Spatiotemporal Stacked Sequential Learning for Pedestrian Detection

Alejandro González[1,2] Sebastian Ramos[1,2] David Vázquez[1]
Antonio M. López[1,2] Jaume Amores[1,3]

[1] Computer Vision Center, Barcelona [2] Universitat Autònoma de Barcelona
[3] United Technologies Research Center

{agalzate,sramosp,dvazquez,antonio,jaume}@cvc.uab.es

Abstract

Pedestrian classifiers decide which image windows contain a pedestrian. In practice, such classifiers provide a relatively high response at neighbor windows overlapping a pedestrian, while the responses around potential false positives are expected to be lower. An analogous reasoning applies for image sequences. If there is a pedestrian located within a frame, the same pedestrian is expected to appear close to the same location in neighbor frames. Therefore, such a location has chances of receiving high classification scores during several frames, while false positives are expected to be more spurious. In this paper we propose to exploit such correlations for improving the accuracy of base pedestrian classifiers. In particular, we propose to use two-stage classifiers which not only rely on the image descriptors required by the base classifiers but also on the response of such base classifiers in a given spatiotemporal neighborhood. More specifically, we train pedestrian classifiers using a stacked sequential learning (SSL) paradigm. We use a new pedestrian dataset we have acquired from a car to evaluate our proposal at different frame rates. We also test on a well known dataset: Caltech. The obtained results show that our SSL proposal boosts detection accuracy significantly with a minimal impact on the computational cost. Interestingly, SSL improves more the accuracy at the most dangerous situations, i.e. *when a pedestrian is close to the camera.*

1. Introduction

Localizing humans in images is key for applications such as video surveillance, avoiding pedestrian-to-vehicle collisions, collecting statistics of players or athletes in sport videos, etc. Developing a reliable vision-based pedestrian detector is a very challenging task with more than a decade of history by now. As a result, a plethora of features, models, and learning algorithms, have been proposed to develop the pedestrian classifiers which are at the core of pedestrian detectors [15].

The research for boosting the accuracy of pedestrian classifiers has followed different lines. Some authors have researched image descriptors well-suited for pedestrians (*e.g.*, HOG [7], HOG+LBP [31], HOG+CSS+HOF [30], OppHOG [26], Haar+EOH [16], Integral Channels [10], Macrofeatures [22]), others have researched different image modalities (*e.g.*, appearance + motion [32], appearance+depth+motion [12]), others have focused on the pedestrian model (*e.g.*, deformable multi-component part-based models [14, 25, 19], multi-resolution [24, 2]), others on the classification architecture (*e.g.*, HOG-SVM/LRF-MLP cascades [23], Haar + EOH-AdaBoost cascades with meta-stages [4], random forest of HOG+LBP-SVMs [21]), and others in the process of collecting good samples for training (*e.g.*, generative approach [13], active learning [1], virtual-world data with domain adaptation [28]).

The outcome of each of the above mentioned proposals is a pedestrian classifier, termed here as *base classifier*, which determines if a given image window contains a pedestrian or background. In practice, such classifiers provide a relatively high response at neighbor windows overlapping a pedestrian, while the responses around potential false positives are expected to be lower. Note that, in fact, non-maximum suppression (NMS) is usually performed as last detection stage in order to reduce multiple detections arising from the same pedestrian to a single one. An analogous reasoning applies for image sequences. If there is a pedestrian located within a frame, the same pedestrian is expected to appear close to the same location in neighbor frames. Therefore, such a location has chances of receiving high classification scores during several frames, while false positives are expected to be more spurious. In fact, this may allow removing such undesired spurious by the use of a tracker.

In this paper we propose to exploit such expected *response correlations* for improving the accuracy of the classification stage itself. In other words, instead of only exploiting spatiotemporal coherence by means of general post-classification stages like NMS and tracking, we propose to add such a type of reasoning in the classification stage itself as well. In particular, we propose to use a two-stage classification strategy which not only rely on the image descriptors required by the base classifiers, but also on the response of the own base classifiers in a given spatiotemporal neighborhood. More specifically, we train pedestrian classifiers using a stacked sequential learning (SSL) paradigm [5].

Temporal SSL involves the analysis of window volumes. The different types of temporal volumes can be potentially useful for different applications depending on the motion of the camera and the targets of interest, as well as the working frame rate and the targets size. In this paper, we are specially interested in on-board pedestrian detection within urban scenarios. Therefore, camera and targets are in movement. Accordingly, in this paper we test our SSL approach for a fixed neighborhood (*i.e.*, fixed spatial window coordinates across frames) and for an scheme relying on an ego-motion compensation approximation (*i.e.*, varying spatial window coordinates across frames). Moreover, in order to assess the dependency of the results with respect to the frame rate, we acquired our own pedestrian dataset at 30fps by normal driving in an urban scenario. This new dataset is used as main guide for our experiments, but we also complement our study with other challenging dataset publicly available: Caltech.

