Olmeda, D., Premebida, C., Nunes, U., Armingol, J.M., & Escalera, A . (2013). Pedestrian Detection in Far Infrared Images. *Integrated Computer-Aided Engineering*, 20 (4), pp. 347-360.
DOI: 10.3233/ICA-130441

Pedestrian Detection in Far Infrared Images

Daniel Olmeda[a,*], Cristiano Premebida[b], Urbano Nunes[b], Jose Maria Armingol[a] and Arturo de la Escalera[a]

[a]*Department of Systems Engineering, Intelligent Systems Lab, Universidad Carlos III de Madrid, Spain.*
Email:{dolmeda,armingol, escalera}@ing.uc3m.es
[b]*Department of Electrical and Computer Engineering, Institute of Systems and Robotics, University of Coimbra, Portugal.*
Email:{cpremebida,urbano}@isr.uc.pt

Abstract. This paper presents an experimental study on pedestrian classification and detection in far infrared (FIR) images. The study includes an in-depth evaluation of several combinations of features and classifiers, which include features previously used for daylight scenarios, as well as a new descriptor (HOPE - Histograms of Oriented Phase Energy), specifically targeted to infrared images, and a new adaptation of a latent variable SVM approach to FIR images. The presented results are validated on a new classification and detection dataset of FIR images collected in outdoor environments from a moving vehicle. The classification space contains 16152 pedestrians and 65440 background samples evenly selected from several sequences acquired at different temperatures and different illumination conditions. The detection dataset consist on 15224 images with ground truth information. The authors are making this dataset public for benchmarking new detectors in the area of intelligent vehicles and field robotics applications.

Keywords: Pedestrian Detection, Far Infrared, Advanced Driver Assistance Systems

1. Introduction

Image analysis and computer vision is lately being incorporated in civil engineering [29, 48, 14]. The applications are plenty and include freeway work zone analysis [37, 1, 38], automatic image search [35], human detection and modeling [12, 70] and face recognition [7]. Object recognition in images has become a very important topic in the fields of traffic infrastructure and driving assistance system [31, 54]. Applications such as traffic signs recognition [56, 13, 65, 33], obstacle avoidance [20] and traffic surveillance [59] have gotten the attention of the industry for some time now. The case of people detection is an exceptionally relevant case, as it leads to a number of important applications, some of which strive for saving lives.

Pedestrian recognition in images is geared toward a variety of applications, which include safety focused road infrastructures [36], driver assistance sys-tems [26] and autonomous robotic vehicles [55]. It is also useful in security, be it for automatic surveillance or people counting [63, 69]. Those applications can be extended to low visibility conditions by using FIR images. Another application that could benefit of automatic recognition of people in low light conditions is unmanned aerial vehicles aimed to rescue missions [71].

Most of the recent research in this topic is based on visible light (VL) images. FIR images share some key characteristics with their VL images counterparts. They both are 2D representations of a scene captured by redirecting electromagnetic waves by means of a lens, light in the first case and infrared radiation, which is proportional to the objects temperature, in the second. Some of the key ideas on pedestrian classification in VL images can be extended to work on FIR images, exploiting common characteristics of both, or adapt them to take benefit of the different kinds of information provided by FIR images.

---

[*] Corresponding author. E-mail: dolmeda@ing.uc3m.es.

Regarding pedestrian detection in VL images, there exists a reasonable number of benchmark datasets publicly available, such as: MIT [53], CVC [27], TUD-det [8], INRIA [16], DC [47], ETH [22] and Caltech [18]. For an overview of recent work on pedestrian detection on these datasets, the authors refer to [26, 63, 19, 21]. In the case of FIR images, the authors found a lack of a complete pedestrian dataset that could serve as a tool to benchmark new features and methods.

In this paper the authors present an in-depth analysis of several well-known VL pedestrian classifiers applied to FIR images: Principal Component Analysis (PCA), Local Binary Patterns (LBP) [49] and Histogram of Oriented Gradients (HOG) [16]. Moreover, experimental results are compared with the proposed descriptor specifically targeted to FIR images: Histograms of Oriented Phase Energy (HOPE) [51].

The results derived from this study where obtained from a new pedestrian dataset, that the authors are making public[1]. Our dataset is divided in two parts, classification and detection. The Classification Dataset contains a preset of cropped images of positives (pedestrians) and negatives (background), rescaled to the same dimensions. The Detection Dataset contains full size images and labels indicating the position and dimensions of each pedestrian.

In short, there are three main contributions in this paper: (i) a new pedestrian dataset in far infrared images; (ii) an extensive study on the pedestrian descriptors using FIR images, and (iii) an adaptation of a latent variable SVM approach [24] to FIR images. This paper is structured as follows. Section 2 includes a brief discussion on pedestrian datasets and the relevant state of the art in pedestrian detection. It also covers an overview of descriptors and FIR image-based classification methods. Characteristics of the FIR image-based pedestrian dataset are discussed in sections 3 and 4, including the methodology of acquisition and sample selection, as well as useful statistics. Section 5 and section 6 focus on the features and methods used for classification and detection performance assessment. In sections 7 and 8 experiments for pedestrian classification and detection, respectively, are presented and further discussed in section 9.

---

[1] http://www.uc3m.es/islab/repository

2. Related Work

2.1. Pedestrian Datasets

The availability of publicly released datasets for pedestrian classification has been a key element that helped advances in this area. It provides a way for researchers to test and benchmark new classification algorithms in a way that can be directly compared with other works. It is also useful for replicating experiments performed by other research groups.

In this domain, datasets are usually divided into two types: classification and detection datasets. In the first one, a fixed set of cropped windows containing pedestrians and background is provided, while detection datasets consist on full images with annotated locations of pedestrians. Usually, a subset of full-frames, with no positives (pedestrians), is provided for negative examples extraction. The method for background sample extraction varies from one author to the other, so the classifiers are not really trained on the same data.

A classification dataset is useful for approaches based on the sliding window paradigm. This detection technique consists on analyzing an image by shifting a fixed sized window in the horizontal and vertical axis. This approach can be extended to a multi resolution search by incrementally resizing the original image. Each window analysis becomes independent from all the others and, as such, the detection turns into a classification problem. Improving the classifier performance would also improve detection performance. The classification performance is usually expressed in terms of miss rate vs. false negative rate per window, while per frame is more suitable for detection performance.

In [47] Munder and Gavrila introduced the DC classification dataset. It consists of 4000 up-right pedestrian and 25000 background samples captured in outdoor urban environments. All of them are resized to 18×36 pixels. In their work, the authors evaluate Haar, Principal Component Analysis (PCA) [30] and Local Receptive Fields (LRF) in combination with neural networks and Support Vector Machine (SVM) classifiers [15] [41]. From their results it can be concluded that the size of the dataset is a key element in improving the classification performance. For the extraction of a large number of background images they apply bootstrapping [64] techniques. The dataset is split into 3 train and 2 test subsets, for cross-validation purposes.

In [16] Dalal et al. presented the INRIA dataset, which is still widely used nowadays. It consists on 2478 128 × 64 cropped images of people for training, and 566 for testing, along with full images for negative extraction. The images were selected from a collection of photographs acquired in urban and rural scenes, and not initially thought to serve as a dataset for driving assistance systems.

