RNPS 2487 ▪ ISSN 2708-3411
Vol. 2 ▪ Nro. 1 ▪ enero-marzo 2021 ▪ págs. 33-45

ARTÍCULO ORIGINAL

Un estudio de la generalización en la clasificación de peatones

A Study of Generalization in Pedestrian Classification

Franco Ronchetti
fronchetti@lidi.info.unlp.edu.ar ▪ http://orcid.org/0000-0003-3173-1327
Facundo Quiroga
fquiroga@lidi.info.unlp.edu.ar ▪ https://orcid.org/0000-0003-4495-4327
Genaro Camele
gcamele@lidi.info.unlp.edu.ar ▪ https://orcid.org/0000-0001-6979-9103
Waldo Hasperué
whasperue@lidi.info.unlp.edu.ar ▪ https://orcid.org/0000-0002-9950-1563
Laura Lanzarini
laural@lidi.info.unlp.edu.ar ▪ https://orcid.org/0000-0001-7027-7564
III-LIDI, UNLP, ARGENTINA

Recibido: 2020-11-15 ▪ Aceptado: 2020-12-22

RESUMEN

Desde el surgimiento de los Histogramas de Gradientes Orientados en 2005 como el descriptor más utilizado para la detección de peatones, se hna producido numerosas mejoras en el área. Sin embargo, las bases de datos disponibles para el entrenamiento no suelen ser suficientemente representativas, lo que dificulta su uso en un entorno real, diferente al original. Este artículo presenta un protocolo para evaluar la generalización de los modelos de detección de peatones entre diferentes bases de datos. Dicho protocolo consiste en entrenar un modelo con cada uno de los conjuntos de datos o combinaciones de estos, así como evaluar con la base de datos restante. Se analizó la eficacia de los modelos de clasificación de peatones basado en descriptores de Histogramas de Gradientes Orientados y/o Patrones Binarios Locales, y una Máquina de Vectores Soporte como clasificador base. Sucesivamente, también se hizo uso de un modelo convolucional actual (ConvNets) para verificar que los resultados del protocolo son acordes al conjunto de datos y no al modelo. Se evaluó los modelos con las tres bases de datos más utilizadas en el estado del arte: INRIA, Daimler y TUD-Brussels. Los resultados obtenidos muestran que si bien cada conjunto de datos contiene imágenes del mundo real, también contienen sesgos que dificultan que el modelo logre generalizar con otras bases de datos. Los modelos entrenados con dos bases de datos combinadas logran una eficacia ligeramente mejor al evaluar con el tercer conjunto restante frente a los modelos entrenados con un único conjunto de datos, ambos con los clasificadores SVM y ConvNets.

Palabras clave: detección de peatones; ResNet; SVM; transferencia de aprendizaje.

ABSTRACT

Since the surge in popularity of Histogram of Oriented Gradients (HOG) in 2005 as the de facto feature vector for pedestrian detection, there have been many improvements in the detection pipeline that have enabled real-world applications of this technique. Nonetheless, the datasets available for training models have many biases, making it hard to use to detect pedestrians from videos and images obtained from other sources than the datasets.

This article presents a protocol to evaluate how pedestrian models generalize between different datasets. The protocol roughly consists of training a model with each dataset or dataset combination and evaluating with the remaining dataset in each case.

We use the protocol to evaluate the performance of a typical pedestrian classification model based on Histogram of Oriented Gradients and/or Local Binary Pattern features and a Support Vector Machine classifier. Alternatively, we also use a modern ConvNets model, to verify that the results of the protocol are due to the datasets and not the models.

We evaluate the models with the three most used pedestrian classification datasets: INRIA, Daimler, and TUD-Brussels. Our results show that while each dataset presents real-world scenes, there are significant biases in each dataset that prevent models trained on one dataset to generalize to other datasets. Models trained on two fused datasets perform only marginally better on the third dataset than models trained on individual datasets, both for SVM and ConvNet classifiers.

Keywords: *pedestrian detection; ResNet; SVM; transfer learning.*

INTRODUCTION

Within computer vision, pedestrian detection is an important topic that has made many advances recently, fueled by its impact in social and urban safety applications (Yan, *et al.,* 2013; Braun, *et al.,* 2019; Benenson, *et al.,* 2015). Having reached a very high-performance level relative to humans, the research topics have become more specialized, focusing on the detection of partially hidden pedestrians, and blurry or low-resolution images, typical of many security cameras and surveillance systems (Ouyang, *et al.,* 2013).