In this paper we start by using a competitive baseline in pedestrian detection [11], namely a holistic base classifier based on HOG+LBP features and linear SVM. Note that HOG/linear-SVM is the core of more sophisticated pedestrian detectors as the popular deformable part-based model (DPM) [14]. Moreover, HOG with LBP are also used as base descriptors of multi-modal multi-view pedestrian models [12], and HOG+LBP/linear-SVM has been used for classifiers with occlusion handling [31, 20], as well as for acting as node experts in random forest ensembles [21]. In addition, it has recently been shown that HOG+LBP/linear-SVM approaches are well suited for domain adaptation [28]. Altogether, we think that HOG+LBP/linear-SVM is a proper baseline to start assessing our proposal. Moreover we have extended this baseline with the HOF [30] motion descriptor that complements the appearance and texture features of the baseline.

Overall, the obtained results show that our spatiotemporal SSL proposal boosts detection accuracy significantly. Especially, when the pedestrians are close to the camera, *i.e*. in the most critical situations. Therefore, encouraging to augment the study for other pedestrian base classifiers as well as other object categories.

The rest of the paper is organized as follows. In Sect. 2 we review some works related to our proposal. Section 3 briefly introduces the SSL paradigm. In Sect. 4 we develop our proposal. Section 5 presents the experiments carried out to assess our spatiotemporal SSL, and discuss the obtained results. Finally, Sect. 6 draws our main conclusions.

2. Related work

The use of motion patterns as image descriptors was already proposed as an extension of spatial Haar-like filters for video surveillance applications (static zenital camera) [29, 6, 17] and for detecting human visual events [18]. In these cases, original spatial Haar-like filters were extended with a temporal dimension. Popular HOG descriptor was also extended to encode temporal information for detecting humans [8], in this case using optical flow to compensate motion. In the same spirit the histograms of flow (HOF) were also introduced for detecting pedestrians [30]. In all cases motion information was complemented with appearance information (*i.e.*, Haar/HOG for luminance and/or color channels).

In contrast with these approaches, our proposal does not involve to compute new temporal image descriptors as new features for the classification process. As we will see, we use the responses of a given base classifier in neighbor frames as new features for our SSL classifier. In fact, our proposal can also be applied to base classifiers that already incorporate motion features. Therefore, the reviewed literature and our proposal are complementary strategies.

Focusing on single frames, it has been recently shown how pedestrian detection accuracy can be boosted by analyzing the image area surrounding potential pedestrian detections. In particular, [9, 3] follow an iterative process that uses contextual features of several orders (*e.g.*, involving co-occurences) for progressively enhancing the response of base classifiers for true pedestrians and lowering it for hallucinatory ones. Our SSL proposal does not require new image descriptors of pedestrian surroundings and is not iterative, which makes it inherently faster. Moreover, we treat equally spatial and temporal response correlations, *i.e.*, under the SSL paradigm, giving rise to a more straightforward method.

Finally, we would like to clarify that our SSL proposal is not a substitute for NMS and tracking post-classification stages. What we expect is to allow these stages to produce more accurate results by increasing the accuracy of the classification stage. For instance, tracking must be used for predicting pedestrian intentions [27], thus, if less false positives reach the tracker, we can reasonably expect to obtain more reliable pedestrian trajectories and so guessing intentions in the very short time this information is required (*i.e.*, around a quarter of second before a potential collision).

