Unsupervised scene adaptation for faster multi-scale pedestrian detection

Federico Bartoli, Giuseppe Lisanti, Svebor Karaman, Andrew D. Bagdanov and Alberto Del Bimbo
Media Integration and Communication Center (MICC)
University of Florence - Florence, Italy
{federico.bartoli, giuseppe.lisanti, svebor.karaman, alberto.delbimbo}@unifi.it
bagdanov@dsi.unifi.it

***Abstract*—In this paper we describe an approach to automatically improving the efficiency of soft cascade-based person detectors. Our technique addresses the two fundamental bottlenecks in cascade detectors: the number of weak classifiers that need to be evaluated in each cascade, and the total number of detection windows to be evaluated. By simply observing a soft cascade operating on a scene, we learn scale specific linear approximations of cascade traces that allows us to eliminate a large fraction of the classifier evaluation. Independently, this time by observing regions of support in the soft cascade on a training set, we learn a coarse geometric model of the scene that allows our detector to propose candidate detection windows and significantly reduce the number of windows run through the cascade. Our approaches are unsupervised and require no additional labeled person images for learning. Our linear cascade approximation results in about 28% savings in detection, while our geometric model gives a saving of over 95%, without appreciable loss of accuracy.**

I. Introduction

Person detection provides the basic measurement model for tracking and person re-identification and is therefore a fundamental component of most modern surveillance systems. However, due to its computationally onerous nature it is also the bottleneck in many systems. The general problem of detection has emerged as one of the major themes of modern computer vision research. Person detection in particular is an highly active topic of research. It has received a lot of attention in recent years, but remains an extremely difficult problem.

Person detection in unconstrained scenes is computationally expensive for several reasons. First of all, without knowledge of the geometry of the scene, every location and scale must be scanned for potential detections. Second, in the *soft cascade* detection architecture, currently the state-of-the-art for efficient person detection, a cascade of *weak classifiers* must be evaluated at each of these locations and scales to obtain a detection score. These two factors conspire to render unconstrained detection computationally onerous.

Improvements in the computational cost of person detection often address only one of these factors and rely on supervision such as manual calibration of the camera. We believe that it is crucial in practice that both factors be addressed with only weak or no supervision. In this work we propose two approaches to scene adaptation for soft cascade pedestrian detectors that need only to observe an already trained detection on the scene of interest. Our first adaptation strategy performs linear cascade approximation to avoid evaluating all stages of the soft cascade, while our second strategy minimizes the number of candidate windows evaluated using a statistical model of scales and position of likely detections in the scene.

In the next section we discuss the state-of-the-art in person detection. In section III we describe the soft cascade detection architecture which represents the current state-of-the-art. We describe our approach to learning how to detect faster in section IV, and in section V we report on a number of experiments we performed to evaluate our approach. We conclude in section VI with a discussion of our contribution.

II. Related Work

Most state-of-the art methods follow the pipeline depicted in figure 1. Recently many techniques have been proposed that improve the detection process both in terms of accuracy and efficiency. These methods can be roughly grouped based on the domain on which they act: the multi-scale feature representation, the method used for proposing detection windows or exploiting scene geometry, and the classifier used.

An approach to computational saving in the feature domain was proposed in [1]. The Integral Channel Feature for integral images uses a combination of different heterogeneous information channels information to speedup the detection process while maintaining high accuracy. While a feature pyramid is mandatory for multi-scale detection, the authors of [2] proposed an approximation that avoid the direct computation of all levels of the feature pyramid by extracting them only for the median layer of each octave and approximating the remaining scales. This approximation, namely Aggregated Channel Features (ACF), takes the form of an exponential function that depends both on the type of the feature and on the position of the level in the octave. However, this preserves detector robustness only for an octave. In [3] the authors exploit a trained classifier for each octave and the approximation in [2] to avoid the computation of the features for each level in the octave.

Several methods have been proposed to speed up the computation by reducing the number of detection windows evaluated. In [4], [5] the authors propose to first compute a sparse set of detector responses and then sample more densely around promising locations. In [6] the Crosstalk Cascade was proposed to simultaneously evaluate multiple candidates at a time exploiting two type of cascade: *excitatory* cascades that encourage a detection window with a neighborhood of possible positive responses and *inhibitory* cascades that reject detection windows with low partial scores in the neighborhood.