Pedestrian detection systems, as well as other detection systems, usually employ a pipeline that consists of three base steps. First, a windowing scheme is applied to the input image so that pedestrians can be found at multiple locations in the scene. Secondly, for each window, a sub-image is extracted and evaluated to determine if it contains a pedestrian. For each window, the sub-image is preprocessed and a feature vector is calculated from it. Afterward, the feature vector is fed to a binary classifier that determines if that sub-image contains a pedestrian (positive) or not (negative). The feature vector step is there to provide an alternative representation of the image more suited for classification. Third, and finally, the coordinates of each window with a positive result are mapped back to the original image to find the bounding box of the pedestrians. Given that the same pedestrian may be detected by different windows, a redundancy filter algorithm such as non-maximum suppression must be applied to the resulting bounding boxes.

An important and general assumption in the evaluation of pedestrian detection systems is that the distribution of the images in the testing dataset matches those of the training dataset. Therefore, these systems are typically evaluated using a single dataset or various datasets but in an independent fashion. This assumption makes it difficult to evaluate if a model will generalize properly when applied to a real problem (Benenson, *et al.,* 2015; Hasan, *et al.,* 2020; Cygert, *et al.,* 2020).

There are several reasons why the distribution of the various datasets may vary. First, capture devices have different resolutions, quality, and distance or angle to the pedestrians. Then, there are illumination or climate variations that modify the intensity values of the images. The context of the dataset recording might also differ. This may cause a significant difference in the appearance of the pedestrians and the background scene. For example, some datasets are recorded in cities, while others have trees or highways as background; some datasets are recorded in winter, others in mild weather; some come from surveillance type cameras while others are recorded from inside a moving vehicle (Dalal, *et al.,* 2005; Enzweiler, et al.; 2008, Wojek, *et al.,* 2009). Appropriate preprocessing and feature design can mitigate these issues but currently cannot achieve complete invariance to all (Cygert, *et al.,* 2020).

Given this situation, it is surprising that very few works use a different testing dataset to evaluate their models. To the best of our knowledge, only two articles have tackled this issue. Cao, *et al.,* (2013) uses multiple datasets to generate a model that is more robust to image variations, but they again focus on improving the performance on a single testing dataset and not on the general ability of the model to generalize between datasets and therefore distributions. Benenson, *et al.,* (2015) do analyze the generalization of common pedestrian detection models to new unseen datasets. For this purpose, they employ a full detection task to evaluate the performance. However, since detection is the final goal of such a model, it includes various windowing and redundancy filter techniques which are not entirely relevant to the task of analyzing whether an image is a pedestrian or not. This introduces additional variability factors that difficults the interpretation of the results. Also, they do not consider the effect of fusing two or more datasets to improve the generalization ability of the model. At the same time, new datasets with increasingly more complex tasks are being developed (Shao, *et al.,* 2020), which further increases the necessity for meta-comparisons that include various datasets.

In this work, we propose to evaluate the generalization properties of pedestrian datasets, but for a classification task. That is, we will use pre-cropped images from the datasets, each containing a pedestrian or a background image. In this way remove the variance introduced by the windowing and non-maximum suppression steps (steps 1 and 3 of the pipeline as described above).

We also systematically evaluate the effect of fusing datasets to improve performance. We do this by introducing the notion of generalization matrices, that clearly show the similarity between different datasets as well as their complementarity. In this way, we simplify the decision of choosing the dataset(s) to train a robust pedestrian classification model able to generalize to new contexts.

Since the number of proposed models for pedestrian detection is very large, we focus on the two that represent the most important milestones in pedestrian detection research.

In previous work, we experimented with a traditional model for pedestrian detection (Camele, *et al.,* 2018) so that the results would be easily understood. That model employs a Support Vector Machine (SVM) as a classifier, with either Histogram of Oriented Gradients (HOGs), Local Binary Patterns (LBP) or both combined as features (Dalal, *et al.,* 2005; Gan, *et al.,* 2011; Dollar, *et al.,* 2012; Wang, *et al.,* 2009; Pei, *et al.,* 2014). To provide a guideline that is more relevant for current applications, in this work, we extend those results by evaluating a Convolutional Neural Network (CNN) model, which are currently the state of the art in many image classification problems, including pedestrian classification (Zhang, *et al.,* 2016).

In both cases, we perform experiments with the three most used datasets for pedestrian detection research: INRIA (Dalal, *et al.,* 2005), Daimler (Enzweiler, *et al.,* 2008), and TUD-Brussels (Wojek, *et al.,* 2009).

Section *Methodology* describes the datasets, features, and models employed in the experiments, including the new CNN model. Section *Results and Discussion* presents the results of the experiments and section *Conclusion* the conclusions and future works of this article.

METHODOLOGY

This section presents the three datasets we employ in our experiments: Daimler Mono Pedestrian Detection, INRIA Person, and TUD-Brussels Motionpairs. We also briefly explain the feature vectors we used with the SVM and the CNN architectures tested. Furthermore, we present the experiment protocol carried out to evaluate the generalization of models.

