record the dataset in a host PC.

Due to the fact that the dataset was obtained in outdoor conditions, and since the sensor apparatus has been exposed to weather and environmental conditions, not unexpectedly, some 'difficulties' have occurred namely: light exposure variations, vibrations, oscillations, noise, dust and particles on the air, among others. Perhaps one of the main problems during the data recording was the occurrence of some spots in the images due to dust on the lens.

The manual labeling process, inherent to any supervised dataset, was carried out using the image frames as primary reference for pedestrian and non-pedestrians annotation. The labeled segments, extracted from raw data laser-scans, were validated using the corresponding image frame (for *ground truth* confirmation). All the samples of interest were labeled under user supervision, avoiding some problems invariably presented on realistic situations, such as: data association mistakes, over-segmentation, missing measurements, calibration inaccuracies, road irregularities, tracking inconsistencies, vehicle vibrations, and so on. However, it was not possible to guarantee 100% of that correspondence, due to a series of reasons, namely: it is a human-based task and it is prone to mistakes; the calibration of the sensor set is not perfectly precise; the time synchronization is not perfect and neither with a precise constant interval.

In summary, the LIPD dataset comprises a training dataset ($\mathcal{D}_{train}$) which is composed exclusively of laser-segments and ROIs (see Fig. 2), representing positives and negatives. On the other hand, $\mathcal{D}_{test}$ comprises raw laser-scans and full image frames, necessary to evaluate the detection system under realistic conditions. $\mathcal{D}_{train}$ is used to train the classifier parameters and also to perform cross-validation, bagging and feature selection; and $\mathcal{D}_{test}$ is used to evaluate the performance of the techniques and methods learned using $\mathcal{D}_{train}$. The cardinality of the dataset is defined by the number of samples, and the dimensionality is defined by the number of features.

Definition II.1 *(rough definition): a positive sample is defined by an entire body pedestrian (PED) present in both the camera and laser field of view (FOV). A negative sample is defined by any other object (nPED) present in the FOV of both sensors, while an occluded pedestrian denotes a partial occluded PED.*

Definition II.2 *A ROI is defined in the image plane by the projection of a laser-segment. ROIs are defined considering the extremes of a segment rather than individual laser-points. Due to the Alasca XT laser sensing principle, the vertical component in the underlying data is strongly limited, therefore the top and bottom part of an object can not be estimated directly from the laser measurements. Under the assumption of flat surface, knowing the distance of the laser setup from the ground and considering the maximum height of the objects as 2.5m, the ROIs coordinates are calculated.*

Figure 2 illustrates laser-segments (clusters of range points: left part of the figure) and their projections onto the image plane, represented by dashed regions. From a total of more than 100K raw samples collected in different days during the Autumn season, at day-light conditions in sections of 4-5 hours, a more restricted amount of data was selected, remaining 26417 samples. Finally, 14367 images were used to compose the training dataset and 12050 the testing part. This selection process was carried out to reduce the cardinality of the dataset to a tractable value and, at the same time, to keep a representative dataset in terms of the universe that defines the problem of pedestrian detection in outdoor scenarios.

The training part of the dataset contains 5237 manually labeled positives (image's cutouts of pedestrian in up-right entire body), and 6328 full-frames, 640x480 resolution images, without any pedestrian evidence. Thus, the elements of the training dataset are the aforementioned 5327 positive bounding-boxes and a free-number of negative ROIs which