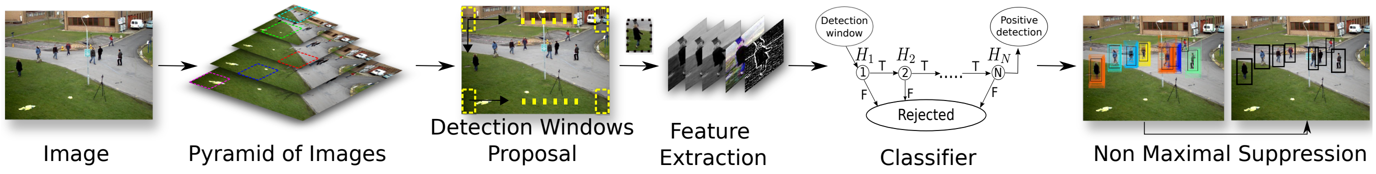


Fig. 1. Standard execution pipeline of a multi-scale pedestrian detector. Given an image I, a pyramid is computed from it by progressively sampling by a fixed factor to obtain the set of levels. For each level are selected the detection windows and then from each of these are extracted the feature that will be considered by the classifier. Finally, for all detection windows not rejected, a non maximum suppression process is performed to obtain the final positive detection windows.

The geometry of the scene is also extensively exploited in the to speedup the computation and improve detection accuracy. For example, the method proposed in [7] exploits a calibration of the scene to improve and speedup a person detector by spatially filtering detection windows based on the expected height of a person. In [8] the authors propose a probabilistic inference model to merge pre-trained detector responses with scene geometry knowledge. However, this method requires that the vanishing lines be always visible in the image in order to estimate a coarse camera viewpoint from objects in the scene. The Stixels model used in [3] exploits a stereo vision system to extract depth information of the scene and then reduce the set of candidate detection windows.

In the classifier domain the Hard Cascade [9] improves both the accuracy and efficiency of the classic AdaBoost algorithm [10] by specializing the first stages in order to reject the majority of the negatives detection windows. The authors of [11] proposed the Soft Cascade architecture in which the evaluation of each detection window depends on the sum of all stages partial scores up to the current stage. The rejection threshold at each stage is learned considering the ROC surface, thus taking into account conjointly the speedup, the detection rate and the false positive rate.

We propose a framework to speed up the detection process by acting both in the classifier domain and in the scene geometry domain. The result is a significant reduction in the total number of stages evaluation required in the soft cascade detection process. To do this we exploit the *regions of support*, which refers to the suppressed positive detections that occur around a local maxima, to improve detector efficiency in:

- the classifier domain through linear approximation of soft cascades in order to estimate a final detection score without calculating all stages;
- the pyramid domain by locally modeling the scene-dependent statistics of detection windows and their scale distribution in order to focus effort on the evaluation of detection windows that are more likely to be a local maxima in the image.

Our approach does not require any *a priori* information about the scene and all learning is done by mining statistics about the soft cascade detector operating on a scene.

III. Pedestrian detection with soft cascade

In figure 1 we show the standard pipeline for person detection. Since the process of capturing an image from a scene can introduce changes in the scale of a pedestrian, a multi-scale detector is required. This is usually performed by constructing a pyramid of images, which is a set of images obtained by progressively upsampling and downsampling the original image (referred to as the *levels* of image pyramid). Then each level is processed to extract the features. In particular, candidate regions are usually obtained using a sliding window at a fixed step size over all image levels. A classifier is then applied to each window for each level to assign a score. Finally, non maximum suppression is performed on positive candidates to obtain the final detection windows.

A. Multi-scale detection complexity

Without any optimization strategies, the evaluation of the whole pyramid of images in terms of total number of detection windows can be very expensive. Let L be the total number of levels of the pyramid, with m levels per octave, extracted for an image of $n \times n$ pixels, then the total number of windows that must be evaluated is:

$$\begin{aligned} \sum_{l=0}^{L-1} \mathcal{O}(n^2)\; 2^{\frac{-2l}{m}} &\approx n^2 \sum_{l=0}^{L-1} (4^{-\frac{1}{m}})^l \\ &= n^2 \left(\frac{1-4^{-\frac{L}{m}}}{1-4^{-\frac{1}{m}}} \right) \qquad (1) \end{aligned}$$

Note that eq. (1) converges to $n^2/(1-4^{-\frac{1}{m}})$ for $L \rightarrow \infty$. Thus, for an image of 640×480 pixels with a pyramid of 3 octaves of 8 levels each, a total of $285,944$ detection windows must be evaluated.