Datasets

INRIA Person

The INRIA Person dataset (Dalal, *et al.,* 2005) (Fig. 1) contains images of people (not necessarily pedestrians) in various situations. It was generated by aggregating and curating various other image datasets. The sizes of the images have been normalized to 64×128 pixels and are full color. The training set has 3.030 and 34.892 positive and negative samples, while the testing set has 1.028 and 453 respectively.

Daimler Mono Pedestrian Detection

The dataset Daimler Mono Pedestrian Detection (Enzweiler, *et al.,* 2008) (Fig. 2) focuses on pedestrian detection from monocular cameras. It contains grayscale images recorded from a moving car. The training set contains pedestrian 15.560 images with a 48×96 pixels resolution. The testing set contains 6.744 full-size images without pedestrians to extract negative images. We generated 32.000 negative images of the same size as the training images by cropping random sub-images. Afterward, 90% of all the images were used for training and the rest for evaluation.

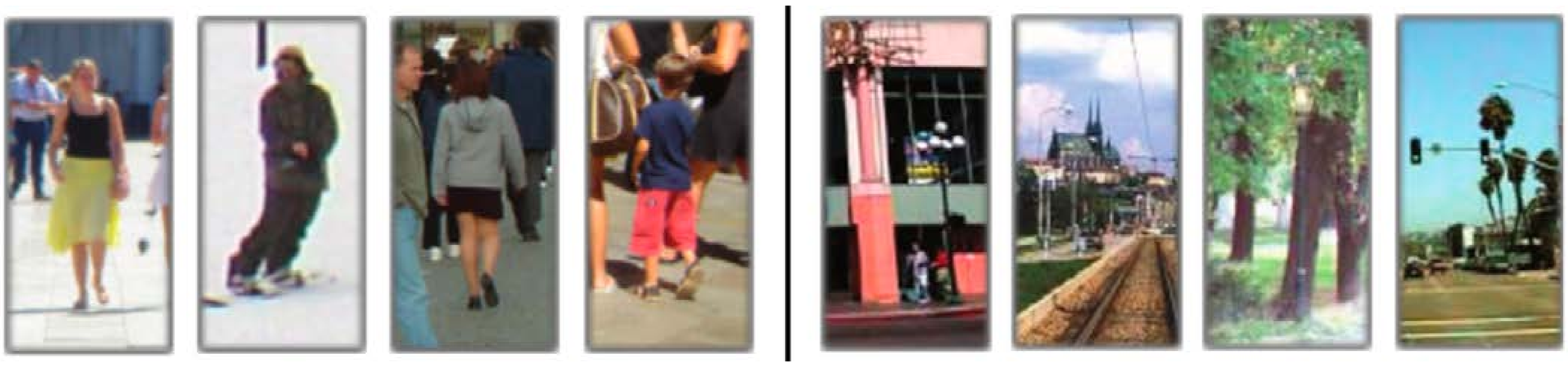

Figure 1. Positive samples (left) and negative (right) of the INRIA Person dataset (Dalal, *et al.,* 2005).

