Exploring Sensor Fusion Schemes for Pedestrian Detection in Urban Scenarios

C. Premebida, O. Ludwig, J. Matsuura and U. Nunes

***Abstract*— This work explores three schemes for pedestrian detection in urban scenarios using information gathered by a LIDAR and a monocular camera mounted on an electric vehicle. In the first scheme, pedestrian detection is conduct by a set of single classification methods trained with LIDAR and/or vision-based features. In the second scheme, the likelihoods from the single-classifiers are fused by means of three fusion rules: average, maximum value and a Naive-product rule. Finally, the third scheme is a cascade of classifiers with four-stages, two per feature space. All these pedestrian detection strategies were compared through off-line experiments conducted using our dataset, that is available on the Web for public usage.**

I. INTRODUCTION

The detection system presented in this work employs a multilayer LIDAR (an Ibeo-Alasca XT) and a monocular camera (Allied Guppy), mounted on an electrical vehicle, with the purpose of pedestrian detection in urban scenarios. Two sensor fusion architectures, a centralized (Fig. 1) and a decentralized one (Fig. 2), are used as platforms to design the pedestrian detection schemes. In the former, the LIDAR and the vision feature spaces are combined within a central/common feature space; in the latter, each sensor-based feature space is processed separately.

Based on the architecture to be used, three detection approaches have been investigated: one with single classifiers, a second scheme were the classifier's likelihoods are combined by means of fusion rules, and a third approach where a cascade of classifiers is used to carry out the pedestrian detection. We have named these detection schemes as: "single classifiers", "fusion scheme", and "cascade of classifiers".

Pedestrian detection using a single classifier, the most usual solution investigated so far [1] [2] [3], is characterized by the utilization of single classification methods to achieve the best possible classification score. On the other hand, the combination of classifiers involves the utilization of two or more classifiers that can be combined by means of many possible methods or algorithms; see [4] [5] as references on combining classifiers in the field of pattern recognition.

In the decentralized architecture discussed here, single classifiers are employed in both feature spaces separately, namely LIDAR-feature space (15 dimensional feature vector), and vision-based space (HOG, COV, HOG-COV); for the centralized case, a combined feature vector (composed by HOG-COV descriptors and LIDAR features) is employed. Moreover, in some specific cases, we have used the Peng method [6] to select 25% of the features with maximum relevance and minimal redundance. The features and the descriptors used throughout this paper are described in Section II.

Regarding the classifiers fusion methods, three fusion rules have been considered: average, max-value, and a product rule. More specifically, the "final" classification decision is obtained by means of fusion rules applied over the single classifiers likelihood responses. The single classification methods, the fusion rules and the proposed four-stage cascade of classifiers are described in Section III.

These strategies for pedestrian detection using LIDAR and vision data are compared in terms of the Area Under ROC-Curve (AUC), Accuracy (Acc) and the Balanced Error Rate (BER), whose results are presented in Section IV. Finally, conclusions are drawn in Section V.

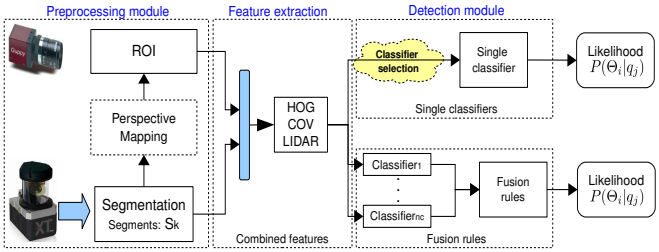


Fig. 1. Centralized sensor fusion architecture: LIDAR and vision information are combined at the feature level, and the pedestrian detection can be carried out by single-classifiers or fusion rules.