[1]http://www.isr.uc.pt/~cpremebida/PoloII-Google-map.pdf

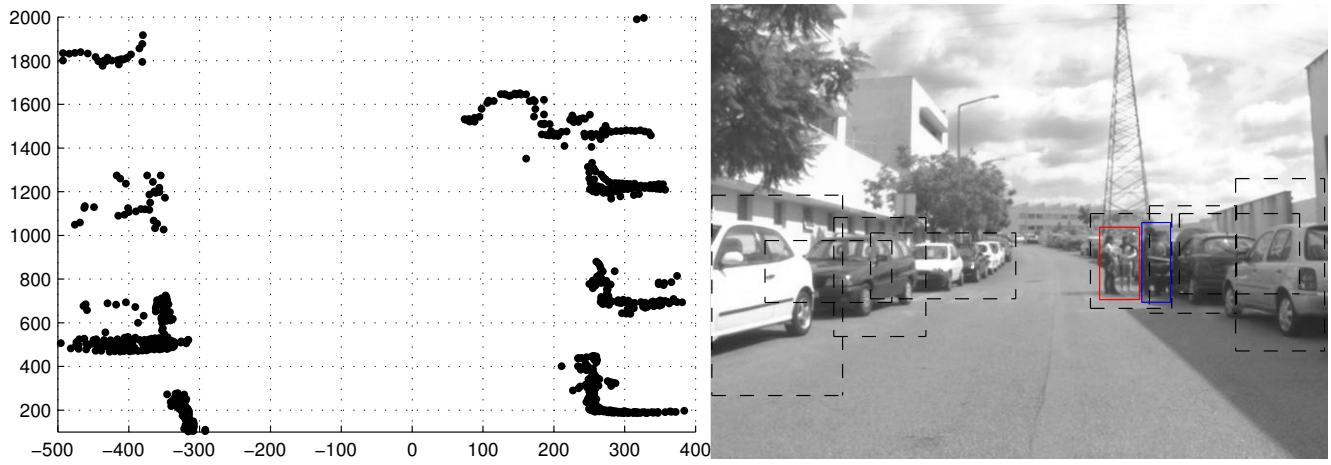


Fig. 2. Laser-based ROI projections in the image frame. Entire body pedestrians (class type 1) and partially occluded pedestrians (class type 0) are highlighted in the image by red and blue bounding-box respectively.

can be extracted from the negative frames. The current testing dataset contains 4823 full-frame images, which correspond to the frames with DGPS information extracted from the entire set (with 12050 samples). To be consistent with recent benchmarking datasets, detailed annotations regarding the pedestrian appearance (in terms of occlusion) were done, namely: occluded/partial pedestrians (class type 0) and entire body pedestrians (class type 1). A summary of the dataset is shown in Table II.

TABLE II
STATISTICS OF LIPD

Set	Npos	Nneg	Description
Training set			
$\mathcal{D}_{train}$	5237	6328	Sunny days, Autumn season. Negative ROIs should be extracted from the Nneg frames.
Testing set			
$\mathcal{D}_{test}$	593	2.76M	Sunny days, Autumn season. The number of negatives (Nneg) depends on the detection approach to be used (which can take advantage of the LIDAR information or not).

III. METHOD TO DEFINE ROIS IN THE IMAGE

A key problem in monocular image-based pedestrian detection, in the field of ADAS applications [4], [5], is the huge amount of negatives (potential false alarms) in contrast with the number of positives, what demands a vast processing time consumption and a high confidence detector. An usual solution adopted to avoid using brute-force multiscale sliding windows approach is to combine, in a sensor fusion architecture, the image-based sensors with active sensors like LIDAR (or laserscanners).

We propose to use a LIDAR-based system acting as a primary object detection sub-system, where each detected object (represented by a laser-segment) constitutes a hypothesis of being a positive (PED) or a negative (nPED). This system outputs a set of laser-segments that are transformed into image plane as regions of interest (ROIs - **Definition** II.2). The functional block diagram of the detection system is divided in LIDAR and image-based sub-systems, as shown in Fig. 3, where the main processing modules are described below (for details see [9]):

1) LIDAR-data processing: a module comprising a set of pertinent data processing tasks, necessary to decrease complexity and processing time of subsequent stages, such as: filtering-out isolated/spurious range-points, discarding measurements that occur out a predefined FOV, and data alignment.
2) LIDAR segmentation: this module outputs a set of segments obtained by a range-data segmentation method applied per LIDAR layer, and a subsequent segment association approach used to combine the set of segments extracted from the layers.
3) Coordinate transformation: defined as the set of rigid coordinate transformations, obtained by system calibration [10], necessary to project laser-segments into the image plane. This module outputs a set of ROIs.
4) Cascade detection: involves a cascade ensemble of SVMs, trained with HOG and COV descriptors, which is employed to detect, using window detectors, potential pedestrians inside the ROIs.

The number of window detectors, used to scan the ROIs in searching for pedestrian evidence, is limited and it is defined by the size of the ROI. These window detectors are shifted by horizontal and vertical step factors, and the window scale is estimated using the depth information provided by the ROIs, that is, LIDAR measurements are also used for scale estimation. This approach decreases the computational processing time, restricting the areas of interest in the image, at most, to a dozen ROIs, and keeping the false positives at low values. For instance, the number of window detectors generated by this LIDAR-based approach is, in average, thousands times lower than the usual full-scanning image approach.