More recently, Dollar et. al introduced in [18] the Caltech Detection Dataset, as well as a benchmark of several pedestrian detection algorithms. Their results were further extended in [19]. This dataset contains approximately 250k labeled pedestrians within several video sequences acquired from a moving vehicle in urban traffic. In their work, the authors directly compare results with other publicly available pedestrian datasets, as the Daimler detection dataset [21].

2.2. Image Based Descriptors

In [53] Papageorgiou et al. introduced a sliding window detector based on SVM trained with a Haar wavelet feature vector. This work inspired Viola and Jones face detector, presented in [66], which accelerates the classification step by applying a cascade approach.

Along with the INRIA dataset, Dalal et al. introduced the HOG descriptor [16]. It became a benchmark for pedestrian classification, due to their excellent results. Since then, there have been many new features using local histograms [63], including classifiers targeted to FIR images [51].

Color or gray level value has also been exploited as a descriptor of the shape of a person. In [12] Ciarelli at al. create a live color model with a modified Mean Shift algorithm. More recently, Self Similarity (SS) has been proposed as a descriptor for pedestrian classification in visible light images in [67]. This method encodes the distribution of color as repetition across the image. In [46] this approach was adapted to work in mono channel images, and tested in a FIR pedestrian classification problem.

2.3. FIR image-based methods

Given that FIR images represent the heat emitted by people, and do not need external illumination, some authors have developed classification methods based on the temperature distribution of the human body. Most systems take advantage of this feature and select regions of interest based on the distribution of the warm parts of the image [9,10]. A review

of techniques pedestrian automotive infrared detection can be found in [52]. In these systems, the discriminating feature of pedestrians is the body shape, and the regions of interest are validated by correlation with predefined probabilistic models [50].

3. Classification Dataset

One of the contributions of this work is our pedestrian classification dataset, which consists of FIR images collected from a vehicle driven in outdoors urban scenarios. The dataset was recorded in Leganés, Spain and Coimbra, Portugal. Images were acquired with an Indigo Omega imager, with a resolution of 164 × 129 pixels, a grey-level scale of 14 bits, and focal length of 318 pixels. The camera was mounted on the exterior of the vehicle, to avoid infrared filtering of the windshield.

Recorded images were manually annotated, where each pedestrian is labeled as a bounding box. To prevent bias introduced by border artifacts their height is subsequently upscaled by 5%. Fig. 1 shows some cropped-image examples of positives and negatives of the classification dataset. The pedestrians appear in an up-right position. In order to not introduce a pose bias, some tolerance is allowed. Specifically, images where head or torso is at a slant due to the normal walking gate are not discarded.

3.1. Dataset Statistics

3.1.1. Number of samples
The dataset comprises 81592 14-bit one-channel images, divided in 16152 positives and 65440 negatives. The train set contains 10208 positives and 43390 negatives, while the test set contains 5944 positives and 22050 negatives.

3.1.2. Aspect ratio
Out of the annotated images, the bounding boxes are resized to a constant aspect ratio (w/h) = 0.5 by changing their width (w) and height (h) appropriately. Figure 2 contains histograms for heights, widths and areas of positive and negative bounding boxes. The height of positive bounding boxes has a mean of 40 pixels. Those bounding boxes refer to pedestrians standing at approximately 10m from the camera as seen in the histogram of Fig. 3. However, less frequent cases appear up to 50m. Any bounding box below 10 pixels in height is ignored. The remaining

Figure 1. Example cropped images of the classification dataset. The upper row contains examples of pedestrians acquired under different temperatures and illumination conditions. The lower row contains randomly selected windows from images containing no pedestrians. For visualization purposes the contrast has been enhanced.

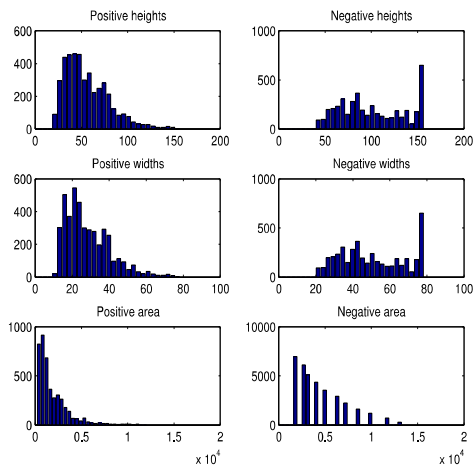bounding boxes are resized to 64×32 pixels using bilinear interpolation.

Figure 2. Histograms of bounding boxes sizes and areas for positive and negative samples of the train dataset. X axis represent the measure (positive and negative height, width and area) and Y axis the number of samples having those value.

3.1.3. Density

Images were acquired from the usual point of view of the driver. As such, pedestrians appear more often in the center of the image as shown in Fig 4, which represents the logarithmic density of the centers of the bounding boxes. In the case of negative samples, the bounding boxes are randomly selected, so the centers appear all over the image, with less density near the borders.
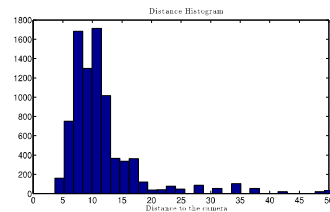

Figure 3. Histograms of pedestrian's distance to the camera for the train and test dataset.
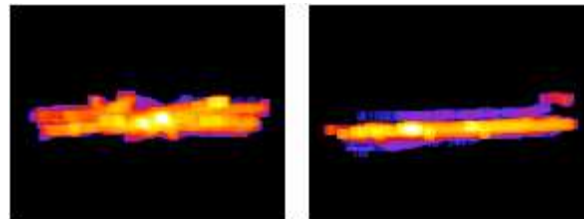

Figure 4. Centers of bounding boxes for positives of the train and test dataset on a logarithmic scale. Left image: Positives in Train Dataset. Right image: Positives in Test Dataset.

4. Detection Dataset

The detection dataset contains the full frames from which the classification dataset was extracted, along with manual annotations of the pedestrian's positions.

The detection dataset was acquired in 13 different sessions, each containing a varying number of images. It comprises 15224 14-bit one-channel images, with dimension 164×129 pixels. The train set contains 6159 images, and the test set contains 9065 images. Only non-occluded pedestrians are considered for future evaluation, therefore images containing pedestrians with more than 20% of the area of the

original bounding box occluded behind other obstacles are disregarded.

Each session occurred at a different location and with different illumination and temperature conditions. Out of those sessions 6 were used to compose the train set, leaving the remaining 7 for test set. This ensures that Train and Test are independent from one another. The temperature at which they were shot, which in turn affects the grey level and the histogram spread, causes the most important difference in appearance between sequences. Fig. 5 contains the histogram of the mean grey level value of the train and test detection datasets.

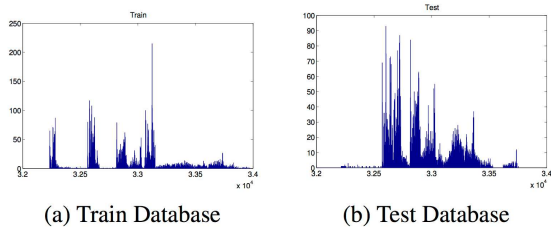(a) Train Database   (b) Test Database

Figure 5. Histogram of mean gray level of the images in the Train and Test Datasets. X axis all possible values of gray level and Y axis the number of pixels having those levels. Left Image: Train Dataset. Right Image: Test Dataset.