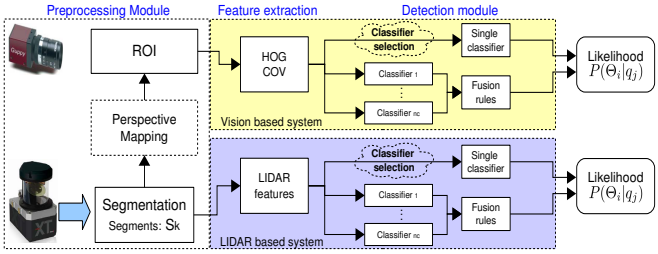


Fig. 2. Decentralized sensor fusion architecture: LIDAR and vision data are fused in the preprocessing module, and the further stages are processed separately. The final classification decision can be achieved per feature space or fusing the likelihoods from each sensor-based module.

II. PREPROCESSING AND FEATURE EXTRACTION

The sensor fusion architectures described in this paper are composed by three main modules: preprocessing; feature extraction; and detection (as illustrated in Figs. 1 and 2).

The authors are with the Department of Electrical and Computer Engineering, Institute of Systems and Robotics, University of Coimbra, Portugal. {cpremebida,oludwig,jmatsuura,urbano}@isr.uc.pt

The main processes involved in the first two modules are concisely described in this Section, and the detection module will be detailed in Section III.

The segmentation process, in conjunction with other pre-processing tasks, is the primary stage to detect the entities of interest, where each entity/object constitutes a hypothesis of being a positive (pedestrian) or a negative (non-pedestrian). The segmentation process is performed in the LIDAR space, where the detected objects are characterized by a group/cluster of laser-points, here named *segment*.

Expressing a 2D full scan as a sequence of $N_{\mathcal{S}}$ measurement points in the form $Scan = \{(r_l, \alpha_l) | l = 1, ..., N_{\mathcal{S}}\}$, where (r_l, α_l) denotes the polar coordinates of the l^{th} *scan point*, a group of scan points that constitute a segment S_k can be expressed as

$$S_k = \{(r_n, \alpha_n)\}, \quad n \in [l_i, l_f], \quad n = 1, \ldots, np \qquad (1)$$

where np is the number of points in the current segment, l_i and l_f are the initial and the final scan points that define the segment. A segment can also be defined in Cartesian coordinates $\mathbf{x} = (x_k, y_k)$, where $(x_k = r_n \cos \alpha_n, \quad y_k = r_n \sin \alpha_n)$. Henceforth, a *segment* is explicitly related to a group of range-points related to one, unambiguously, object of interest and expressed by S_k.

As the main objective of this work is to concentrate on pedestrian detection using LIDAR and vision based features, all the experiments were conducted using data sets where the segments were extracted under user supervision, avoiding some problems invariably presented on realistic situations, such as: data association errors, over-segmentation, measurements missing, tracking inconsistencies, etc. On the other hand, the images extracted from the ROIs in the image plane were not post-processed; it means that all the cropped images in the data set were extracted automatically from ROIs obtained using LIDAR segments projected in the image plane and, as consequence, are prone to error due to calibration imprecision, road irregularities, vehicle vibrations, and so on. Nevertheless, we decided to allow those cropped images with no user intervention or any correction, resulting in a closer realistic image-based data set. Therefore, our interest is in the situation where an entity, characterized by a segment (in the laser space) and by a projected ROI (in the image plane), has been already detected and should be classified as pedestrian or non-pedestrian.

A. LIDAR-based features

Features extracted from LIDAR data and its utilization for urban scenario interpretation is a subject that was investigated by [7], [8] and [9]; although the latter one was concentrated on indoor environments, many of the features used here are based on Arras's work. The components of the laser-based feature vector are detailed in [10]. As an example of a pedestrian detected by the LIDAR, with its corresponding segment and the region of interest (ROI) in the image plane, consider Fig. 3.