B. The soft cascade classifier

An evolution of the cascade classifier used in [9] is the *Soft Cascade* proposed in [11]. To train a Soft Cascade, a set of rejection thresholds is learned in order to perform early stopping during the evaluation of negative detection windows. Given the feature vector $x \in \mathbb{R}^D$ of a sample detection window, and let $H : \mathbb{R}^D \rightarrow \mathbb{R}$ be a classifier composed of T stages, where each stage is a function $h_i : \mathbb{R}^D \rightarrow \mathbb{R}$. The partial score up to stage t is computed as:

$$H_t(x) = \sum_{i=1}^{t} h_i(x). \qquad (2)$$

Let $\{\tau_t\}$ be the set of rejection thresholds, x is classified as *positive* with score $H_T(x)$ if $H_t(x) \geq \tau_t \quad \forall t \in [1, T]$. In this way the evaluation of each sample depends also on the scores obtained in the previous stages. Thus, considering the number of detection windows estimated in (1), it follows that using a soft cascade with 1024 stages requires the evaluation of approximately 10^9 stages for a single second of a video at 25 fps. This enormous number of cascade stages evaluated renders real-time pedestrian detection extremely challenging.

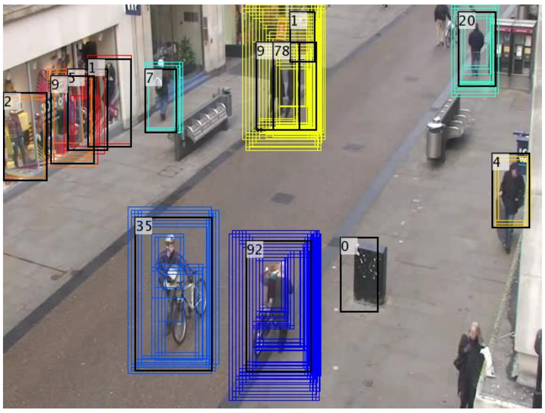


Fig. 2. The Region of Support (ROS) around strong detections (black detection window) on a frame extracted from *Oxford*. The windows inside the same ROS have the same color and at the top-left of each strong detection window we report the cardinality of each ROS.

IV. Unsupervised scene adaptation of soft cascade detectors

To avoid the computation of a very high number of stages as described in section III we propose a strategy to reduce the entire process by acting on both the classification and the detection windows proposal on the pyramid. The first contribution regards the total number of weak classifiers that must be evaluated to obtain a score for each positive detection window. In particular, we propose a solution to approximate the final score of a detection window without considering all the stages of a soft cascade. The second contribution provides an alternative strategy to the classic detection windows proposal that is able to avoid the sampling on the scene of those detection windows with a low probability of being a local maximum, in particular by filtering out those windows with a scale not consistent with the geometry of the scene. Both strategies are unsupervised and require only some frames extracted from the observed scene as a training set.

A. Leveraging region of support information

As reported in [6], the responses of the classifier on near positions and scales of the pyramid are related. A region of support (ROS) represents the extension of the sub-regions of an image in which all the detection windows (with different scale) are classified as positives. In general a ROS is composed of many intersecting detection windows, each with a different score. The window with the highest score is called local maximum (*strong*) because it is the only one that will survive the non maximum suppression procedure. Figure 2 shows some *strongs* with their respective ROS extracted from a soft cascade on a frame from the Oxford dataset [12]. The ROS shown can be very indicative of both the detector precision and the scene geometry, as well as the targets location inside the scene. In fact, the cardinality of each ROS can be used as a estimate of true positive for a detection window since the objects with a low rank in the frame are often false positive, e.g. the garbage and the mannequins. The location and scale of *strongs* can be considered to learn a model able to describe the geometry and perspective of the scene. All this information are very discriminative and can be extracted at no additional cost during the non maximum suppression process.