Additionally, some experiments have been performed using the image-based focus of attention method proposed in

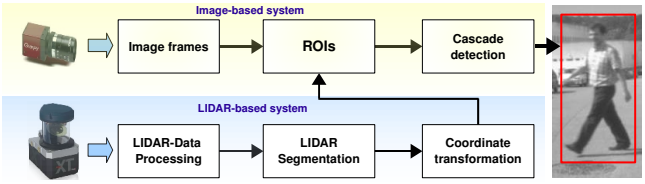


Fig. 3. Pedestrian detection composed of laser and vision subsystems.

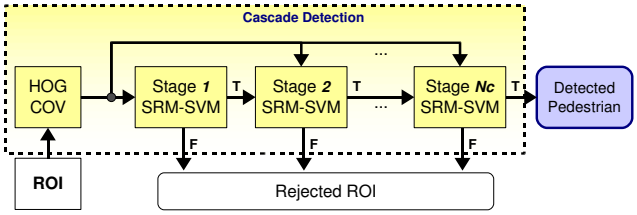


Fig. 4. Functional diagram illustrating the negative-rejection cascade detector: each stage *k* rejects the samples classified as negative (F: false), while those classified as positives (T: true) pass to the next stage k+1.

[11] as an attempt to restrict the system to salient regions, namely, to avoid regions in the image typically covered by the sky and the road and, at the same time, to compare this method with the proposed laser-based approach. That method can in fact reduce drastically the number of searching regions in the image but, as counterpart, it was verified that a significant number of positives (pedestrians) tends to be missed. Moreover, the scale can not be directly estimated. To summarize, the approach to define ROIs in the image has two major advantages: (1) decreasing the time-processing and complexity; (2) enabling range (scale) information.

IV. Background

Considering a training dataset S with l pairs $(x_1, y_1), \ldots, (x_l, y_l)$, where $x \in U$ represents the input vectors and y denotes the targets, which are considered to be drawn randomly and independently according to a unknown joint distribution $F(x, y) = F(y|x)\,F(x)$, we define the learning procedure as the process of choosing an appropriate function $f(x, \alpha^*)$ [12], in the sense of the adopted objective function of the training algorithm, from a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$ which can contain a finite number of elements (e.g., in the case of decision trees) or an infinite number of elements (e.g. syntactic classifiers, such as RNN, which have a set Λ of adjustable parameters that can assume any real value).

Taking into account the following loss function

$$L(f(x,\alpha), y) = \{^{0 \text{ if } f(x,\alpha)=y}_{1 \text{ if } f(x,\alpha)\neq y}, \tag{1}$$

the problem is to determine the probability that the expected risk $R(\alpha) = \int L(f(x,\alpha), y)\, dF(x,y)$ will deviate from the empirical risk $R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} L(f(x_i,\alpha), y_i)$. Such problem was approached by Vapnik and Chervonenkis which provided an upper-bound on the expected risk:

Theorem 1 *[12] Let h denote the VC-dimension of the set of functions $f(x,\alpha)$, $\alpha \in \Lambda$. For all α, all $l > h$, and all $\sigma > 0$, the inequality bounding the expected risk*

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\ln \frac{2l}{h} + 1) - \ln \frac{\sigma}{4}}{l}} \tag{2}$$

holds with probability of at least $1-\sigma$ over the random draw of the training samples.

V. Cascade ensembles

The cascade ensembles applied in this work, using SVM as component classifiers, perform rejection of negatives in series [7] (see Fig. 4), *i.e.*, input data classified by the current stage as negative occurrence is immediately rejected, however, input data classified as positive example is propagated through the cascade stages. The cascade ensembles are trained in a bootstrap fashion: the cascades have their first stage trained in the usual way (using the entire dataset), and the subsequent stages are trained by using a sub-set composed by the aggregation of the positive examples, which are the same for all the stages, and the false positives gathered from previous stages.

A. Boosting-SVM cascade

The Boosting-SVM cascade employs a simple training method where each stage of the cascade is trained in such way that a constant true positive rate (TP), e.g. $TP = 0.98$, is achieved by adjusting the bias (or threshold) of the SVM in the current stage. The number of features is incremented as the number of stages increases; that is, the number of features (and the complexity) of a given stage is increased by adding n_f features *wrt* the previous stage. The feature vector was previously ordered as function of the maximum relevancy and minimal redundancy (mRMR [13]), and the number of features n_f was selected to be constant and equal to 10; thus the first stage has 10 features, the second 20, and so on.