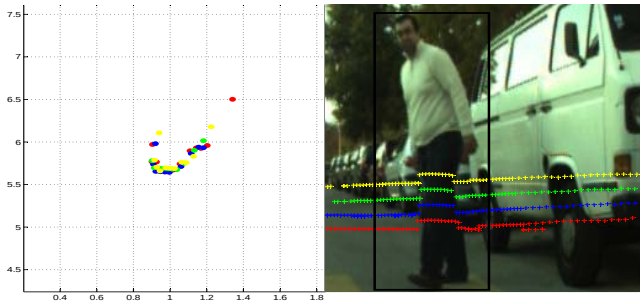


Fig. 3. An example that illustrates a pedestrian perceived by the laser, as a segment of range points, and the corresponding ROI in the image.

B. Vision-based features

A 261 dimensional vision-based feature vector, with 81 components due to HOG and 180 using COV image-descriptors, is addressed in the next subsections.

1) HOG features: Histogram of Oriented Gradients (HOG) [2] is inspired on Scale-Invariant Feature Transform (SIFT) descriptors proposed by [11]. To compose HOG, the cell histograms of each pixel within the cell casts a weighted vote, according to the gradient $L2$-norm, for an orientation-based histogram channel. In this work the histogram channels are calculated over rectangular cells (*i.e.* R-HOG) by the computation of unsigned gradient. The cells overlap half of their area, meaning that each cell contributes more than once to the final feature vector. In order to account for changes in illumination and contrast, the gradient strengths were locally normalized, *i.e.* normalized over each cell. The HOG parameters were adopted after a set of experiments performed over the training data set using a Neural Network as classifier (NN-MCI [12]). The higher Area Under ROC Curve (AUC), computed over the validation data set, was achieved by means of 9 rectangular cells and 9 bin histogram per cell. The nine histograms with nine bins were then concatenated to make a 81-dimensional feature vector.

2) COV features: The utilization of covariance matrices descriptors in classification problems was followed by [3] and [13]. Let I be the input image matrix, and z_p the corresponding d-dimensional feature vector calculated for each pixel p,

$$z_p = \left[x, y, |I_x|, |I_y|, \sqrt{I_x^2 + I_y^2}, |I_{xx}|, |I_{yy}|, arctan\frac{|I_y|}{|I_x|}\right] \qquad (2)$$

where x and y are the pixel p coordinates, I_x and I_y are the first order intensity derivatives regarding to x and y respectively, I_{xx} and I_{yy} are the second order derivatives, and the last term in (2) is the edge orientation.

In this work, four sub-regions are computed within a region R, which represents the area of a cropped image. Each sub-region overlaps half of its area, meaning that each sub-region contributes more than once to the final feature

vector. For the i^{th} rectangular sub-region R_i, the covariance matrix C_{R_i} is expressed by,

$$C_{R_i} = \frac{1}{S-1}\sum_{i=1}^{S}(z_i - \mu)(z_i - \mu)^T \quad (3)$$

where μ is the statistical mean of z_i and S is the number of sub-regions (in this case $S = 4$). Notice that, due to the symmetry of C_{R_i}, only the upper triangle part need to be stored, hence the covariance descriptor of a sub-region is an 8×8 matrix. The features of the whole region R are also calculated, therefore a feature vector with 180 features is generated, *i.e.*, 4 sub-regions R_i, totalizing 144 features, plus 36 features of the whole region R.

III. PEDESTRIAN DETECTION

In this Section we will cover the following subjects: the single-classifiers, the fusion scheme, and the cascade of classifiers.

A. Single classifiers scheme

Five classifiers (FLDA, Naive-Bayes, GMMC, SVM-RBF and NN-MCI), see [14] for a concise description, have been used as decision functions to separate the feature space in two parts, that is, pedestrians and non-pedestrians classes. The single classifiers have been trained considering three feature spaces: LIDAR, vision, and combined-features (see Fig. 4). In the two latter feature spaces, the Naive-Bayes and the GMMC classifiers were trained with 25% of the features to avoid inconsistencies (*e.g.* singularities on the covariance matrix, or likelihoods tending to zero). The minimal-redundancy-maximal-relevance (mRMR) method [6] has been used to selected 25% of the most relevant and less redundant features. Table I summarizes, for both proposed architectures, the percentage of feature utilization per classification method.