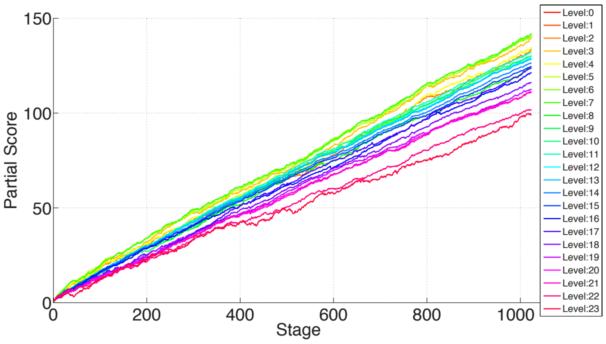


Fig. 3. Average positive traces extracted from a soft cascade of 1024 stages on the Oxford dataset. Traces are colored based on their level membership in the pyramid with 3 octaves of 8 levels each.

B. Linear cascade approximation

In figure 3 we plot the average positive traces from the Oxford sequence. Note how all traces are basically linear. They are subject to local perturbations of limited energy. Considering this trend, we estimate a linear function that approximates the trend of the traces from each level separately.

In particular, we define a linear score estimation function $\tilde{H}_{t\to T}(x) \in \mathbb{R}$ that requires the evaluation of only a fixed number $t < T$ of cascade stages and such that:

$$\tilde{H}_{t\to T}(x) \approx H_T(x), \tag{3}$$

where $H_T(x)$ represents the true cascade output obtained by evaluating all stages on input x. We use linear regression and estimate the slope and intercept parameters for each trace. Formally, we solve the following minimization problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\|\mathbf{S}^{\mathbf{T}}\mathbf{w} - \mathbf{h}_{t\to T}(x)\right\| \tag{4}$$

where $\mathbf{w} \in \mathbb{R}^2$, $\mathbf{w} = [w_0 \quad w_1]$ with w_0 the intercept and w_1 the slope and with:

$$\mathbf{S} = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 \\ t & t+\Delta & t+2\Delta & \cdots & T \end{bmatrix} \tag{5}$$

$$\mathbf{h}^{\mathbf{T}}_{t\to T}(x) = [H_t(x) \quad H_{t+\Delta}(x) \quad \cdots \quad H_T(x)] \tag{6}$$

where Δ is the sampling step for the stages used in the regression. Under the maximum rank hypothesis of $\mathbf{S}$ the problem in Eq. (4) admits a unique solution $\hat{\mathbf{w}} = (\mathbf{S}\mathbf{S}^{\mathbf{T}})^{-1}\mathbf{S}\ \mathbf{h}_{t\to T}(x)$.

We compute $\{\hat{\mathbf{w}}^l_i\}$ through eq (4) for each trace in each level l of the pyramid and then estimate the final parameter $\overline{\mathbf{w}}^l$ as the average. The final score approximation is:

$$\tilde{H}_{t\to T}(x) = \overline{\mathbf{w}}^l \cdot [0 \quad T-t] + H_t(x) + \overline{\epsilon}^l \tag{7}$$

where l is the level of x, ϵ^l is an error obtained as $E[H_T(x) - (\overline{\mathbf{w}}^l \cdot [0 \quad T-t] + H_t(x))]$ for $x \in V$, and V is a validation set. Note that Eq. (7) does not consider w_0, since the approximation is constrained to pass through $H_t(x)$ in that $\tilde{H}_{t\to t}(x) \simeq H_t(x)$.

Eq. 7 is easy and fast to compute and can be used to obtain an approximation of the final score of a detection window. This approximation requires the evaluation of only the first t stages of the soft cascade. Note also that it is completely unsupervised in that we only require a sample of cascade evaluations from an already trained soft cascade detector and do not require additional labeled training data to fit the model parameters.