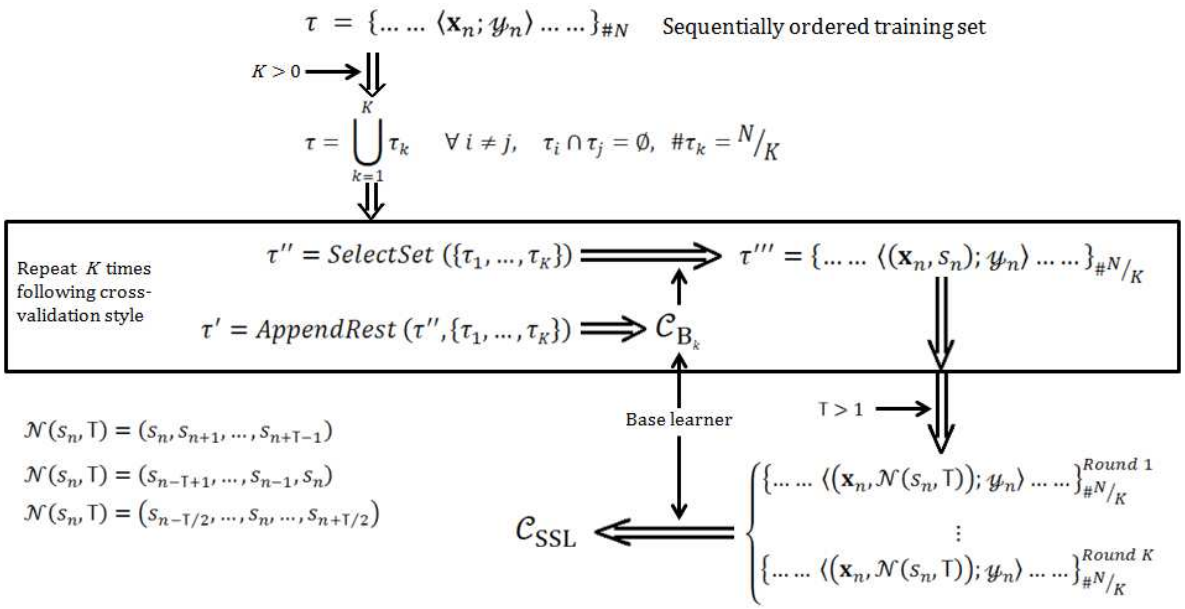

Figure 1. SSL learning. See main text in Sect. 3 for details.

3. Stacked sequential learning (SSL)

Stacked sequential learning (SSL) was introduced by Cohen *et al*. [5] with the aim of improving base classifiers when the data to be processed has some sort of sequential order. In particular, given a data sample to be classified, the core intuition is to consider not only the features describing the sample but also the response of the base classifier in its neighbor samples. Figure 1 summarizes the SSL learning process that we explain in more detail in the rest of this section.

Let τ be an ordered training sequence of cardinality N. The SSL approach involves to select a sub-sequence for training a base classifier, $\mathcal{C}_B$, and the rest to apply $\mathcal{C}_B$ and so training the SSL classifier, $\mathcal{C}_{SSL}$. If this is done once, then the final classifier $\mathcal{C}_{SSL}$ would be trained with less than N samples. Thus, to avoid this, it is followed a cross-validation style were τ is divided in $K > 0$ disjoint sub-sequences, $\tau = \cup_{k=1}^{K} \tau_k \wedge i \neq j \Rightarrow \tau_i \cap \tau_j = \emptyset$, and K rounds are performed by using a different subset each round to test the $\mathcal{C}_{B_k}$ and the rest of subsets for training this $\mathcal{C}_{B_k}$. At the end of the process, joining the K sub-sequences processed by the corresponding $\mathcal{C}_{B_k}$, we can have N *augmented* training samples for learning $\mathcal{C}_{SSL}$. $k = 1$ means to train the C_B and C_SSL on the same training set, without actually doing partitions.

Let us explain what means *augmented* training samples. The elements of τ, *i.e.*, the initial training samples, are of the form $< \mathbf{x_n}; y_n >$, where $\mathbf{x_n}$ is a vector of features with associated label y_n. Therefore, the elements of each sub-sequence τ_k are of the same form. As we have mentioned before, during each round k of the cross-validation-style process, a sub-sequence τ'' is selected among $\{\tau_1, \ldots, \tau_K\}$, while the rest are appended together to form a sub-sequence τ'. From τ' it is learned $\mathcal{C}_{B_k}$ and applied to τ'' to obtain a new τ'''. The elements of τ''' are of the form $< (\mathbf{x_n}, s_n); y_n >$, where we have augmented the feature $\mathbf{x_n}$ with the classifier score $s_n = \mathcal{C}_{B_k}(\mathbf{x_n})$. Therefore, after the K rounds, we have a training set of N samples of the form $< (\mathbf{x_n}, s_n); y_n >$. It is at this point when we can introduce the concept of neighbor scores into the learning process. In particular, the final training samples are of the form $< (\mathbf{x_n}, \mathcal{N}(s_n, T)); y_n >$, where $\mathcal{N}(s_n, T)$ denotes a neighborhood of size $T > 1$ anchored to the sample n. For instance, $\mathcal{N}(s_n, T) = (s_{n-T+1}, \ldots, s_{n-1}, s_n)$ is a *past* neighborhood, $\mathcal{N}(s_n, T) = (s_n, s_{n+1}, \ldots, s_{n+T-1})$ is a *future* neighborhood, and $\mathcal{N}(s_n, T) = (s_{n-\frac{T}{2}}, \ldots, s_n, \ldots, s_{n+\frac{T}{2}})$ is a *centered* neighborhood, which are analogous concepts to the ones of filtering, extrapolation and smoothing, resp., used in the classical tracking nomenclature.