A single-classifier can also be used as basis for a trainable fusion method; basically, a trainable fusion algorithm is also a classifier that receives the likelihoods from single-classifiers, which has been trained previously, and outputs the likelihood of the ensemble classifier [10], [15].

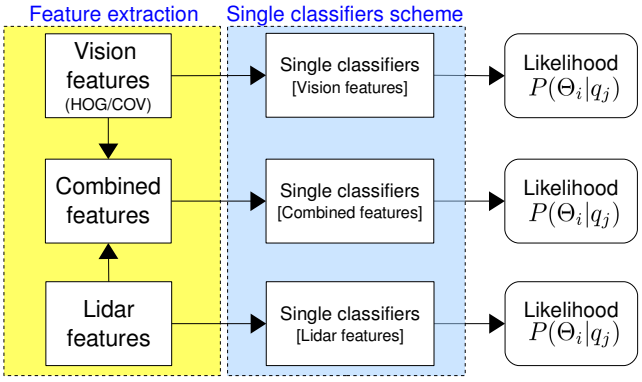


Fig. 4. Functional diagram representation for the single-classifiers scheme.

TABLE I

SINGLE CLASSIFIERS ARRANGEMENT IN TERMS OF THE ARCHITECTURE AND THE FEATURES UTILIZATION (IN %).

Features	Naive	GMMC	FLDA	SVM-RBF	NN-MCI
Decentralized architecture					
LIDAR	100%	100%	100%	100%	100%
Vision	25%	25%	100%	100%	100%
Centralized architecture					
LIDAR + Vision	25%	25%	100%	100%	100%

B. Fusion scheme

Three fusion rules were applied to combine the likelihood response yielded by each classifier (as shown in Fig. 5). More specifically, the classifiers likelihood outputs ($P(\Theta_i|q_j), i = 1, nc$), where nc is the number of classifiers and j is the number of classes, are combined using the average rule $Avage(\Theta_i)$, the maximum rule $Max(\Theta_i)$ and a Naive-product rule $Nprod(\Theta_i)$ (Naive Bayes inspired rule). Given $q_1 \equiv$ *pedestrian* and $q_2 \equiv$ *non-pedestrian*, the average rule is expressed by

$$Avage = \frac{\sum_{i=1}^{nc} P(\Theta_i|q_j)}{nc} \quad (4)$$

and the maximum rule

$$Max = \max(P(\Theta_i|q_j)) \quad (5)$$

The Naive-product rule is expressed as the product of the likelihoods normalized by the *evidence* ($P(\Theta_i)$):

$$Nprod = \frac{\prod_{i=1}^{nc} P(\Theta_i|q_j)}{P(\Theta_i)} \quad (6)$$

where $P(\Theta_i) = \prod_{i=1}^{nc} P(\Theta_i|q_1) + \prod_{i=1}^{nc} P(\Theta_i|q_2)$.

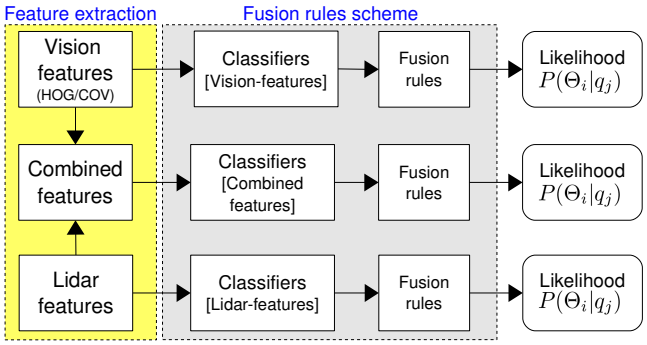


Fig. 5. Fusion scheme diagram representation. The likelihoods, yielded by the single-classifiers, are combined by means of fusion rules.