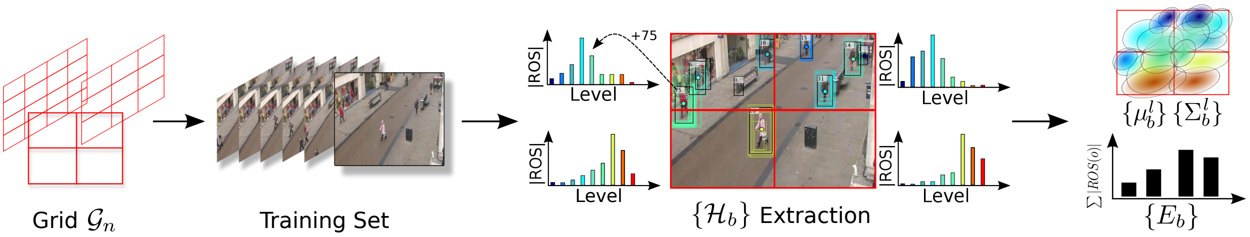


Fig. 4. Pipeline for training the candidate window proposal model. After selecting the grid resolution, for each frame of the training set we extract the histogram of levels $\mathcal{H}_b$ and $\{\mu_b, \Sigma_b\}$ considering the ROI information of strong detections. Finally, for each block we estimate the energy parameter E_b to accentuate the research in sub-regions of the scene.

C. A generative model for candidate window proposal

The naive soft cascade approach to detection achieves scale invariance by exhaustively scanning all locations and scales in an image. In practice, especially in typical surveillance scenarios using fixed cameras, not all scale/location combinations are feasible due to the geometry of the scene. Our second strategy is to learn a generative model for candidate window proposal in order to reduce the number of candidate windows extracted from the pyramid. We do this without relying on calibration or any additional information. As shown in figure 2, the presence and scale of targets is highly dependent on the geometry of the scene. Since the geometric information of the scene is directly related to the level of the pyramid, we argue that the complete evaluation of all possible levels of the pyramid in all subregions of the image is wasteful. Instead, we will exploit the ROS for observed strong detections (i.e. those that survive non maximum suppression) in order to propose candidate windows for each scale and position combination in the scene.

Learning the generative model The pipeline of the proposed model is shown in figure 4. To extract the statistics of subregions of the scene we divide each frame of the training set into $n \times n$ rectangular blocks. Inside each block b, the strong detections observed in the training set are used to compute a histogram $\mathcal{H}_b$ where each bin $\mathcal{H}_b^l$ represents a level of the pyramid. Specifically, the strong detections in the block b contribute with the cardinality of their ROS in the corresponding bin level. The cardinality of the ROS is the number of detections that are suppressed by the overlapping strong detection. This provides a robust local description of the frequent scales in a block.

To extract information about the representative locations in a block for a certain level we thus compute the average centroid position μ_b^l and its covariance Σ_b^l on the strong detections. This is useful to estimate the real locations in the scene where person detections occur with high probability. Finally, for each block we compute an energy factor E_b, such that:

$$E_b = \frac{\sum_{l=1}^{L} \mathcal{H}_b^l}{\sum_{\tilde{b} \in \mathcal{G}_n} \sum_{l=1}^{L} \mathcal{H}_{\tilde{b}}^l}, \tag{8}$$

where $\mathcal{G}_n$ indicate the set of blocks. This factor emphasizes the research for certain sub-regions by generating the detection windows proportionally. The final model is:

$$M_n = (\mathcal{G}_n, \{\tilde{\mathcal{H}}_b^l\}, \{\mu_b^l, \Sigma_b^l\}, \{E_b\}), \tag{9}$$

where $\tilde{\mathcal{H}}_b^l$ indicates $\mathcal{H}_b^l$ normalized over all levels in block b.

Candidate window proposal at detection time The number of detection windows of a pyramid to be evaluated is chosen proportionally to a parameter $\gamma \in [0, 1]$. This parameter is an estimate of target speedup. There is clearly a tradeoff between speed ($\gamma \to 0$) and accuracy ($\gamma \to 1$). In particular, given a test frame I, the number of detection windows that we evaluate for each block b and level l in the pyramid $\mathcal{P}(I)$ is:

$$N = \gamma \ |\mathcal{P}(\mathcal{I})| \ E_b \ \tilde{\mathcal{H}}_b^l. \tag{10}$$

where $|\mathcal{P}(\mathcal{I})|$ corresponds to the total number of detection windows in pyramid $\mathcal{P}(\mathcal{I})$.