4. SSL for pedestrian detection

In this section, without loosing generality, we will assume the use of the *past neighborhood* (Sect. 3) to illustrate and explain our SSL approach. From the viewpoint of the processing of image sequences, this means to use previous images to do detection in the current one (*i.e.*, in the last one acquired when processing directly from a camera). Actually there is no need to save the previous images. The detection scores of the neighbouring windows, that were al-

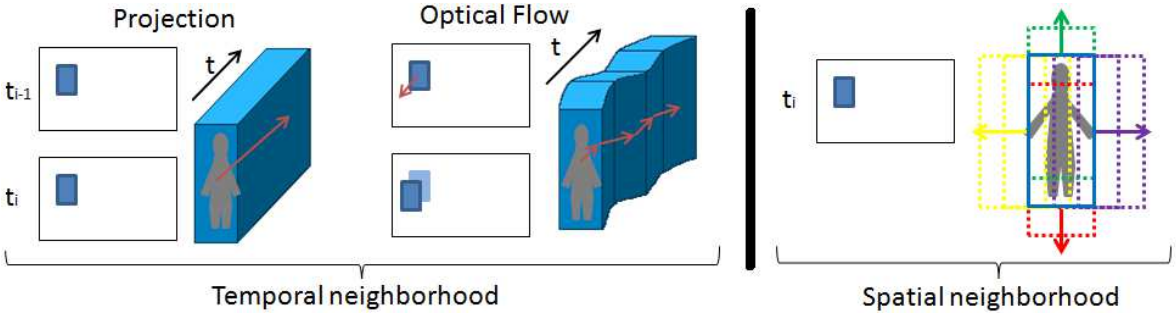

Figure 2. Different types of neighborhood for SSL. See main text in Sect. 4.1 for details.

ready computed, are enough to compute the current SSL descriptor making the computation of SSL very computational efficient.

4.1. Spatiotemporal neighborhoods for SSL

For object detection in general and for pedestrian detection in particular, applying SSL starts by defining which are the neighbors of a given window under analysis. In learning time, such a window will correspond either to the bounding box of a labeled pedestrian or to a rectangular chunk of the background. In operation time (*i.e.*, testing), such a window will correspond to a candidate generated by a pyramidal sliding window scheme or any other candidate selection method. In this paper we assume the processing of image sequences and, consequently, we propose the use of a spatiotemporal neighborhood.

Temporal SSL involves the analysis of window volumes. Therefore, there are several possibilities to consider (see Fig. 2). Let us term as W_f the set of coordinates defining an image window in frame f, and $\mathbf{V}_f = \text{vol}(\cup_{t=0}^{T-1} W_{f-t})$ the window volume defined by a temporal neighbor of T frames. The simplest volume is obtained by assuming fixed locations across frames, which we term as *projection* approach. In other words, $W_f = W_{f-1} = ... = W_{f-(T-1)}$. Another possibility consists in building volumes taking into account motion information. For instance, $W_f = W_{f-1} + t_{OF(W_{f-1})}$, where $t_{OF(W_{f-1})}$ is a 2D translation defined by considering the *optical flow* contained in W_{f-1}, and $'+'$ stands for summation to all coordinates defining W_{f-1}.

Spatial SSL involves the analysis of windows spatially overlapping the window of interest (see Fig. 2). For instance, we can fix a 2D displacement $\Delta = (\delta_x, \delta_y)$ and n_x displacements in the x axis, to the left and to the right, an analogously for the y axis given a n_y number of up and down displacements.

Our proposal combines both ideas, *i.e.*, the temporal volumes and the spatial overlapping windows, in order to define the spatiotemporal neighborhood required by SSL (Sect. 3).

4.2. SSL training

As usual, we assume an image sequence with labeled pedestrians (*i.e.*, using bounding boxes) for training. Negative samples for training are obtained by random sampling of the same images, of course, these samples cannot highly overlap labeled pedestrians. The cross-validation-style rounds of SSL (Sect. 3) are performer with respect to the images of the sequence, not with respected to the set of labeled pedestrians and negative samples as it may suggest the straightforward application of SSL (note that pedestrian/negative labels are for individual windows not for full images). Moreover, as we have seen in Sect. 4.1, the neighborhood relationship is not only temporal but spatial too. The training process is divided in two stages. First, we train the auxiliary classifiers ($\mathcal{C}_{B_k}$) as usual using three bootstraping rounds. Then we train the SSL classifier (using final $\mathcal{C}_{B_k}$ as auxiliary), again we run three bootstrapping rounds for obtaining the final classifier ($\mathcal{C}_{SSL}$).

Using the full training dataset, we also assume the training of a base classifier $\mathcal{C}_B$. Another possibility is to understand the different $\mathcal{C}_{B_k}$ as the result of a bagging procedure and ensemble them to obtain $\mathcal{C}_B$. Without loosing generality, in this paper we have focused on the former approach.