C. Cascade of classifiers

The cascade scheme is based on four single classifiers: the NN-MCI and the GMMC forming the first two-stages (in the LIDAR-feature space), and two FLDA trained with COV and HOG-COV features respectively (vision-feature space). Figure 6 illustrates the proposed cascade of classifiers. The classifiers composition, used to form the cascade scheme,

was selected considering the sub-optimal separation in the likelihood distributions, within the feature spaces, using the training data set.

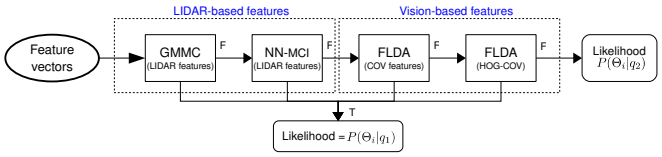


Fig. 6. Functional diagram of the cascade of classifiers for pedestrian detection. F and T denotes False and True respectively.

IV. EXPERIMENTAL RESULTS

The proposed detection system was evaluated in terms of Acc, AUC, BER, and ROC curves. Data sets, employed in the training and evaluation of the classifiers, are summarized in Table II. The $ISR - UC\_train$ and $ISR - UC\_test$ data sets were acquired in the ISR-UC *Campus*, under the following configuration: the LIDAR field of view (FOV) was restricted to 180^o, with a horizontal angular resolution of 0.5^o, vertical resolution of $[-1.2^o, -0.4^o, 0.4^o, 1.2^o]$, and measurement range up to $30m$; the camera FOV was 67^o approximately. The data sets and the corresponding ground truth, generated under user supervision, are online available (http://www.isr.uc.pt/~cpremebida/dataset). The data set has been recorded in the ISR-UC *Campus*[1] open surrounding areas, with many static and moving pedestrian and cars around, using the sensor apparatus and the electric vehicle shown in the Fig. 7, which was driven manually with a maximum speed of 40Km/h approximately.

The training and the testing data sets were collected on different dates. Although both data sets were acquired around the same area, the positives (pedestrians) and the negatives are clearly different; another relevant aspect is that on the testing part, some samples were acquired at dusk, where the illumination condition changed drastically. Some images of the data sets are shown in Fig. 8.

Fig. 7. Electric vehicle and the sensors setup used in the data set acquisition.

The FLDA and the Naive-Bayes are base classifiers, not requiring parameters adjustments. Eventually, for the case of Naive-Bayes classifiers, the likelihood tends to zero depending on the feature distributions. Based on the results obtained in the training phase, the number of Gaussian components for the GMMC classifier was set to 4, and the margin parameter

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

Fig. 8. Some samples to illustrate the different conditions and situations in which the data sets have been acquired.

TABLE II
DATA SET USED TO TRAIN AND TO EVALUATE THE CLASSIFIERS

Training data set				
Name	Total	Npos.	Nneg.	Description
$ISR - UC\_train$	2680	1340	1340	Sunny day, winter, collected between 15:30 to 16:30
Testing data set				
$ISR - UC\_test$	1400	400	1000	Sunny day, winter, collected between 12:00 to 17:30

used in the RBF-SVM was 1000. Moreover, the MCI-NN was configured with 30 and 200 neurons in the first and second layers respectively.

A. Single classifiers scheme

The experimental results regarding the single classifiers, depending on the fusion architecture and the feature space, are summarized in Table III, where the best results are highlighted in bold. In the decentralized architecture, the Naive-Bayes classifier had the best performance in the LIDAR space, while the GMMC and the NN-MCI achieved the best scores in the vision feature space. Moreover, in the centralized architecture, the GMMC obtained the best results. ROC curves of the best single classifiers using the decentralized architecture are depicted in Figs. 10(a), 11(a) and 12(a).