5. Image-based features

In this section the feature selection is discussed, along with implementation details.

LBP. Local Binary Patterns (LBP), as introduced in [49], represents the image as a similarity vector of each pixel with their surroundings. This descriptor encodes information as a binary number. For each pixel, the neighbors with a gray value higher or equal contribute with one in their position in the binary number, otherwise with zero. Each sample is divided in 3×3 pixel non-overlapping cells.

HOG. In this work we have used 5×5 pixel non-overlapping cells. In our 64×32 dataset this means 10×4 cells per image, once removed the border blocks. Within each cell a 9 bins histogram of orientation between 0 and 2π radians is calculated.

HOPE. Defining the shape of a pedestrian in FIR images is challenging using one dimensional point derivatives. This is due to the much wider infrared spectrum, compared with visible light. Another difficulty is that the sensitivity curve of an uncooled microbolometer sensor changes very quickly with min-

imum changes of its temperature [32]. To overcome these challenges, we proposed in [51] a contrast invariant descriptor for pedestrian classification in FIR images called HOPE. Basically, the HOPE descriptor encodes a grid of local oriented histograms extracted from the phase congruency of the images, which is computed from a joint of Gabor filters.

The histograms are calculates in 5×5 pixel non-overlapping cells with 9 bins of orientation between 0 and 2π radians, for a total of 10×4 cells per image. No normalization step is applied. Phase congruency was calculated out of a set of 30 complex Gabor filters, divided in 5 scales ranging between a minimum wavelength of 2 pixels and a maximum of 10, and 6 orientations, ranging from 0 to 2π radians.

PCA. We treat PCA [45, 30] eigenvectors as a grey-level feature vector. The initial motivation for applying this approach is that PCA tends to disregard small details at high frequency, as seen in Fig. 6, while FIR images usually have poor levels of detail, as they present softness due to motion blur, especially at low resolutions. We retain the 30 most significant eigenvectors, that is, those with the largest eigenvalues.

Figure 6. First five eigenpedestrians

Feature Concatenation. Descriptor fusion is explored as feature vector concatenation, resulting in a new higher dimension feature vector with different kinds of complementary information, which can improve the overall performance.

Selecting those features in the train dataset with a higher score in Welch's t-test shortens high-dimensional feature vectors. The minimum number of features selected is set so that the resulting feature vector has an accuracy on the test dataset within 1% of the unabridged vector. This usually results in feature vectors with half the dimensions of the original.
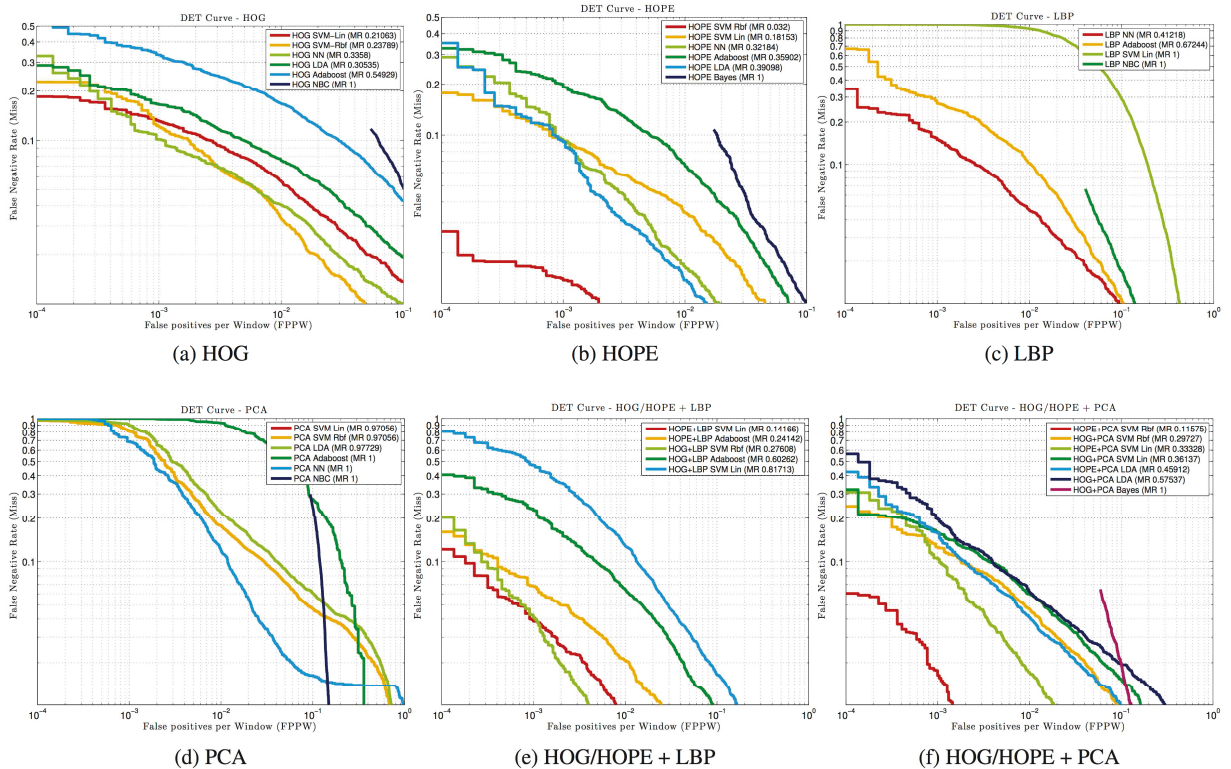
(a) HOG

(b) HOPE

(c) LBP

(d) PCA

(e) HOG/HOPE + LBP

(f) HOG/HOPE + PCA

Figure 7. DET curves for each feature vector in the classification dataset. Legend states Miss Rate (MR) at 10^{-4} FPPW.

6. Classification Methods

Pedestrian classification is treated as a supervised pattern recognition problem, where the set x D $= \{\{x_1, y_1\}, \cdots, \{x_n, y_n\}\}$, is the collection of manually labeled examples, $x_i \in R^d$ is a feature vector and $y_i \in \{\pm 1\}$ is a binary label. In this section five kinds of classification methods have been used: SVM, Naïve Bayes Classifier (NBC), Quadratic Discriminant Analysis (QDA), Neural Networks (NN) and Adaboost. The parameters selected for the different classifiers are discussed in the sequel.

Support Vector Machines. Concerning SVM [11], two different kernels were used for benchmarking: a linear classifier, hereafter called SVM-Lin, and a radial basis function kernel (RBF) [2, 42], designated by SVM-Rbf. In this implementation the radial Gaussian function kernel $K(x, y) = e^{-\gamma \|x-y\|^2}$ has a scale parameter $\gamma = 1$. Both linear and RBF kernel have a regularization trade-off parameter C = 0.05.

Naïve Bayes Classifier. NBC [40] is a member of the bayesian classifiers family [57] [39]. It is designed for use when features are independent of one

within each class, but it appears to work well in practice in other circumstances. Naive Bayes classification is based on estimating the conditional probability of the feature vector given the class.