4.3. SSL detector

The proposed pedestrian detection pipeline is shown in Fig. 3. As we can see there are two main stages. The first stage basically consists in a classical pedestrian detection method relying on the learned base classifier $\mathcal{C}_B$. In Fig. 3 we have illustrated the idea for a pyramidal sliding window approach, but using other candidate selection approaches is also possible. Detections at this stage are just considered as potential ones. Then, the second stage applies the spatiotemporal SSL classifier, $\mathcal{C}_{SSL}$, to such potential detections in order to reject or keep them as final detections.

There are some details worth to mention. First, the usual non-maximum suppression (NMS) step included in pedestrian detectors is not performed for the output of the first stage, but it is done for the output of the second stage. Second, for ensuring that true pedestrians reach the second

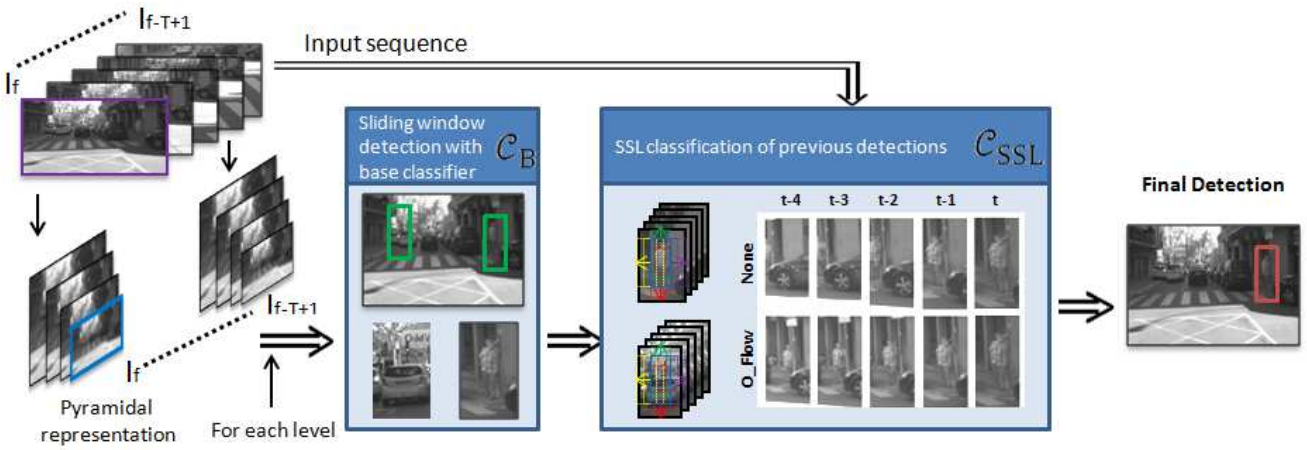

Figure 3. Two-stage pedestrian detection based on SSL. See main text in Sect. 4.3 for details.

stage, we apply a threshold on $\mathcal{C}_B$ such that it guarantees a very high detection rate even having a very high rate of false positives. In our experiments this usually implies that while the $\mathcal{C}_B$ processes hundred of thousands windows (for pyramidal sliding window), $\mathcal{C}_{SSL}$ only process a few thousands. Third, although in Fig. 3 we show pyramids of images for a temporal neighborhood of T frames, what we actually keep from frame to frame are the already computed features, so that we compute them only once. However, this depends on the type of temporal neighborhood we use (Sect. 4.1). For instance, using projection style no feature are needed to keep (*i.e.*, keeping the classification scores is enough). However, if we use optical flow we may need to compute features in previous frames if the window under consideration does not map to a location where they were already computed.

5. Experimental results

Protocol. As evaluation methodology we follow the defacto Caltech standard for pedestrian detection [11], *i.e.* we plot curves of false positives per image (FPPI) *vs.* miss rate. The miss rate average in the range of 10^{-2} to 10^{0} FPPI is taken as indicative of each detector accuracy, *i.e.* the lower the better. Moreover, during testing we consider three different subset based on the pedestrian height. *Near* subset include pedestrians with height equal or higher than 75 pixels, *medium* subset include pedestrian between 50 and 75 pixel height. Finally we group the two previous subset in the *reasonable* subset (height $>=$ 50 pixels).

Our own dataset (OurDS). Since the temporal axis is important for the SSL classifier, we acquired our own dataset to be sure we have stable 30 fps sequences. The sequences were acquired on-board under normal urban driving conditions. The images are monochrome and of 480×960 pixels. We used a 4mm focal length lens, so providing a wide field of view. We drove during 30 minutes approximately, giving rise to a sequence of around 60,000 frames. Then, using steps of 10 frames we annotated all the pedestrians. This turns out in 7,900 annotated pedestrians, 5,400 reasonable and non occluded. We have divided the video sequence into three sequential parts, the first one for training, the last one for testing, in the middle we have leaved a gap for avoiding testing and training with the same persons. Overall we train with 3,600 reasonable pedestrians, and test on 1,300 reasonable ones.