It is important to note that the performance metrics (BER, Acc and AUC) are calculated over all samples presented on the testing data set, hence those scores serve as a global indicator of the classification performance. Hence, to support specific analysis, we also used the value $TPr_{10\%}$, i.e. the true positives rate related to 10% of false positives (FPr = 0.1).

TABLE III

PERFORMANCE RESULTS FOR SINGLE-CLASSIFIERS SCHEME: TESTING DATASET

	FLDA	Naive-Bayes	GMMC	SVM-RBF	NN-MCI
Decentralized architecture: LIDAR-based features					
AUC	0.866	**0.928**	0.909	0.876	0.915
Acc	0.745	**0.880**	0.851	0.835	0.877
BER	0.181	**0.109**	0.133	0.183	0.127
$TPr_{10\%}$	0.432	**0.835**	0.825	0.655	0.835
Decentralized architecture: Vision-based features					
	FLDA	Naive-Bayes	GMMC	SVM-RBF	NN-MCI
AUC	0.892	0.874	0.876	0.768	**0.911**
Acc	0.848	0.811	**0.864**	0.821	0.839
BER	0.169	0.159	0.191	0.239	**0.146**
$TPr_{10\%}$	0.657	0.5	**0.842**	0.620	0.657
Centralized architecture: combined features					
	FLDA	Naive-Bayes	GMMC	SVM-RBF	NN-MCI
AUC	0.915	**0.931**	0.928	0.774	0.909
Acc	0.881	0.897	**0.914**	0.827	0.844
BER	0.112	0.091	**0.090**	0.242	0.175
$TPr_{10\%}$	0.775	0.885	**0.943**	0.614	0.767

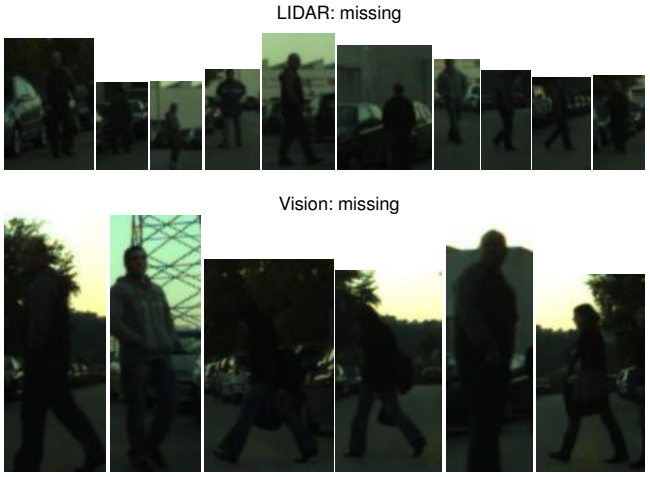


Fig. 9. Some examples of miss detections using single classifiers with LIDAR and vision features.

B. Fusion rules scheme

The fusion scheme performance, where the final likelihood is obtained by means of the fusion rules, is summarized in Table IV. ROC curves *w.r.t.* decentralized and centralized architectures are shown in Figs. 10(b), 11(b) and 12(b). The average rule achieved the best result with the LIDAR and the combined feature spaces; on the other hand, the Nprod-rule had the best scores with vision features.

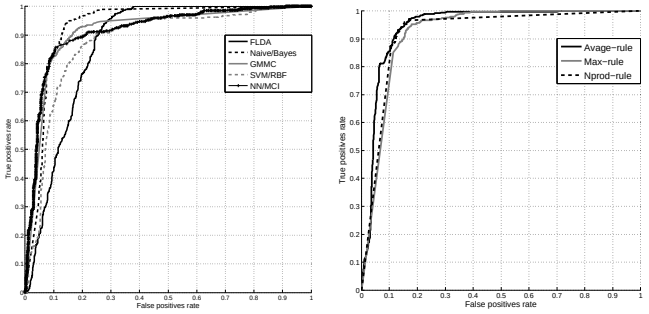
C. Cascade of classifiers

The cascade scheme performance is summarized in Table V, and the ROC curve is shown in Fig. 13. Regarding this proposed cascade of classifiers, many configurations can be experimented during the training stage to decide the best arrangement to be used with the testing data set. Although we have trained multiple variations for the cascade scheme (changing the classifier method and/or the feature space), some of them with perfect separation within the training data set, the final *selected* cascade structure did not achieved remarkable results in the testing data set.