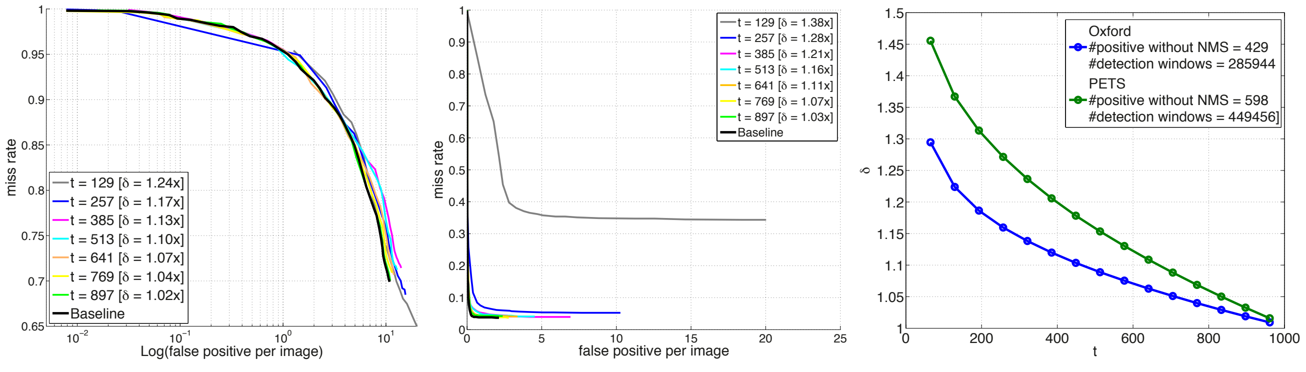
At detection time we sample detection windows using an iterative procedure. In the first iteration we randomly sample N detection windows from the normal distribution $\mathcal{N}(\mu_b^l, \Sigma_b^l)$. From this set of detection windows we remove duplicates and if necessary perform another iteration, expanding the covariance matrix by a fixed factor s along the principal directions of the covariance matrix Σ_b^l.

Note that this strategy for improving the efficiency of soft cascade detection is completely unsupervised. We build our scale- and position-local generative models by analyzing the behavior of strong detections and their regions of support on a training set of detection outputs. At detection time we can control the number of candidate windows proposed and thus control the efficiency/accuracy tradeoff of the final detector.

V. Experimental Results

In this section we report the performance of our linear cascade approximation, our candidate windows proposal model, and the combination of both. We use the soft cascade detector implemented in the OpenCV repository[1] as our baseline. We use two datasets in our experiments: Oxford [12] and PETS [13]. The Oxford dataset is a challenging full HD video sequence due to high variation of pedestrian scale, occlusions and confusion with shopping window mannequins. For the PETS dataset we considered the *s2.l1-view1* sequence with an image resolution of 768 × 576 pixels. We extracted 180 frames from Oxford by sampling one over fifteen frames and 199 frames from PETS by sampling one over four frames. From these frames we a third for the training and the remaining for the test. All comparisons between different detectors are given using ROC curves in terms of miss rate versus false positive per image. The baseline is represented by the soft-cascade with 1024 stages using a classifier for each octave and a pyramid of images consisting of 3 octaves of 8 levels each.

[1] Open source computer vision library, https://github.com/Itseez/opencv



(a) (b) (c)

Fig. 5. ROC curves of baseline using the linear cascade approximation, for different values of t, in sequence Oxford (a) and PETS (b). In bracket we show the obtained saving. (c) Saving (delta) for different values of t, using the linear cascade approximation. The maximum reduction is under the 40% (1.5x).

The performance of our proposed approaches is measured as function of a savings factor δ that is computed as:

$$\delta = \frac{\sum_{\forall x \in \mathcal{P}} [H(x)]}{\sum_{\forall x \in \mathcal{X}} 1_{\{c=0\}} \, [H(x)] + 1_{\{c=1\}} \, [\tilde{H}_{t \to T}(x)]} \qquad (11)$$

where the operator $[\cdot]$ returns the number of stages computed, c indicates if the linear cascade approximation is used ($c = 1$) and $\mathcal{X} = \mathcal{P}$ when all sliding detection windows are considered or $\mathcal{X} = \tilde{\mathcal{P}}$ when the set of detection windows is obtained from our generative model for candidate window proposal.