Caltech dataset. We have also used other popular dataset acquired on-board. The Caltech dataset [11], which contain 3,700 reasonable pedestrians for training.

Base detectors. For the experiments presented in this section we use our own implementation of HOG and LBP features, which provides significant better results than the one proposed in [31], *i.e.*, removing the occlusion handling reasoning. Moreover, using TV-L1 [33] for computing optical flow, we obtain HOF features [30] as well. These features complement HOG and LBP by motion information. We call Base to the HOG+LBP/Linear-SVM and Base+HOF to the HOG+LBP+HOF/Linear-SVM.

SSL. The experiments are based on the spatiotemporal SSL (with past temporal window style) and settings $(\Delta x, \Delta y, \Delta f) = (3, 3, 5)$. In preliminary experiments we tested several values of K (Fig. 1), ranging from $K = 4$ to $K = 1$. The obtained results were very similar, thus we decided to set $K = 1$ (i.e., omitting the partition of the training sequence) since then the training is faster.

Table 1. Evaluation of SSL over different datasets, frame rates and pedestrian sizes. For FPPI $\in [0.01, 1]$, the miss rate average % is indicated.

Dataset	FPS	Experiment	Near	Medium	Reasonable
OursDS	Any	Base: HOG+LBP	39.71	50.83	45.91
	3	SSL(Base) Proj. - OptFl.	36.03 - 36.72	50.01 - 50.04	44.40 - 44.02
		Base+HOF	47.98	56.65	50.88
		SSL(Base+HOF) Proj.	37.62	52.21	45.47
	10	SSL(Base) Proj. - OptFl.	35.49 - 34.79	50.22 - 49.42	43.56 - 42.10
		Base+HOF	39.24	52.37	42.43
		SSL(Base+HOF) Proj.	29.42	44.62	37.13
	30	SSL(Base) Proj. - OptFl.	34.18 - 34.01	49.84 - 48.04	42.90 - 41.73
		Base+HOF	37.81	53.39	38.78
		SSL(Base+HOF) Proj.	27.37	46.53	35.85
Caltech	25	Base	45.4	82.3	59.4
		SSL(Base) Proj. - OptFl.	40.6 - 38.9	81.2 - 80.4	59.4 - 57.6
		Base+HOF	33.8	78.4	52.9
		SSL(Base+HOF) Proj.	32.0	77.1	51.6