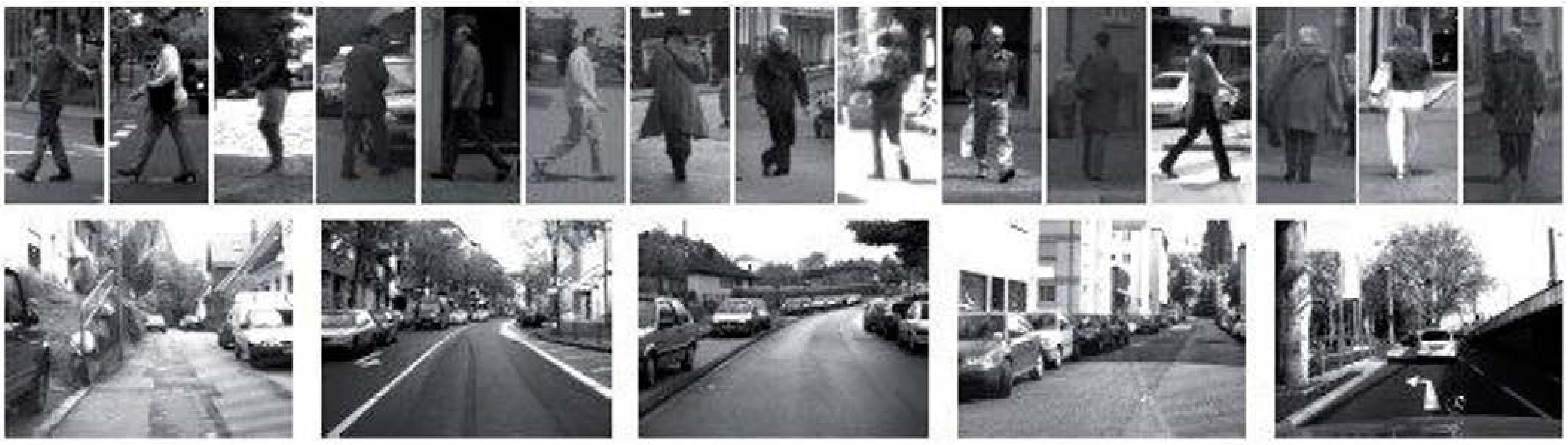

Figure 2. Positive samples (up) and negative (down) of the Daimler Mono Pedestrian Detection dataset (Enzweiler, *et al.,* 2008).

TUD-Brussels Motionpairs

The dataset TUD-Brussels Motionpairs (Wojek, *et al.,* 2009) (Fig. 3) is smaller, but also focused on pedestrian detection from a moving car, but with color images of 640×480 pixels resolution. It contains 551 annotated pedestrians for the training set and 311 for the testing set.

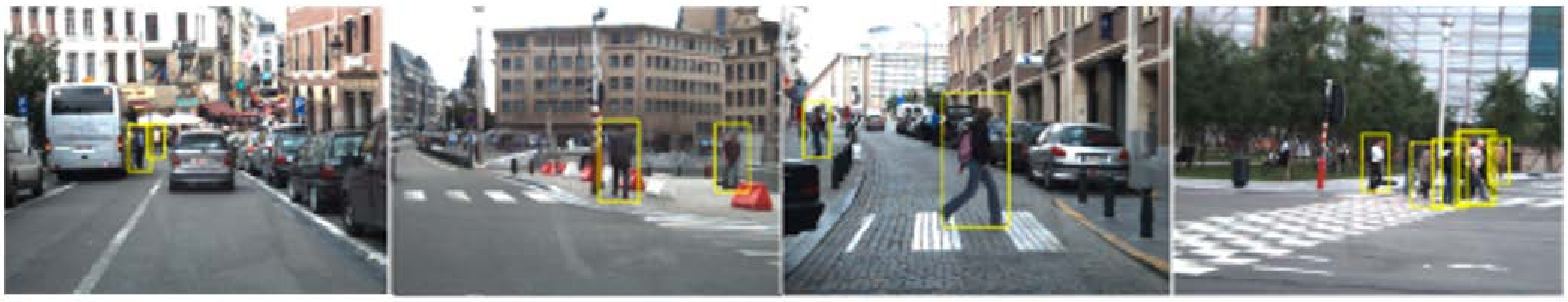

Figure 3. Samples of the dataset TUD-Brussels Motionpairs (Wojek, *et al.,* 2009).

Table 1. Distribution of samples in the datasets used in the experiments.

Dataset	Training			Testing		Color	Context
	Positive	Negative		Positive	Negative		
Daimler	14,000	30,000		1,560	2,000	No	Pedestrian
INRIA	3,030	34,879		1,028	453	Yes	Person
TUD-Brussels	982	9,064		110	1,010	Yes	Pedestrian

Features

Histogram of Oriented Gradients (HOGs)

HOG (Dalal, *et al.,* 2005) is a widely employed feature that has served many object classification tasks. The first step of its calculation is to obtain the gradient of the input image. Given that the gradient tends to be larger in magnitude near an object's border, HOGs have been used to detect object shapes. The gradient is represented in polar form, and the angles are roughly perpendicular to the object's boundary or salient characteristics. Since individual gradient angles are too variable to use effectively, the image is divided into cells and the histogram of gradients is calculated for each one. The histogram bins are made up by a quantization of the angles of the gradients, usually into 8 bins. Each bin, therefore, corresponds to a range of angles, and each gradient value contributes to a certain bin proportionally to its magnitude. Afterward, the histograms are grouped by blocks, usually overlapping to capture more structure, and are normalized. Finally, the blocks are concatenated into a 1D feature vector. The last row of figure 4 shows a visualization of the HOG cells for positive and negative images of each dataset.

Local Binary Patterns (LBP)

LBP (Mu, *et al.,* 2008) is a texture-based descriptor. They are calculated by dividing the images into cells, and for each pixel of a cell, the neighborhood gradient of the pixel is calculated by concatenating n neighboring pixels. Afterward, a histogram with 2^n bins is calculated for the entire cell, where each pixel increments with value 1 the corresponding bin. The cell histograms are concatenated in the same way as HOGs to form the final feature.