B. SRM-SVM cascade

The training method for the SRM-SVM cascade ensemble is summarized in **Algorithm 1**. Linear classifiers, in this case SVM, offer a good opportunity to apply structural risk minimization schemes, because their Vapnik-Chervonenkis dimension (VCd) can be conveniently determined, what enables the perfect control of the classifier complexity and, consequently, a better control of the upper bound on the expected risk, namely in case of usual linear classifiers, $VCd = N+1$, where N is the number of features. However, because this work applies linear SVMs, this value will be used as an upper bound, since the VCd of SVM is limited by the margin constraints, *i.e.* in the case of linear SVM $VCd \leq N+1$.

The training process starts by collecting non-pedestrian samples using a multiscale sliding window approach inside

ROIs which are defined by a LIDAR-based detection system. From each window detector a set of features are extracted by using two image descriptors: HOG[2] [14] and COV [15]. This ROI-based sliding window detection method is used, similarly, in the experiments. The training process continues stage-by-stage. For each stage n an iterative process is applied, in order to determine the optimal number of features, what in our case is the number of features that results in the minimal upper-bound of the expected risk, $R(\alpha)$. Therefore, the feature selector, which uses the mRMR method, is applied iteratively in order to select an increasing number of features from both HOG and COV descriptors. The bigger the number of features, N, the bigger the SVM complexity. The ensemble is trained by using a dataset with a different number of features for each stage, in order to obtain the empirical risk, $R_{emp}(\alpha)$. Then, the inequality (2) is computed by replacing the number of training data, l, and the upper bound of the VCd of the SVM, *i.e.* $h = N + 1$. The set composed by the "selected features + SVM" with the smallest $R(\alpha)$ is chosen to compose the current ensemble stage.

Regarding the training set, except for the first stage, each stage is trained by using a dataset which was generated by the previous stages. Namely, the training set which is used in the current stage n is composed by all the pedestrian training images (positives) and the false positives (negatives) which passed through all the previous stages, $1, \ldots, n-1$, during the scanning process of the training frames without pedestrians. Notice that, the training dataset used in stage n is a sub-set of the set which was applied in the training of the previous stage, $n-1$. In this context, it is important to change the set of features selected at each stage, what may enable the correct classification of the patterns which were miss-classified by the previous stage.

Algorithm 1 SRM-SVM cascade training process

Input: $\mathcal{D}_{train} = \mathcal{D}_{pos} \cup \mathcal{D}_{neg}$, N_c, and TP_{des}: training dataset with cropped images of pedestrians ($\mathcal{D}_{pos}$) and frames without pedestrians ($\mathcal{D}_{neg}$), number of cascade stages, and desired TP respectively;
Output: $\{w_n^*\}$, $\{b_n^*\}$, and $\{s_n^*\}$: sets of SVM parameters and set of vectors with the indexes of the selected features for each stage n, respectively;

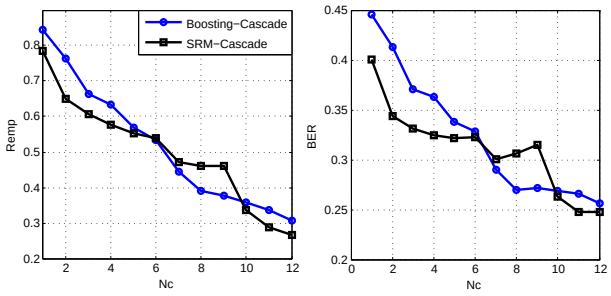
extract n_f HOG and COV features from $\mathcal{D}_{pos}$, in order to generate the training dataset S_{pos} of n_f-dimensional exemplars;
collect non-pedestrians samples from $\mathcal{D}_{neg}$, by using sliding window approach, in order to compose a training dataset with cropped images of non-pedestrians $\mathcal{D}_{neg}^*$;
compose the set of labels $\mathcal{L}_{train}$;
extract n_f HOG and COV features from $\mathcal{D}_{neg}^*$, in order to generate the training dataset S_{neg} of n_f-dimensional exemplars;
$n \leftarrow 1$: n is the stage index;
$S_n \leftarrow S_{neg} \cup S_{pos}$: training dataset of the current stage n;
$step \leftarrow 10$: increment in the number of features;
while $n \leq N_c$ **do**
  $R^* \leftarrow 1$: R^* is the maximum risk;
  for $N = 1 : step : n_f$ **do**
    select a set of N of the n_f features by applying the feature selector method [13] to the S_n dataset, and compose the training dataset $S_{(n,N)}$, whose exemplars are obtained from the exemplars of S_n by retaining only the N selected features;
    store the indexes of the selected features in vector s;
    apply $S_{(n,N)}$ and $\mathcal{L}_{train}$ to train a linear SVM, in order to obtain the SVM parameters $w_{(n,N)}$ and $b_{(n,N)}$;
    apply $S_{(n,N)}$, $\mathcal{L}_{train}$, $w_{(n,N)}$, and $b_{(n,N)}$ to compute $R_{emp}(\alpha)$ and TP;
    while $TP < TP_{des}$ **do**
      $b_{(n,N)} \leftarrow b_{(n,N)} + 0.05$: increasing the bias in order to increase the TP;
      recalculate TP and the empirical risk, R_{emp}, from $S_{(n,N)}$ and $\mathcal{L}_{train}$, using the learned SVM model with the current bias $b_{(n,N)}$;
    end while
    $h \leftarrow N + 1$: VC-dimension of the current SVM;
    replace $l = |S_{(n,N)}|$, R_{emp}, and h in (2), in order to obtain the expected risk $R(\alpha)$;
    if $R(\alpha) < R^*$ **then**
      $R^* \leftarrow R(\alpha)$;
      $w_n^* \leftarrow w_{(N,n)}$;
      $b_n^* \leftarrow b_{(N,n)}$;
      $s_n^* \leftarrow s$; // feature indexes
    end if
  end for
  scan the current training dataset, S_n, by using the current cascade stage, i.e. using the SVM with parameters w_n^*, b_n^*, and the set of features s_n^*, in order to collect a set of false positive occurrences, S_{FP};
  $S_{n+1} \leftarrow S_{FP} \cup S_{pos}$: composing the dataset for the next stage;
  $n \leftarrow n + 1$;
end while