Discriminant Analysis. Linear Discriminant Analysis [45, 68, 58] is used as a linear classification model in terms of dimensionality reduction. Considering a two-class separation problem the D-dimensional input vector x can be projected down to one dimension as $y = w^T x$, where w is the components weight vector. Selecting appropriate weights the projection can be done over the dimension that maximally separates both classes, avoiding the exceeding overlapping that can occur due to dimensionality reduction. Over this projection a threshold w_0 is selected, where values $y \leq -w_0$ are classified as pedestrians, whereas values $y > -w_0$ are classified as background. In this implementation the coefficient matrix of the boundary equation is quadratic thus, the discriminant analysis takes a quadratic form, designated QDA, assuming normally distributed classes. The multivariate normal densities are fitted with covariance estimates stratified by group.

6

Neural Network. A neural network pattern recognition scheme [34] is used with a two-layer feed-forward network, with ten hidden and one output sigmoid neurons. The network is trained with scaled conjugate gradient backpropagation. The overall network function follows equation 1, where σ is the sigmoid function, *rnk* is the output ranking, N is the number of inputs and M is the maximum number of linear combinations of the N inputs.

$$rnk(x,w) = \sigma\left(\sum_{j=1}^{M} w_j^{(2)} h\left(\sum_{i=1}^{N} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_0^{(2)}\right) \quad (1)$$

Neural Networks are a very popular pattern recognition technique in the field of ITS and infrastructure engineering [28] [3] [43], and many variations of it have been developed in these fields [62] [61] [5] [6] [44].

Adaboost. We use Real AdaBoost as described in [60]. The key idea is that the combined response of a set of weak classifiers can build a strong one, improving the performance that a complex classifier alone would have. Iteratively, Adaboost selects a threshold that best separates each feature set xi in one of the classes y_i, applying a higher weight to misclassified samples. In this implementation the maximum number of iterations is set to 50. The final ranking of each feature vector is $rnk = \sum_{i=1}^{N} x_i(f_i)$. In the case of HOG and HOPE, each bin in the orientation histograms is treated as a weak feature.

7. Evaluation on the classification dataset

In this section we compare the performance of each feature discussed in section 5 by applying the classification techniques described in section 6.

The train dataset is used to train the feature-classifier combinations. Likewise, testing is performed on the whole test dataset. Classification performance is evaluated by means of Detection-Error Trade-off (DET) curves, which quantify the trade-off between miss rate and false positive rate.

Results for different ensembles of features and classifiers are shown in figure 7. From these curves it can be observed that approaches based on local orientated histograms, such as HOG and HOPE, get better results than PCA or LBP. The best performing feature seems to be HOPE, with a miss rate of 0.3%

at 10^{-4} false positives (FP) for the SVM-Lin classifier, followed by HOG with a miss rate of 0.38% at 10^{-4} FP. With an RBF kernel performance improves up to 0.06% miss rate at 10^{-4} FP in the case of HOPE and 0.25% at 10^{-4} FP in the case of HOG. DET curves of the SVM-Lin results are shown in Fig. 8. SVM-Rbf results are plotted in case of a significant gain.
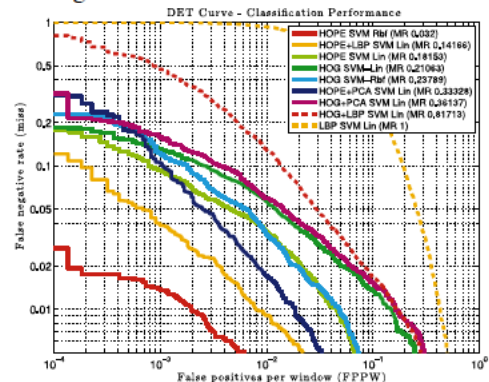

Figure 8. DET curves of the best performing SVM-Lin classifiers. Results with Rbf kernels are plotted where they represent a significant gain. Legend states Miss Rate (MR) at 10^{-4} FPPW.

Regarding feature combination, we have used an SVM-Lin to assess the impact of the features in the classification performance. LBP features combined with HOPE significantly reduce the miss rate, by 18% at 10^{-4} FP. PCA, although getting better results as an independent classifier than LBP, does not improve significantly classification when merged with other features. When merged with HOG the miss rate is reduced by 6% at 10^{-4} FP, while the improvement of merging it with HOPE is negligible.

Concerning the classification methods, SVM-Rbf generally has the best performance followed by SVM-Lin. LDA classifier performs almost as well, or better than Linear SVM for the HOG and HOPE descriptors. NN showed some promising results, though performance could be improved by using a robust training scheme, as the one presented in [4]. The NBC showed the worst performance, except for LBP features.

Additionally we performed some experiments to assess the impact on classification performance by varying the number of negative examples on the train set. Fig. 9 shows that, for the HOPE SVM-Rbf classifier, the performance gets significantly better by increasing the number of train negatives and no saturation is appreciated.
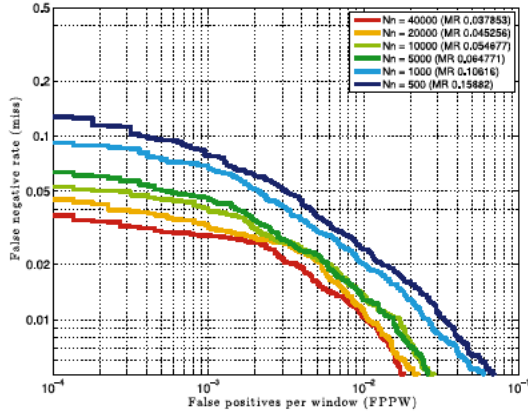
Figure 9. DET curve of the HOPE SVM-Rbf classifier trained with an increasing number of negatives. Legend states Miss Rate (MR) at 10−4 FPPW.

7.1. Statistical Significance of the Results

Statistical significance is assessed with McNemar's approximate test [23]. It is used to compare two classifiers at a particular value of bias. To determine whether classifier (C1) is significantly better than (C2), the χ^2 statistic is used (equation 2).

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \qquad (2)$$

Where n_{01} is a number of cases misclassified by C1 and classified correctly by C2, and n_{10} is a number of cases misclassified by C2 and classified correctly by C1. The null hypothesis H_0 states that the performance of both classifiers is the same. H_0 hyphothesis may be rejected if χ^2 falls below a probability of 5%, i.e. $\chi^2_{1,0.95} \geq 3.841$. If that is the case, it can be assumed that one classifier performs significantly better than the other.

Table 1 contains χ^2 values for every pair of clas-

sifiers used. The bias of all classifiers has been $b = 0$, after rescaling, as it is the value that maximally separates both classes. From these results it can be concluded that the null hypothesis can be rejected for all classifier pairs.

8. Performance on the detection dataset

Pedestrian detection resumes on finding the position and scale of an a priori unknown number of pedestrian on a set of full images. Detection is evaluated by applying the best performing feature-classifier constructs explained in sections 5 and 6, in a sliding window approach. For brevity, only best performing ensembles (HOG and HOPE in combination with SVM-Lin, SVMRbf and Adaboost) are further discussed.

For multi-resolution detection purposes the input image is resized to 6 different scales per octave, with an average scaling step between scales of $ss = 1.1$. This results in a scanning window that searches for pedestrians between 120 and 24 pixels in height. At each scale, the detection window is moved sideways so that each one has an 80% overlap with the previous one. Likewise, it is moved from top to bottom so that the overlapping is 90% in the vertical axis. This results in 2377 windows per image.