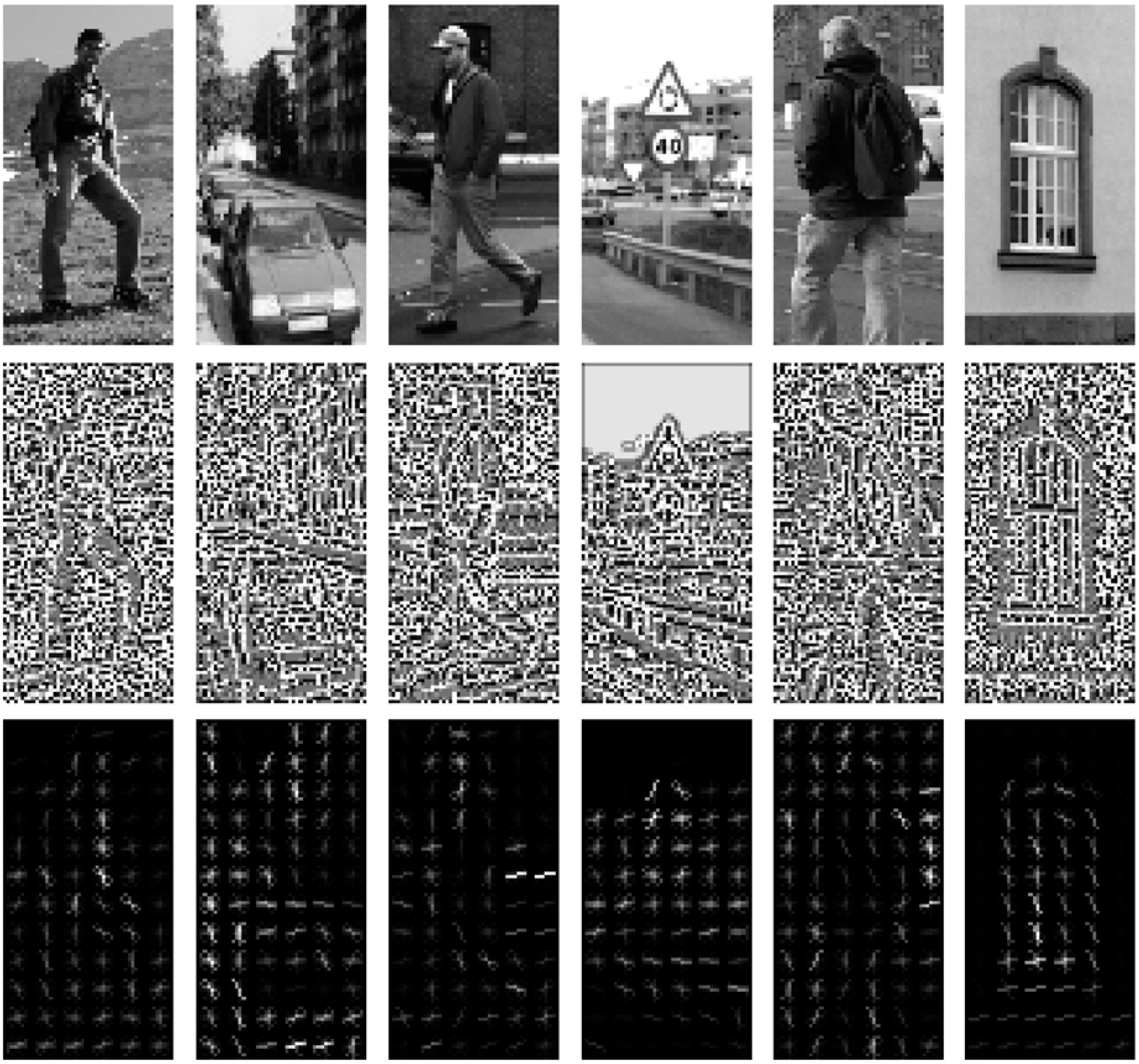

Figure 4. Positive and negative images of each dataset (row 1), with corresponding. LBPs and HOGs (rows 2 and 3), for datasets INRIA (columns 1 and 2), Daimler (columns 3 and 4), and Brussels (columns 5 and 6).

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are the current state of the art in image classification and other machine learning fields. ResNet (He, *et al.,* 2016) is a CNN architecture that features residual connections, bottlenecks, ReLu activation functions, global average pooling, strided convolutions instead of max pooling, and batch normalized layers. While more recent and powerful CNN architectures exist, we chose ResNets because they are widely known, represent many important milestones in the evolution of CNNs, and have been employed extensively in pedestrian classification (Zhang, *et al.,* 2016). Therefore, we propose ResNets as an adequate modern match to the traditional SVN+HOG model for pedestrian classification.