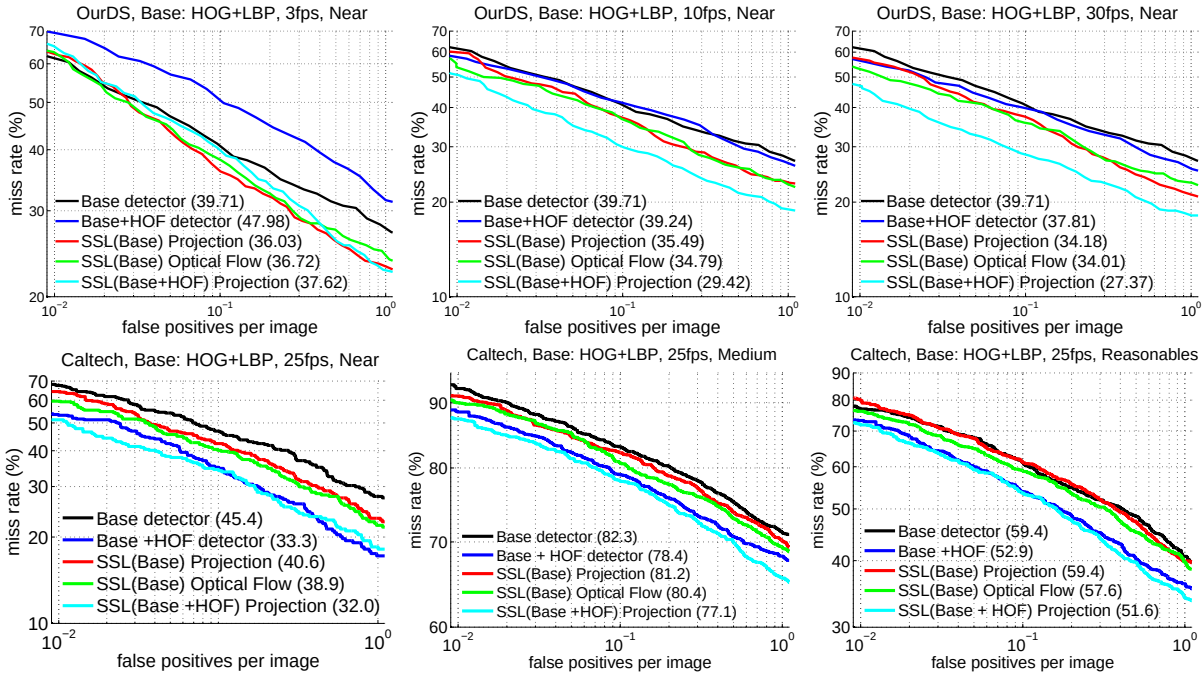

Figure 4. Results for OursDS and Caltech datasets. At the top row there are the 30fps, 10fps and 3fps cases of OursDS using the *near* testing subset. The last two cases are obtained by sub-sampling the video sequence, but always keeping the same training and testing pedestrians. At the bottom row there are the experiments over the *near*, *medium* and *reasonable* testing of Caltech dataset.

Experiments. In table 1 we show the results for the SSL experiments. As baseline detectors we use the Base and Base+HOF. The experiments are run over the different datasets, and different frame rates for the OursDS case. We tested them for different ranges of pedestrian sizes. We observe significant accuracy improvements for all the tested datasets comparing the baseline detector and its SSL counterpart. For instance, in OurDS near with SSL(Base+HOF) we obtain an accuracy improvement of ten points approximately. Also, significant accuracy improvements are obtained for all the tested frame rates (30 fps, 10 fps, 3 fps) of OurDS dataset. Besides, we observe an improvement due to the optical flow in the volume generation at high frame rates. However, no significant difference is observed at low frame rates. The SSL accuracy improvement is more clear for the near pedestrians. In Fig. 4 we plot the accuracy curves obtained for some representative experiments.

Discussion. SSL approach outperforms its baseline in almost all the tested configurations. However, the improvement is more clear for near pedestrians at high frame rates. If we generate the *past neighborhood* over the far away pedestrians, we should expect a *past neighborhood* with pedestrians smaller than the minimum pedestrian size that the base detector can detect. That is why the SSL improvement is not so clear for the medium subset. However, in

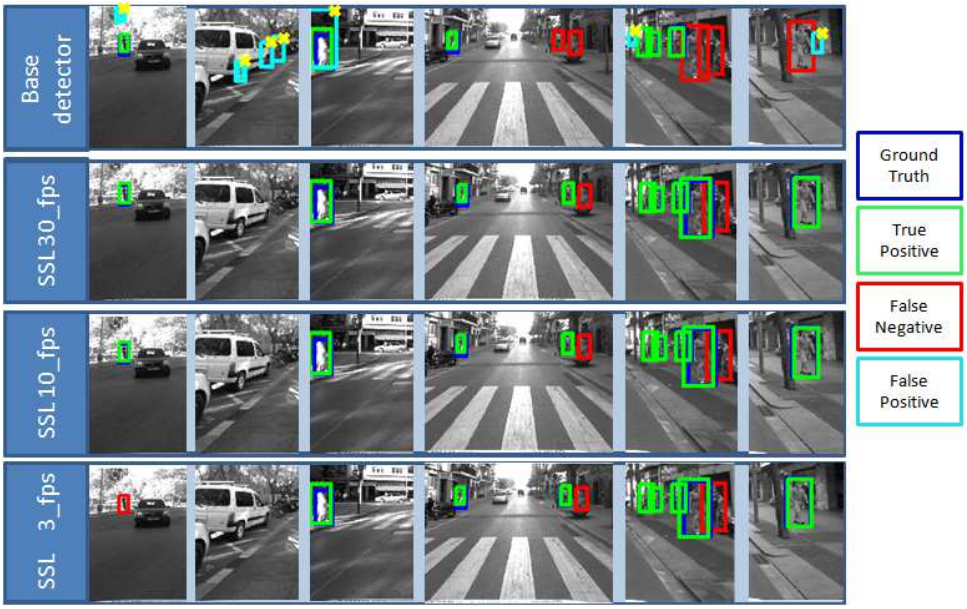

Figure 5. Qualitative results from the OursDS dataset comparing the base classifier and the SSL for 3, 10 and 30 fps. The first three columns focus on improvements regarding false positives rejection, while the rest focus on examples where SSL avoids missing pedestrians. The non-detected pedestrians with the SSL approach (last two columns) correspond to occluded pedestrians.

near pedestrians *past neighborhood* is more probable to find a history of confident responses. This is a very relevant improvement since for close pedestrians the detection system has less time to take decisions like braking or doing any other manoeuvre. Regarding the neighborhood generation approaches, the optical flow slightly improves the projection one as it captures the movement of the pedestrians in the temporal neighborhood.