TABLE IV

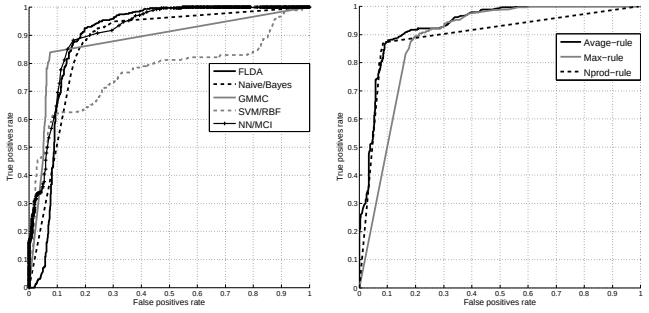
PERFORMANCE RESULTS FOR THE FUSION SCHEME: TESTING DATASET

	Avage-rule	Max-rule	Nprod-rule
Decentralized architecture: LIDAR-based features			
AUC	**0.940**	0.923	0.915
Acc	**0.884**	0.681	0.877
BER	**0.101**	0.224	0.105
$TPr_{10\%}$	**0.855**	0.736	0.812
Decentralized architecture: Vision-based features			
	Avage-rule	Max-rule	Nprod-rule
AUC	**0.928**	0.884	0.890
Acc	0.881	0.702	**0.894**
BER	0.149	0.215	**0.123**
$TPr_{10\%}$	0.875	0.504	**0.875**
Centralized architecture: combined features			
	Avage-rule	Max-rule	Nprod-rule
AUC	**0.948**	0.940	0.930
Acc	**0.917**	0.793	0.917
BER	0.083	0.147	**0.081**
$TPr_{10\%}$	**0.953**	0.922	0.946



(a) ROC: single classifiers scheme. (b) ROC: fusion scheme.

Fig. 10. The ROC curve for single-classifiers and fusion schemes in the LIDAR-based feature space.

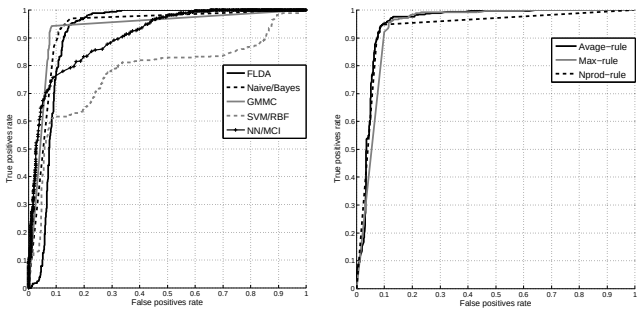


(a) ROC: single classifiers scheme. (b) ROC: fusion scheme.

Fig. 11. The ROC curve for single-classifiers and fusion schemes in the vision-based feature space.

V. CONCLUSIONS

In this paper we have investigated different schemes to combine information from a LIDAR and a camera, mounted on an electric vehicle, for pedestrian detection in an urban



(a) ROC: single classifiers scheme. (b) ROC: fusion scheme.

Fig. 12. The ROC curve for single-classifiers and fusion schemes for the combined feature space.