VI. Results

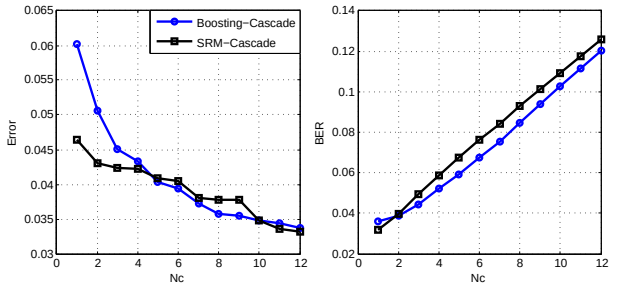
This paper presents experimental results obtained with the LIPD dataset. The empirical risk $Remp$ is calculated on the training set and, since we have the ground truth of our testing set, it is possible to calculate the error rate $Error$ of the cascade ensembles (with an increasing number of stages Nc). Since the dataset is unbalanced, the Balanced Error Rate (BER) was also used to evaluate the cascade performance. In our experiments a miss is computed when a full-visible pedestrian (type 1) is not detected, and a false positive occurs when an area with no label, *i.e.* a negative, is wrongly classified as pedestrian. The methodology used to assess the detection performance of the cascades is similar to the one described in [4]: the matching criterion is based on the intersection area, in pixels, between a window detector and the ground-truth bounding-box; if the area of a window detector covers more than 25% of the area of a ground-truth event, then a correct detection is considered. Moreover, a nonmaximum suppression method is used to discard multiple-detectors at close/identical locations. Thus, from the set of window detectors with a ratio of intersection area above 0.6, the detection with the greatest confidence output is retained and the remaining are discarded.

The cascade ensembles were trained with approximately 1.5×10^5 samples extracted from the 6328 negative frames of $\mathcal{D}_{train}$. The sliding window parameters used to extract the negative samples followed the definitions of the set S_4, from [4], which are: spatial stride ($\Delta_x = 0.1, \Delta_y = 0.025$) and scale step $\Delta_s = 1.25$. Notice that the samples (window detectors) were extracted from the ROI projections, instead of the whole image. From approximately 1.8×10^7 negative examples extracted using the methodology described above, a sample selection algorithm (which selects the training

[2]http://www.mathworks.com/matlabcentral/fileexchange/28689-hog-descriptor-for-matlab



(a) Empirical risk $Remp$ and training BER as function of Nc, for an adopted $Thr_{tp} = 0.98$.



(b) Testing error ($Error$) and testing BER as function of Nc, for an adopted $Thr_{tp} = 0.98$.