6. Conclusion

In this paper we have presented a new method for improving pedestrian detection based on spatiotemporal SSL. We have shown how even simple projection windows can boost the detection accuracy in different datasets acquired on-board. We have shown that our approach is effective for different frame rates. In this paper we have focused on HOG+LBP/Linear-SVM and HOG+LBP+HOF/Linear-SVM pedestrian base classifiers, thus, our immediate future work will focus on testing the same approach for other base classifiers of the pedestrian detection state-of-the-art. Regarding the improvement obtained using optical flow neighborhood, we want to further explore different approaches for dealing with the neighborhood generation for moving pedestrians.

Acknowledgements

This work is supported by the Spanish MICINN projects TRA2011-29454-C03-01 and TIN2011-29494-C03-02 and Sebastian Ramos' FPI Grant BES-2012-058280.

References

[1] Y. Abramson and Y. Freund. SEmi-automatic VIsuaL LEarning (SEVILLE): a tutorial on active learning for visual object recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[2] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[3] G. Chen, Y. Ding, J. Xiao, and T. Han. Detection evolution with multi-order contextual co-occurrence. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, Oregon, USA, 2013.

[4] Y.-T. Chen and C.-S. Chen. Fast human detection using a novel boosted cascading structure with meta stages. *IEEE Trans. on Image Processing*, 17(8):1452–1464, 2008.

[5] W. Cohen and V. de Carvalho. Stacked sequential learning. In *Int. Joint Conferences on Artificial Intelligence*, Edinburgh, Scotland, 2005.

[6] X. Cui, Y. Liu, S. Shan, X. Chen, and W. Gao. 3d haar-like features for pedestrian detection. In *IEEE Int. Conf. on Multimedia & Expo*, Bejing, China, 2007.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[8] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conf. on Computer Vision*, Graz, Austria, 2006.

[9] Y. Ding and J. Xiao. Contextual boost for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[10] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *British Machine Vision Conference*, London, UK, 2009.

[11] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[12] M. Enzweiler and D.M. Gavrila. A multi-level mixture-of-experts framework for pedestrian classification. *IEEE Trans. on Image Processing*, 20(10):2967–2979, 2011.

[13] M. Enzweiler and D. Gavrila. A mixed generative-discriminative framework for pedestrian classification. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[14] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[15] D. Gerónimo and A. López. *Vision-based Pedestrian Protection Systems for Intelligent Vehicles*. Springer, 2013.

[16] D. Gerónimo, A. Sappa, D. Ponsa, and A. López. 2D-3D based on-board pedestrian detection system. *Computer Vision and Image Understanding*, 114(5):583–595, 2010.

[17] M. Jones and D. Snow. Pedestrian detection using boosted features over many frames. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[18] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Int. Conf. on Computer Vision*, Beijing, China, 2005.

[19] J. Lafferty, A. McCallum, and F. Pereira. Real-time pedestrian detection with deformable part models. In *IEEE Intelligent Vehicles Symposium*, Madrid, Spain, 2012.

[20] J. Marin, D. Vázquez, A. López, J. Amores, and L. Kuncheva. Occlusion handling via random subspace classifiers for human detection. *IEEE Trans. on Systems, Man, and Cybernetics (Part B)*, 2013.

[21] J. Marin, D. Vázquez, A. López, J. Amores, and B. Leibe. Random forests of local experts for pedestrian detection. In *Int. Conf. on Computer Vision*, Sydney, Australia, 2013.

[22] W. Nam, B. Han, and J. Han. Improving object localization using macrofeature layout selection. In *Int. Conf. on Computer Vision - Workshop on Visual Surveillance*, Barcelona, Spain, 2013.

[23] L. Oliveira, U. Nunes, and P. Peixoto. On exploration of classifier ensemble synergism in pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 11(1):16–27, 2010.

[24] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *European Conf. on Computer Vision*, Crete, Greece, 2010.

[25] D. Ramanan. *Part-based Models for Finding People and Estimating Their Pose*. Springer, 2009.

[26] M. Rao, D. Vázquez, and A. López. Color contribution to part-based person detection in different types of scenarios. In *Int. Conf. on Computer Analysis of Images and Patterns*, Seville, Spain, 2011.

[27] N. Schneider and D. Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In *German Conference on Pattern Recognition*, Saarbrücken, Germany, 2013.

[28] D. Vázquez, A. López, J. Marín, D. Ponsa, and D. Gerónimo. Virtual and real world adaptation for pedestrian detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013.

[29] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Int. Conf. on Computer Vision*, Nice, France, 2003.

[30] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[31] X. Wang, T.X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *Int. Conf. on Computer Vision*, Kyoto, Japan, 2009.

[32] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

[33] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *German Association for Pattern Recognition (DAGM) Conference*, Heidelberg, Germany, 2007.