8.1. Evaluation Methodology

The detection task is evaluated by the Pascal Criteria [15], plotting results in DET curves. Detect ions are considered true or false positives based on the area of overlap with ground truth bounding boxes. To be considered a correct detection, the area of overlap a_o between the predicted bounding box B_p and ground-truth bounding-box B_{gt} must exceed 50% by the equation 3, as depicted in Fig. 10.

Table 1. Results of the McNemar's approximate significance test for every pair of classifiers. The value expressed in the table's fields is χ2, as stated in equation 2.

	HOGLin	HOGRbf	HOPELin	HOPERbf	LBP	PCA
HOGLin	0	247.1	81.5	342.1	1852.9	1508.6
HOGRbf	247.1	0	51.6	11.5	2825.5	2443.7
HOPELin	81.5	51.6	0	108.0	2427.6	2056.1
HOPERbf	342.1	11.5	108.0	0	2980.8	2596.7
LPB	1852.9	2825.5	2427.6	2980.8	0	23.8
PCA	1508.6	2443.7	2056.1	2596.7	23.8	0

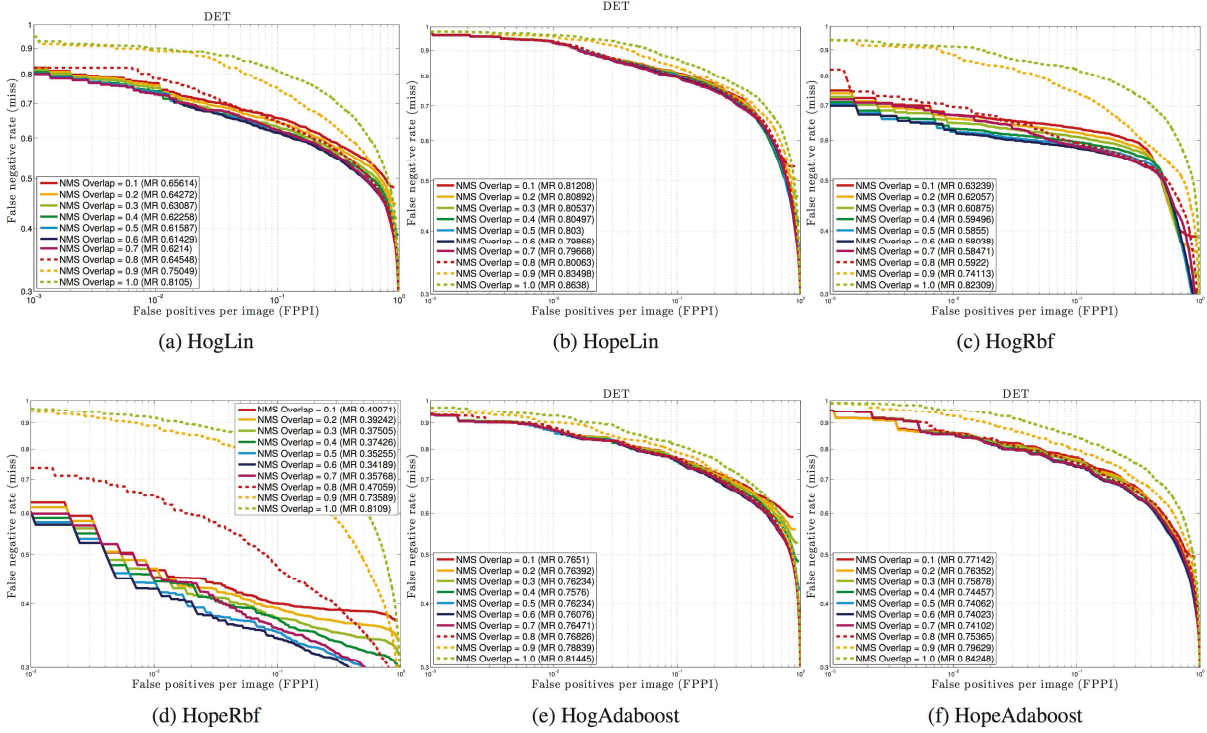| (a) HogLin | (b) HopeLin | (c) HogRbf |
| (d) HopeRbf | (e) HogAdaboost | (f) HopeAdaboost |

Figure 10. DET curves for each feature vector in the detection database. For each non-maximum suppression overlap area threshold, one curve is plotted. Legend states Miss Rate (MR) at 0.1 FPPI.

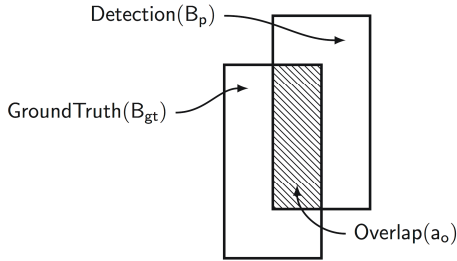$$a_o = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \qquad (3)$$

Figure 11. Overlapping area of Ground Truth and Detection.

Figure 12. Example of non maximum suppression of multiple detections for each pedestrian.

In Fig. 12 one curve is plotted for each overlapping percentage tested, in the range $a_0 = \{0.1,...,1.0\}$ for every detector considered. The legend states Miss Rate at 0.1 FPPI.

8.2. Results

In Fig. 13 the DET curves of the best performing detectors are plotted. Best results were obtained with the HOPE-Rbf detector, with a 35% miss rate at 0.1 False Positives per Image (FPPI), followed by the HOGRbf, with a 58% miss rate at 0.1 FPPI. Using a linear SVM, the HOG descriptor get an slightly higher miss rate of 76% at 0.1 FPPI while for the HOG feature results degrade up to 80% at 0.1 FPPI. With an Adaboost approach the miss rates are 76% at 0.1 FPPI for the HOG descriptor and 74% at 0.1 FPPI for the HOPE descriptor. Based on these results it might seem that there is a correlation between classification and detection results. Per image results present a correlation with the per window results.

For each pedestrian, it is usual that many detections appear in the neighborhood around the ground-truth bounding-box. If two or more detections match the same ground-truth bounding-box, only the one with the higher score would be considered a true positive. Other overlapping detections are considered false positives. To minimize the number of repeated detections, a greedy non-maximum suppression (NMS) algorithm, pairwise max (PM) suppression [25], is applied to all bounding boxes. It selects iteratively detections with higher scores than their neighborhood, discarding detections with lower scores over an overlapping percentage. This overlap is again calculated with equation 3. Fig. 11 shows an example of multiple detections for the same pedestrians, and the result after applying the NMS algorithm.
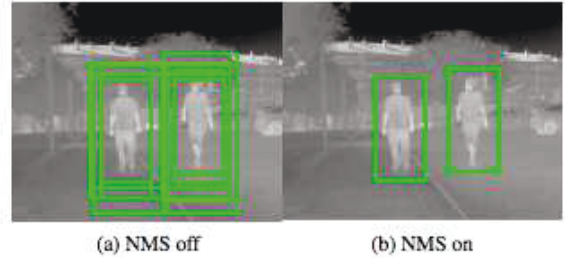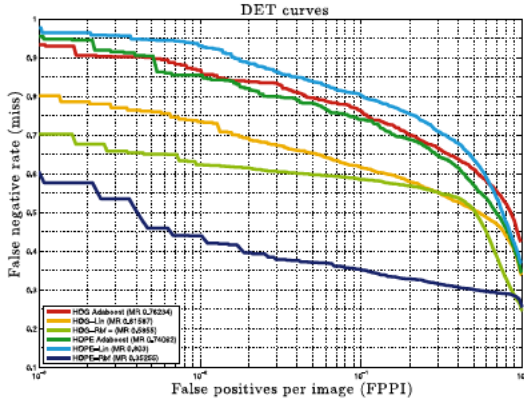
Figure 13. Detection DET Curves after applying the PM NMS algorithm with an overlap threshold of $a_o = 0.5$. Legend states Miss Rate (MR) at 0.1 FPPI