Preprocessing

Given that images have different sizes in different datasets, we resize all images to 48×96 pixels as a standard resolution so that results could be comparable and images could be evaluated by any model. Since the original images of all datasets had an aspect ratio of 1:2, resizing did not introduce any distortion.

Furthermore, we converted INRIA and TUD-Brussels images to grayscale so that the HOG and LBP calculation is the same for all datasets since Daimler had grayscale images. We normalized the range of the images to (0..1). We did not perform histogram normalization or another such additional preprocessing to preserve the original data distribution, since we are interested in inter-dataset generalization.

The HOG's cell size was 8×8 pixels with 2×2 blocks and L2 normalization following Dalal, *et al.,* (2005). LBPs used n= 8 sampling points in a 1-pixel radio, also following (Wang, *et al.,* 2009).

Experiment protocol

We present the protocol to generate the generalization matrices to evaluate the generalizability of a model between datasets. First, we train three versions of the model, one for each of the training sets of each dataset. Then, we evaluate each model with each test set of the three datasets. As a performance metric, we employed the f-score, to combine the precision and recall in a single measure and simplify the results.

In this way, we obtained a generalization matrix D of size 3×3 between the datasets where $D_{i,j}$ indicates the f-score of the model when trained by dataset i and evaluated with dataset j. D facilitates observing the similarity between the distributions of the images of the different datasets. Note that this matrix is not symmetric; it is possible that, for some models and datasets, $D_{1,2}$= 1 and $D_{2,1}$= 0.

Classifiers

For the HOG and LBP features, we used the linear Support Vector Machine, since it is the most used method in the literature. It is also very simple and efficient to evaluate, which is ideal for a detection system that has to run the classifier multiple times per image. Following Dalal, *et al.,* (2005) we set C= 0.1, use a linear Kernel with L2 regularization. The model was trained with the SMO method from the library *liblinear* with the dual formulation.

The CNN model was ResNet50 (He, *et al.,* 2016) (50 layers deep), trained from scratch, using RMSprop and a learning rate of 0.001. The CNN was trained and evaluated with raw images without any further preprocessing or data augmentation.

In both cases, we assigned balanced class weights to the classifiers to deal with the unbalanced nature of the datasets.

RESULTS AND DISCUSSION

Generalization learning

Table 2 shows the generalization matrices between the three datasets for all models. Values in the diagonal of the matrix are generally much greater than off-diagonal entries since models trained with one dataset do not generalize well one to another.

Table 2. Generalization matrices between the three datasets.
Entries show f-scores. Rows: Training set. Columns: Test set.
Abbreviations I = INRIA, D = Daimler, B = TUD-Brussels.

a). SVM+HOG

	I	D	B
I	0.619	0.297	0.064
D	0.841	0.984	0.627
B	0.335	0.396	0.803

b). SVM+LBP

	I	D	B
I	0.176	0.078	0.097
D	0.693	0.887	0.278
B	0.167	0.151	0.218

c). SVM+HOG+LBP

	I	D	B
I	0.611	0.351	0.210
D	0.837	0.979	0.556
B	0.376	0.369	0.748

d). ResNet

	I	D	B
I	0.950	0.885	0.778
D	0.358	0.996	0.752
B	0.285	0.813	0.980

There is a clear difference in the structure of the matrices of the ResNet and the three SVM-based models. ResNets trained with INRIA have the best performance overall, while SVM models show better generalization with Daimler.

Daimler appears to have the best capacity for generalization; however, that could be simply due to having more examples than the other datasets and not due to some inherent property of the distribution of the images in the dataset.

For the SVM classifiers, the best results generally seem to come from the HOG descriptor. LBPs alone behave poorly in most cases. Adding LBPs to HOGs does not improve performance in general. However, concatenating LBPs and HOGs does increase the performance when training with the INRIA dataset. This suggests the possibility that given that INRIA is a person dataset and has a dataset bias (Azulay, *et al.,* 2019; Hasan, *et al.,* 2020), the HOG descriptor captures the shape with this bias; since LBPs give more emphasis to local texture, they may help overcome that bias.

The ResNet model has the best results when training and testing with the same dataset, but does worse than SMV+HOG in many cases. This may be due to deep CNNs having more *dataset bias* than other models (Azulay, *et al.,* 2019; Hasan, *et al.,* 2020).

Generalization learning with multiple datasets

A natural extension of the previous experiment consists of merging the training set of two datasets to train the model. Test sets are not merged but are evaluated in the same way as in the previous experiment. In this way, we can evaluate the complementarity of the datasets. Table 3 shows the resulting generalization matrices. Note that for the SVM+HOG/LBP models, we could not train the SVM with all three datasets combined due to memory constraints.