TABLE V
PERFORMANCE RESULTS FOR THE CASCADE SCHEME: TESTING DATASET

	Cascade of classifiers
AUC	0.846
Acc	0.845
BER	0.185
$TPr_{10\%}$	0.652

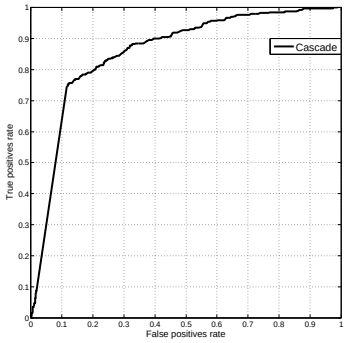


Fig. 13. ROC for the cascade scheme.

scenario. The fusion process was accomplished at the feature level (combining features from LIDAR and vision space) and at the classifiers level (using the likelihoods from the single-classifiers). Regarding the experimental results, the following conclusions can be drawn for each proposed scheme:

1) Single-classifiers scheme: using this approach, the basis for any other classification strategy, the GMMC classifier with 25% of the selected combined-feature vector achieved the best performance, followed by the empirical Naive-Bayes classifier within the LIDAR-based features and the NN-MCI using HOG-COV descriptors in the vision space;
2) Classifiers fusion scheme: regarding the fusion rules, the average of the likelihoods in the centralized architecture attained the best result against other fusion rules. However this scheme obtained the best classification, the results did not differ significantly compared to the single classifiers scheme;
3) Cascade of classifiers: although this scheme seems to be very appealing on pedestrian classification, motivated mainly by Viola-Jones work, this solution did not reveal promising results in the context discussed here; thus further studies will be conducted.

VI. ACKNOWLEDGMENTS

This work is supported in part by Fundação para a Ciência e a Tecnologia de Portugal (FCT), under Grant PTDC/EEA-ACR/72226/2006. C. Premebida is supported by FCT under grant SFRH/BD/30288/2006 and O. Ludwig under grant SFRH/BD/44163/2008. The authors would like to thank Antonio M. Gonçalves for his helpful assistance.

REFERENCES

[1] L. Spinello and R. Siegwart. Human detection using multimodal and multidimensional features. In *Robotics and Automation, ICRA. IEEE International Conference on*, pages 3264–3269, May 2008.
[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, CVPR. IEEE Conference on*, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.
[3] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *In Proc. 9th European Conf. on Computer Vision*, pages 589–600, 2006.
[4] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. In *IEEE Trans. on Pattern Anal. and Machine Intelligent*, volume 20, pages 226–239, 1998.
[5] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.
[6] Hanchuan Peng, Fuhui 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.
[7] D. Streller and K. Dietmayer. Object tracking and classification using a multiple hypothesis approach. In *Intelligent Vehicles Symposium, IVS. IEEE*, pages 808–812, June 2004.
[8] B. Douillard, D. Fox, and F. Ramos. A spatio-temporal probabilistic model for multi-sensor object recognition. In *Intelligent Robots and Systems, IROS. IEEE/RSJ International Conference on*, pages 2402–2408, 29 2007-Nov. 2 2007.
[9] K.O. Arras, O.M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In *Robotics and Automation, IROS. IEEE Int. Conference on*, pages 3402–3407, April 2007.
[10] Cristiano Premebida, Oswaldo Ludwig, and Urbano Nunes. Lidar and vision-based pedestrian detection system. *Journal of Field Robotics*, 26(9):696–711, 2009.
[11] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.
[12] O. Ludwig and U. Nunes. Improving the generalization properties of neural networks: an application to vehicle detection. In *Intelligent Transportation Systems, ITSC. IEEE International Conference on*, pages 310–315, Oct. 2008.
[13] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *Computer Vision and Pattern Recognition, CVPR. IEEE Conference on*, pages 1–8, June 2007.
[14] C. Premebida, O. Ludwig, and U. Nunes. Exploiting lidar-based features on pedestrian detection in urban scenarios. *Intelligent Transportation Systems, ITSC. IEEE Int. Conference on*, October, 2009.
[15] Oswaldo Ludwig, David Delgado, Valter Gonçalves, and Urbano Nunes. Trainable classifier-fusion schemes: an application to pedestrian detection. In *Intelligent Transportation Systems, ITSC. IEEE International Conference on*, 2009.