8.2.1. Impact of Pedestrian size

Pedestrian size has a big impact on detection results. Pedestrians located at a long distance from the vehicle appear at a lower resolution on the image and, as such, have lower detection rates. Some experiments are performed to assess the impact on detection performance by subdividing the test set into several subsets, i.e. small pedestrians ($[10 - 40]$ pixels in height), medium ($[40 - 80]$) and large ($[80 - 120]$). While evaluating a subset, ground-truth bounding boxes not belonging to a given subset are assigned to class 0. A detection window overlapping any of the subsets is not counted as a true positive or as a false positive. Likewise, a missed detection of a class 0 bounding box is not counted as a false negative. Fig. 14 shows the miss rate at 10^{-3} FPPI for the six applied detectors. Higher resolution pedestrians get the lowest miss rate in all cases.

Scanning the image looking for very small pedestrians also have an impact on computational demands of the algorithm in a sliding window detector.
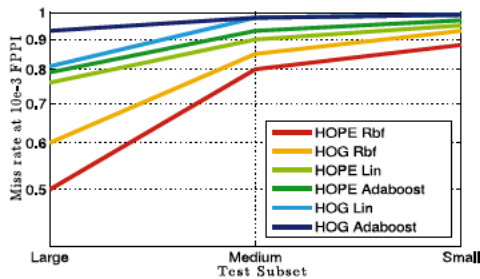

Figure 14. Miss rate at 10^{-3} FPPI for all detectors in the Small, Medium and Large Test Subsets.

8.3. Latent-SVM HOPE

One of the limitations of classifiers based on dense histograms of orientations is that performance degrades for very small cells [16]. Intuitively, finer grained descriptors retain more details, so classification should benefit from it. However, due to non-rigid deformations of pedestrians the overall classification scores for high-resolution descriptors are low. To overcome these limitations Felzenszwalb et al. introduced in [24, 25] their latent SVM detector. This classification method relies on a set of filters: a low-resolution root filter and a set of high resolution part filters that define a hidden or latent structure. The locations of the parts of the pedestrian that best define its presence on the image are the latent variables z. While training, the exact location of the ground-truth bounding-box of positive examples is also a latent variable. This allows for auto-correcting mistakes made while labeling the dataset.
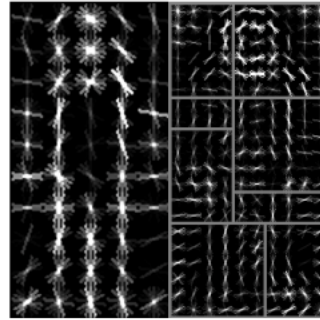

Figure 15. Latent SVM filters using HOG feature.

The detection is treated as a binary classification problem in a sliding window approach. Given a training set $D = \{\{x_1, y_1\}, \ldots \{x_n, y_n\}\}$, where $x_i \in R_d$ is a feature vector and $x_i \in \{\pm 1\}$ is a binary label, each region of interest of the image is assigned a score,

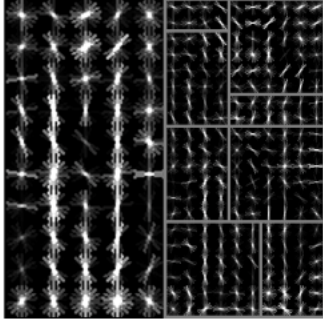$$f_{\beta,Z}(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z) \qquad (4)$$

Figure 16. Latent SVM filters using HOPE feature

Where x is the region of interest, $Z(x)$ is the set of all possible positions of the parts, and β is a vector of model parameters. The function $\Phi(x, z)$ is the feature vector assembled from the root filter and the best latent parts. β should then minimize equation 5.

$$\frac{1}{2}\|\max_{i=1, ,k}\beta\| + C\sum_{i=1}^{n}\max(0, 1 - y_i \cdot f_{\beta,z}(x_i)) \quad (5)$$

Where k is the number of components in the model, and the second term determines the softness of the SVM margin.

To assess the impact on performance of part based detection, two descriptors have been trained using the Latent-SVM approach: HOG and HOPE. Detection performance is evaluated using the same methodology explained in section 8.1, except for the following exception: the minimum overlap threshold between B_p and B_{gt} is set to a more restrictive value of $a_o = 0.7$.

The latent-SVM models for both the HOG and HOPE descriptors have been trained using the same parameters as their standalone counterparts. They have been initialized for k = 8 latent parts and (6 × 6) cells. The descriptors are extracted from a feature pyramid, where the parts filter has double the resolution as the root filter. Fig. 15 represents the filters trained with the HOG descriptor and Fig. 16 the filters trained with the HOPE descriptor.

The latent parts are computed at twice the resolutions as the root filter. In low-resolution images, such as the ones presented in this database, latent parts are hard to find in the smallest objects. Because of this limitation, the following results are based on a sub-sample of the testing database. Only images in the Medium and Large subsets are used. For a pedestrian with less than 40 pixels in height the size of a cell of

the root filter would have (3 × 3) pixels, therefore it is not possible to construct the part filters at double that resolution. All other images are previously resized to double their size in order to satisfy minimum resolution restriction mentioned above.

The DET curves in Fig 17 compares the performance of root, latent and parts detectors for the HOG and HOPE descriptors. Root filters are the original low-resolution descriptors trained with an extended set of warped positives. There is a notable decrease in miss rate when compared with their standalone counterparts, due to the extended training set, allowing the detector to cope with subtle pose variations. Introducing latent detections (i.e. automatic estimations of the pedestrians correct position) seems to degrade slightly the performance. The overall best results for *Large* pedestrians derive from searching for pedestrian parts. This seems to help specially in the case of semi-occluded pedestrians. As the pedestrians get smaller in the images, adding the parts filter loses it edge, due to greater resolution demands. For comparing purposes the DET curve of the original HOG-Lin tested under the same conditions is also included in the figure.
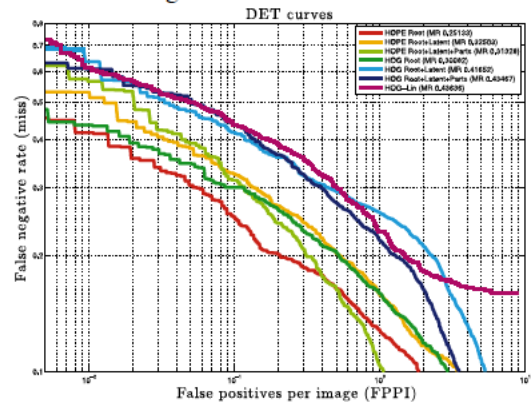

Figure 17. Detection DET Curves of Latent-SVM HOG and HOPE after applying the PM NMS algorithm with an overlap threshold of $a_o = 0.7$. Legend states Miss Rate (MR) at 0.1 FPPI.