Table 3. Generalization matrices training with merged datasets. Entries represent f-scores for the testing set of the given dataset. Rows: Training sets. Columns: Test set. Abbreviations: I = INRIA, D = Daimler, B =TUD-Brussels.

a). SVM+HOG

	I	D	B
I+D	0.837	0.960	0.578
D+B	0.816	0.976	0.850
B+I	0.747	0.645	0.658

b). SVM+LBP

	I	D	B
I+D	0.444	0.709	0.215
D+B	0.581	0.829	0.303
B+I	0.273	0.149	0.176

c). SVM+HOG+LBP

	I	D	B
I+D	0.844	0.946	0.472
D+B	0.789	0.965	0.766
B+I	0.697	0.562	0.497

d). ResNet

	I	D	B
I+D	0.965	0.991	0.708
D+B	0.246	1.00	0.969
B+I	0.860	0.496	0.874
I+D+B	0.964	0.999	0.938

To see more clearly the trend of increased accuracy when merging datasets, figure 5 sums up the results obtained in all experiments using the HOG+SVM and ResNet models only, which were the ones with the best results. The figure also shows that combining the three datasets as a training set is the most robust choice in all cases.

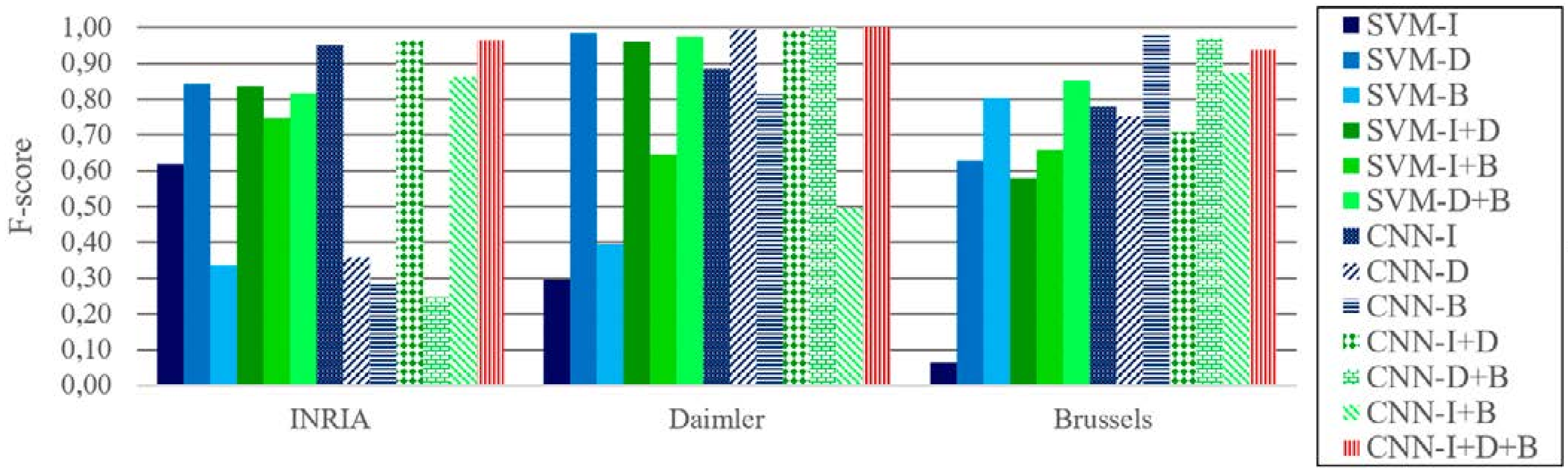

Figure 5. Comparison of the f-scores of each dataset combination when using the HOG+SVM model (solid bars) and ResNet (patterned bars). Each bar represents a training set. Each block of bars represents a test set. Best viewed in color.

By training with a combination of datasets the f-score increases in many cases relative to training with a single one. There is, however, a dataset size effect we need to consider since more data helps to regularize the model. On the other hand, merging two large datasets does not provide as much advantage to either as merging a small dataset with a large one. For example, we can see that TUD-Brussels does not contribute many examples to INRIA+TUD+-

Brussels, but the f-score of its test set is increased for the SVM+HOG case nonetheless. This suggests that including many small datasets can have a large impact on performance relative to the number of training examples. This could be due to the greater variability of including a different dataset, which may act as a stronger regularizer.

In many cases adding new data does not improve performance, such as in INRIA+Daimler for the INRIA test set and SVM+HOG. On the other hand, only in very few cases adding a dataset hinders performance, such as Brussels+INRIA for the test set of INRIA and ResNet. Therefore, the results suggest including as many datasets as possible, especially if these are small and different from each other.