Fig. 5. Results regarding the training dataset (a) and the testing dataset (b), for the Boosting-SVM cascade (blue-circle markers) and for the SRM-SVM cascade (square-black markers). The left part shows the errors, while the right part presents the BER.

support vectors [16]) were employed to obtain the final 151528 negatives that, together with 5237 positive examples, were used to train the cascades. The results on the training set, for $Thr_{tp} = 0.98$ (**Algorithm 1**), are shown in Fig. 5(a) in terms of the empirical risk $Remp$ (left part) and in terms of the BER (right part) as a function of the number of stages Nc. The lowest empirical risk value in the Boosting-SVM cascade was 0.3061, for $Nc = 12$. On the other hand, the SRM-SVM cascade achieved 0.2658 of minimum on the empirical risk for $Nc = 12$. Moreover, the minimum BER for the Boosting-SVM cascade was 0.2569 (at $Nc = 12$), and the best result for the SRM-SVM cascade, for $Nc = 12$, was $BER = 0.2476$.

For the purpose of comparison, although at a preliminary stage[3], experiments on the testing set were performed as a function of the number of stages. Regarding the testing error ($Error$), left part of Fig. 5(b), the best generalization was observed for the SRM-SVM cascade, with an error of 0.0331, for $Nc = 12$, which is equivalent to a $FP = 0.0330$ and false negatives (FN) equal to 0.2189; the results for the Boosting-SVM cascade are: $Error = 0.0337$, for $Nc = 12$, corresponding to $FP = 0.0336$ and $FN = 0.2067$. In terms of BER, right side of Fig. 5(b), the SRM-SVM cascade obtained the minimum value of 0.0316 for $Nc = 1$, representing $FP = 0.0464$ and $FN = 0.01686$; while the Boosting-SVM cascade had a minimum BER of 0.0360.

[3]Our detection solution still demands improvements, in terms of the false-alarms, to achieve a realistic application.

VII. Conclusion

This work presented two methods for designing negative-rejection cascade detectors, composed by linear stages, emphasizing high unbalanced datasets collected by a monocular camera and a LIDAR. Experiments on pedestrian detection using image-based descriptors and range information were presented and the detection errors were compared for both SVM-based cascade ensembles: the Boosting-SVM cascade and the novel proposed SRM-SVM cascade.

A multi-sensor dataset was introduced, with measurements collected from an onboard sensor setup with data from a LIDAR, a monocular camera, an inertial unit, and a DGPS station. The LIPD dataset is specifically devoted for training and evaluation of LIDAR and image-based pedestrian detection methods in the context of IV/ITS safety systems.

The experiments, demanding several weeks of CPU time, were performed up to 12 stages. From the results over the testing dataset, the SRM-SVM cascade showed slightly better results than the Boosting-SVM cascade concerning the detection error.

References

[1] DARPA. The urban and grand challenges. [online]. http://www.darpa.mil/grandchallenge/, (accessed: 2010), 2003.

[2] ELROB. The european robot trial. [online]. http://www.elrob.org/, (accessed: 2011), 2006.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *IEEE Computer Vision and Pattern Recognition (CVPR 09)*, pages 304–311, Los Alamitos, CA, USA, 2009.

[4] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, Oct. 2009.

[5] T. Gandhi and M.M. Trivedi. Pedestrian protection systems: issues, survey, and challenges. *Intelligent Transportation Systems, IEEE Transactions on*, 8(3):413–430, 2007.

[6] David Geronimo, Antonio M. Lopez, Angel D. Sappa, and Thorsten Graf. Survey on pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1239–1258, 2010.

[7] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Vision and Pattern Recognition (CVPR 01)*, volume 1, pages 511–518, Hawaii, Dec. 2001.

[8] Yong Ma and Xiaoqing Ding. Face detection based on hierarchical support vector machines. *Pattern Recognition, International Conference on*, 1:1051–4651, 2002.

[9] C. Premebida, O. Ludwig, and U. Nunes. Lidar and vision-based pedestrian detection system. *Journal of Field Robotics*, 26(9), 2009.

[10] Q. Zhang and R. Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *IEEE Intelligent Robots and Systems Conference (IROS)*, Sept.-2 Oct. 2004.

[11] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[12] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.

[13] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, Aug. 2005.

[14] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Vision and Pattern Recognition (CVPR 05)*, pages 886–893, Washington, DC, USA, 2005.

[15] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *In Proc. 9th European Conf. on Computer Vision*, pages 589–600, 2006.

[16] Hans Peter Graf, Eric Cosatto, Léon Bottou, Igor Durdanovic, and Vladimir Vapnik. Parallel support vector machines: The cascade SVM. In *NIPS*, 2004.