Finally, a qualitative inspection of misclassified samples suggests that ambient temperature has a determinant impact on performance. Sequences collected at a high environment temperature or under direct sun light present the most false positives, and also the highest miss rate. A qualitative evidence of this issue is shown in Fig.18 where we present examples of misclassified positives and negatives. Other sources of misclassification are motion blur, which in FIR images appear frequently, and pose variation. False positives appear mostly in negative examples with a high vertical symmetry.

Figure 18. Misclassified samples. Upper row. False negatives due to low resolution, motion blur and pose variation Lower row: False positives in areas with a high vertical symmetry

9. Conclusions and Discussion

This paper presents an experimental study on pedestrian classification in FIR images. Several combinations of descriptor and classification methods have been tested in a new FIR dataset. By our best knowledge this is the first complete FIR based pedestrian classification and detection dataset publicly available for benchmarking.

From the experimental results reported in the previous sections it can be concluded that histogram based features perform best than LBP or PCA features. Among the features, HOPE performs better both for classification and detection problems. LBP and PCA features get worse overall performance, though some gains can be achieved by merging them with HOG or HOPE. In terms of classification methods, SVM achieved the best performance. The RBF kernel can significantly reduce misclassifications compared with a linear kernel, but is more computationally demanding. This is a critical factor in computer aided transportation applications.

Pedestrian resolution has an important impact on performance, with all detectors having lower hit rates for small pedestrians. For the best performing detector, the miss rate at 10^{-3} goes from 50% for *Large* pedestrians to 88% for *Small* ones. In the case of Latent SVM approach, the detectors are not able to detect any pedestrian in the *Small* subsample. From these results it may be concluded that current pedestrian detectors in FIR images suffer from the same limitations as VL detectors when it comes to detect small pedestrians. This is specially relevant in this case, as FIR cameras tend to have low resolution sensors.

The results presented in this paper suggests that FIR images are a very useful source of information for pedestrian classification and detection, having similar performance to that found in state of the art in VL images, with advantage in low visibility applications.

References

[1] Hojjat Adeli and Samanwoy Ghosh-Dastidar. Mesoscopic-wavelet freeway work zone flow and congestion feature ex traction model. *Journal of Transportation Engineering,* 130(1):94–103, 2003.

[2] Hojjat Adeli and Asim Karim. Fuzzy-wavelet rbfnn model for freeway incident detection. *Journal of Transportation Engineering,* 126(6):464–471, 2000.

[3] Hojjat Adeli and Ashif Panakkat. A probabilistic neural net work for earthquake magnitude prediction. *Neural Networks,* 22(7):1018–1024, 2009.

[4] Mehran Ahmadlou and Hojjat Adeli. Enhanced probabilistic neural network with local decision circles: A robust classifier. *Integrated Computer-Aided Engineering,* 17(3):197–210, 2010.

[5] Sultan Uddin Ahmed, Md Shahjahan, and Kazuyuki Murase. A lempel-ziv complexity-based neural network pruning algorithm. *International Journal of Neural Systems,* 21(05):427–441, 2011.

[6] MAH Akhand and Kazuyuki Murase. Ensembles of neural networks based on the alteration of input feature values. *International journal of neural systems,* 22(01):77–87, 2012.

[7] Mohammad Al-Naser and Ulrik Söderström. Reconstruction of occluded facial images using asymmetrical principal component analysis. *Integrated Computer-Aided Engineering,* 19(3):273–283, 2012.

[8] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on,* pages 1–8. IEEE, 2008.

[9] Massimo Bertozzi, Alberto Broggi, Mike Del Rose, Mirko Felisa, Alain Rakotomamonjy, and Frédéric Suard. A pedestrian detector using histograms of oriented gradients and a support vector machine classifier. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE,* pages 143–148. IEEE, 2007.

[10] Emanuele Binelli, Alberto Broggi, Alessandra Fascioli, Stefano Ghidoni, Paolo Grisleri, Thorsten Graf, and M Meinecke. A modular tracking system for far infrared pedestrian recogni-

tion. In *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE,* pages 759–764. IEEE, 2005.

[11] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery,* 2(2):121–167, 1998.

[12] Patrick Marques Ciarelli, Evandro OT Salles, and Elias Oliveira. Human automatic detection and tracking for outdoor video. *Integrated Computer-Aided Engineering,* 18(4):379–390, 2011.

[13] Gustavo A Peláez Coronado, María Romero Muñoz, José María Armingol, Arturo de la Escalera, Juan Jesús Muñoz, Wouter van Bijsterveld, and Juan Antonio Bolaño. Detection and classification of road signs for automatic inventory systems using computer vision. *Integrated Computer-Aided Engineering,* 19(3):285–298, 2012.

[14] S Rolando Cruz-Ramírez, Yasushi Mae, Tatsuo Arai, Tomohito Takubo, and Kenichi Ohara. Vision-based hierarchical recognition for dismantling robot applied to interior renewal of buildings. *Computer-Aided Civil and Infrastructure Engineering,* 26(5):336–355, 2011.

[15] H. Dai, H. Zhang, W. Wang, and G Xue. Structural reliability assessment by local approximation of limit state functions using adaptive markov chain simulation and support vector regression. *Computer-Aided Civil and Infrastructure Engineering,* 27(9):676–686, 2012.

[16] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on,* volume 1, pages 886–893. IEEE, 2005. 2, 3, 9

[17] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation,* 10(7):1895–1923, 1998.

[18] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on,* pages 304–311. IEEE, 2009.

[19] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 34(4):743–761, 2012.

[20] Valentin Enescu, Geert De Cubber, Kenny Cauwerts, Hichem Sahli, Eric Demeester, Dirk Vanhooydonck, and Marnix Nuttin. Active stereo vision-based mobile robot navigation for person tracking. *Integrated Computer-Aided Engineering,* 13(3):203–222, 2006.

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

[22] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on,* pages 1– 8. IEEE, 2007.

[23] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision,* 88(2):303–338, 2010.

[24] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on,* pages 1–8. IEEE, 2008.

[25] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 32(9):1627–1645, 2010.

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

[27] David Gerónimo, A Sappa, Antonio López, and Daniel Ponsa. Adaptive image sampling and windows classification for onboard pedestrian detection. In *Proceedings of the International Conference on Computer Vision Systems, Bielefeld, Germany,* 2007.

[28] Samanwoy Ghosh-Dastidar and Hojjat Adeli. Wavelet-clustering neural network model for freeway incident detection. *Computer-Aided Civil and Infrastructure Engineering,* 18(5):325–338, 2003.

[29] Samanwoy Ghosh-Dastidar and Hojjat Adeli. Neural network-wavelet microsimulation model for delay and queue length estimation at freeway work zones. *Journal of Transportation Engineering,* 132(4):331–341, 2006.

[30] Samanwoy Ghosh-Dastidar, Hojjat Adeli, and Nahid Dadmehr. Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection. *Biomedical Engineering, IEEE Transactions on,* 55(2):512–518, 2008.