CONCLUSIONS

In this article, we presented a protocol to evaluate how well can models generalize between pedestrian datasets. We evaluated three datasets, Daimler, TUD-Brussels, and INRIA, with both traditional and state of the art models, based on SVM+HOG/LBP and CNNs, respectively. Our results show that in general models trained on one dataset do not perform well when testing with another. The Daimler dataset provides the best transferability overall, which is not surprising given it has the most samples and variability. Also, merging several datasets improves the performance of the model in various cases, especially when also testing generalization with various datasets.

While the techniques presented are simple and straightforward, the results show that these types of studies can help to identify problems in both datasets and models. Therefore, this points to a need to perform more inter-dataset generalization studies in this field, as well as possibly others. In future work, we will include more datasets in the comparison and quantify the effect of their relative sizes in addition to their variability. We also propose to extend the experiments into the detection domain with the same setup and models to quantify the effect of the windowing scheme and non-maximum suppression steps independently.

REFERENCIAS

Azulay, A. y Weiss, Y. (2019). Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research, 20,* 1-25.

Cygert, S. y Czyżewski, A. (2020). Toward Robust Pedestrian Detection With Data Augmentation. *IEEE Access, 8,* 136674-136683.

Benenson, R., Omran, M., Hosang, J., y Schiele, B. (2015). Ten years of pedestrian detection, what have we learned? In: *Computer Vision - ECCV 2014 Workshops.* (pp. 613-627). Springer International Publishing.

Camele, G., Quiroga, F., Ronchetti, F., Hasperué, W., y Lanzarini, L.C. (2018). Transferencia de aprendizaje para la detección de peatones. In: *XXIV Congreso Argentino de Ciencias de la Computación, CACIC 2018.* La Plata. (pp. 52-61). Red de Universidades con Carreras en

Informática (RedUNCI).

Cao, X., Wang, Z., Yan, P., y Li, X. (2013). Transfer learning for pedestrian detection. *Neurocomputing, 100,* 51-57, special issue: Behaviours in video.

Dalal, N. y Triggs, B. (2005). Histograms of oriented gradients for human detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2005. CVPR 2005 (pp. 886-893).

Dollar, P., Wojek, C., Schiele, B., y Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*(4), 743-761, doi: 10.1109/TPAMI.2011.155

Enzweiler, M. y Gavrila, D.M. (2008). Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 31*(12), 2179-2195. doi: 10.1109/TPAMI.2008.260.

Gan, G. y Cheng, J. (2011). Pedestrian detection based on hog-lbp feature. *2011 Seventh International Conference on Computational Intelligence and Security* (pp. 1184-1187). doi:10.1109/CIS.2011.262

He, K., Zhang, X., Ren, S., y Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV,* (pp. 770-778). doi: 10.1109/CVPR.2016.90

Mu, Y., Yan, S., Liu, Y., Huang, T., y Zhou, B. (2008). Discriminative local binary patterns for human detection in personal album. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008* (pp. 1-8). doi:10.1109/CVPR.2008.4587800

Ouyang, W., y Wang, X. (2013). Single-pedestrian detection aided by multi-pedestrian detection. *IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013* (pp. 3198-3205). doi:10.1109/CVPR.2013.411

Pei, W.J., Zhang, Y.L., Zhang, Y., y Zheng, C. H. (2014). Pedestrian detection based on HOG and LBP. In: *Intelligent Computing Theory.* (pp. 715-720). Springer International Publishing.

Wang, X., Han, T.X., Yan, S. (2009). An hog-lbp human detector with partial occlusion handling. In: *IEEE 12th International Conference on Computer Vision, 2009* (pp. 32-39).

Wojek, C., Walk, S., y Schiele, B. (2009). Multi-cue onboard pedestrian detection. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, 2009* (pp. 794-801), doi:10.1109/CVPRW. 2009.5206638.

Yan, J., Zhang, X., Lei, Z., Liao, S., y Li, S. Z. (2013). Robust multi-resolution pedestrian detection in traffic scenes. *2013 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3033-3040). Portland, OR, 2013. doi: 10.1109/CVPR.2013.390.

Braun, M., Krebs, S., Flohr, F., y Gavrila, D. M. (2019). EuroCity persons: A novel benchmark for person detection in traffic scenes. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1844-1861.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., y Sun, J. (2018). Crowdhuman: A benchmark for detecting humans in a crowd. arXiv preprint arXiv:1805.00123.

Zhang, L., Lin, L., Liang, X., y He, K. (2016). Is faster r-cnn doing well for pedestrian detection?. *Computer Vision and Pattern Recognition. ECCV 2016* (pp. 443-457). Springer International Publishing.

Hasan, I., Liao, S., Li, J., Akram, S. U., y Shao, L. (2020). Pedestrian Detection: The Elephant In The Room. arXiv preprint arXiv:2003.08799.

Zhang, S., Benenson, R., y Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3213-3221).