[31] Richard Harrison, Roger Birchall, Dave Mann, and Wenjia Wang. Novel consensus approaches to the reliable ranking of features for seabed imagery classification. *International journal of neural systems,* 22(06), 2012.

[32] MK Hinders and WL Fehlman. Passive infrared thermographic imaging for mobile robot object identification. In *AIP Conference Proceedings,* volume 1211, page 2076, 2010.

[33] Sina Hooshdar and Hojjat Adeli. Toward intelligent variable message signs in freeway work zones: Neural network model. *Journal of Transportation Engineering,* 130(1):83–93, 2003.

[34] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences,* 79(8):2554–2558, 1982.

[35] Jin Hou, Zeng Chen, Xue Qin, and Dengsheng Zhang. Automatic image search based on improved feature descriptors and decision tree. *Integrated Computer-Aided Engineering,* 18(2):167–180, 2011.

[36] Xiaomo Jiang and Hojjat Adeli. Freeway work zone traffic delay and cost optimization model. *Journal of transportation engineering,* 129(3):230–241, 2003.

[37] Xiaomo Jiang and Hojjat Adeli. Freeway work zone traffic delay and cost optimization model. *Journal of transportation engineering,* 129(3):230–241, 2003.

[38] Xiaomo Jiang and Hojjat Adeli. Object-oriented model for freeway work zone capacity and queue delay estimation. *Computer-Aided Civil and Infrastructure Engineering,* 19(2):144–156, 2004.

[39] Xiaomo Jiang, Sankaran Mahadevan, and Hojjat Adeli. Bayesian wavelet packet denoising for structural system identification. *Structural Control and Health Monitoring,* 14(2):333–356, 2007.

[40] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence,* pages 338– 345. Morgan Kaufmann Publishers Inc., 1995.

[41] Vilen Jumutc, Pawel Zayakin, and Arkady Borisov. Ranking-based kernels in applied biomedical diagnostics using a support vector machine. *International Journal of Neural Systems,* 21(06):459–473, 2011.

[42] Asim Karim and Hojjat Adeli. Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with california algorithm. *Journal of Transportation Engineering,* 128(1):21–30, 2002.

[43] Chih-Min Lin, Ang-Bung Ting, Chun-Fei Hsu, and Chao-Ming Chung. Adaptive control for mimo uncertain nonlinear systems using recurrent wavelet neural network. *Int. J. Neural Syst.*, 22(1):37–50, 2012.

[44] Fan Liu and Meng Joo Er. A novel efficient learning algorithm for self-generating fuzzy neural network with applications. *International journal of neural systems*, 22(01):21–35, 2012.

[45] Aleix M Martínez and Avinash C Kak. Pca versus lda. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):228–233, 2001.

[46] Alina Miron, Bassem Besbes, Alexandrina Rogozan, Samia Ainouz, and Abdelaziz Bensrhair. Intensity self similarity features for pedestrian detection in far-infrared images. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 1120–1125. IEEE, 2012.

[47] Stefan Munder and Dariu M Gavrila. An experimental study on pedestrian classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1863–1868, 2006.

[48] Takafumi Nishikawa, Junji Yoshida, Toshiyuki Sugiyama, and Yozo Fujino. Concrete crack detection by multiple sequential image filtering. *Computer-Aided Civil and Infrastructure Engineering*, 27(1):29–47, 2012.

[49] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[50] Daniel Olmeda, Arturo de la Escalera, and José María Armingol. Far infrared pedestrian detection and tracking for night driving. *Robotica*, 29(04):495–505, 2011.

[51] Daniel Olmeda, Arturo de la Escalera, and José María Armingol. Contrast invariant features for human detection in far infrared images. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 117–122. IEEE, 2012.

[52] Ronan O'Malley, Martin Glavin, and Edward Jones. A review of automotive infrared pedestrian detection techniques. In *Signals and Systems Conference, 208.(ISSC 2008). IET Irish*, pages 168–173. IET, 2008.

[53] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.

[54] François Petitjean, Florent Masseglia, Pierre Gancarski, and Germain Forestier. Discovering significant evolution patterns from satellite image time series. *International Journal of Neural Systems*, 21(06):475–489, 2011.

[55] Cristiano Premebida and Urbano Nunes. Fusing lidar, camera and semantic information: A context-based approach for pedestrian detection. *The International Journal of Robotics Research*, 32(3):371–384, 2013.

[56] Rahul Putha, Luca Quadrifoglio, and Emily Zechman. Comparing ant colony optimization and genetic algorithm approaches for solving traffic signal coordination under oversaturation conditions. *Computer-Aided Civil and Infrastructure Engineering*, 27(1):14–28, 2012.

[57] Siamak Saliminejad and Nasir G Gharaibeh. A spatial- bayesian technique for imputing pavement network repair data. *Computer-Aided Civil and Infrastructure Engineering*, 27(8) 594–607, 2012.

[58] A Samant and H Adeli. Feature extraction for traffic incident detection using wavelet transform and linear discriminant analysis. *Computer-Aided Civil and Infrastructure Engineering*, 15(4):241–250, 2000.

[59] A Sánchez, EO Nunes, and Aura Conci. Using adaptive background subtraction into a multi-level model for traffic surveillance. *Integrated Computer-Aided Engineering*, 19(3):239–256, 2012.

[60] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[61] F-M Schleif, Thomas Villmann, Barbara Hammer, and Petra Schneider. Efficient kernelized prototype based classification. *International Journal of Neural Systems*, 21(06):443– 457, 2011.

[62] Liangdong Shi, Yinghuan Shi, Yang Gao, Lin Shang, and Yubin Yang. Xcsc: a novel approach to clustering with extended classifier system. *Int. J. Neural Syst.*, 21(1):79–93, 2011.

[63] D Simonnet, SA Velastin, E Turkbeyler, and J Orwell. Backgroundless detection of pedestrians in cluttered conditions based on monocular images: a review. *Computer Vision, IET*, 6(6):540–550, 2012.

[64] K.K. Sung and Tomaso Poggio. Example-based learning for view-based human face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):39–51, 1998

[65] Yichang Tsai and Yuchun Huang. A generalized framework for parallelizing traffic sign inventory of video log images using multicore processors. *Computer-Aided Civil and Infrastructure Engineering*, 27(7):476–493, 2012.

[66] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[67] Stefan Walk, Nikodem Majer, Konrad Schindler, and Bernt Schiele. New features and insights for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1030–1037. IEEE, 2010.

[68] Andrew P Worth and Mark TD Cronin. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure THEOCHEM*, 622(1):97–111, 2003.

[69] Tao Xiong, Mengping Zhang, Chi-Wang Shu, SC Wong, and Peng Zhang. High-order computational scheme for a dynamic continuum model for bi-directional pedestrian flows. *Computer-Aided Civil and Infrastructure Engineering*, 26(4):298–310, 2011.

[70] Tao Xiong, Mengping Zhang, Chi-Wang Shu, SC Wong, and Peng Zhang. High-order computational scheme for a dynamic continuum model for bi-directional pedestrian flows. *Computer-Aided Civil and Infrastructure Engineering*, 26(4):298–310, 2011.

[71] Chunsun Zhang and Ahmed Elaksher. An unmanned aerial vehicle-based imaging system for 3d measurement of unpaved road surface distresses. *Computer-Aided Civil and Infrastructure Engineering*, 27(2):118–129, 2012.