A. Experiments with linear cascade approximation

In this section we analyze the performance of linear cascade approximation for different t values. Observe in Figure 5(a) how on the Oxford sequence, the curves of the proposed approximation are close to the baseline, with a gradual reduction in loss when the number of stages evaluated increases. The total savings varies from 19% (1.24×) with 129 stages to 2% (1.02×) with 897 stages evaluated. For the PETS sequence, shown in figure 5(b), loss is drastically reduced for $t > 129$ stages. The maximum saving reached with this sequence is 28% (1.38×).

In figure 5(c), we show the savings evolution varying the number of stages evaluated for both sequences. Considering a small number of stages for each detection window, the computational savings is at most 23% (1.3×) in Oxford and 31% (1.45×) in PETS. The savings is modest because the computational cost is mostly dominated by the total number of *negative* windows evaluated, that decreases exponentially with increasing t (the number of stages considered for the linear cascade approximation). Linear cascade approximation helps, but to achieve significant computational cost reduction the total number of the candidate windows must be reduced.

B. Experiments with candidate windows proposal

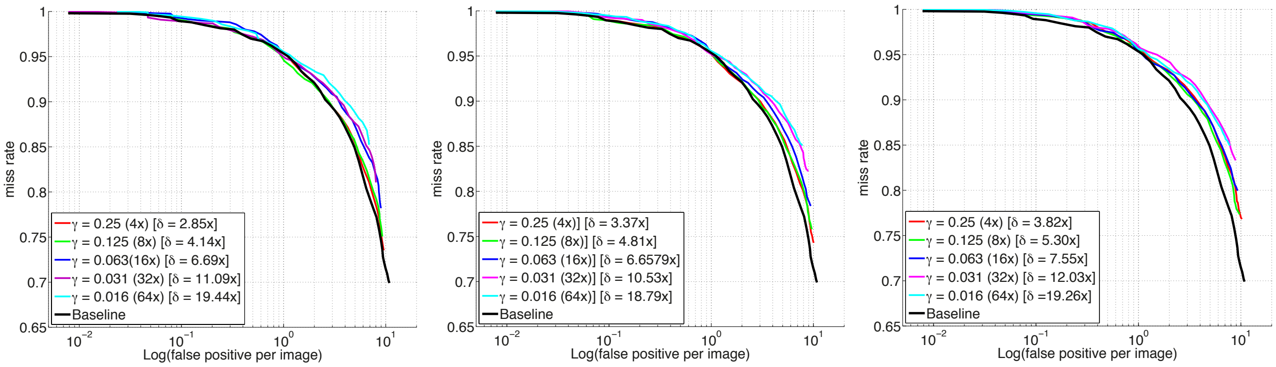
We evaluated the performance of our candidate window proposal model on the Oxford sequence for different values of γ and grid dimensions. The results are shown in figure 6. Each plot shows results for different grid resolutions (2×2, 4×4 and 6×6) and varying the speedup parameter γ. In general, with all configurations we obtain a savings greater then 50% (2×). For example, for a grid size of 2×2, the minimum and maximum saving values is 65% (2.85×) and 95% (19.44×), respectively. Considering the savings in computation, the loss in accuracy with respect to the baseline is very low at 10^{-1} *fppi* (under 0.5%). Increasing the grid resolution results in a small performance drops with respect to the baseline. The grid 2×2 is the best configuration in terms of loss and savings. This is due to the fact that, despite the large blocks in the 2×2 grid configuration, covariance expansion will ensure that the Gaussian will still eventually cover the whole block.

C. Experiments with both strategies

In this section we evaluate the combination of both proposed strategies on the Oxford and PETS sequence (see figure 7). Results are shown for different values of γ and t with a grid resolution of 2×2. On Oxford, with $\gamma = 0.25$ (4×) the maximum savings obtained respect to the baseline is 74% (3.78×), 9% more than the candidate window proposal alone, with no loss. For PETS, with $\gamma = 0.25$ (4×) we obtain a reduction of 81% (5.42×) with respect to the baseline, while with $\gamma = 0.0625$ (16×) we reach the 91% (11.26×) of saving. With both values of γ and $t \geq 513$ the obtained curves are the best in terms of accuracy with respect to the baseline, with a loss under 5%. The combination of the proposed strategies result in higher savings compared to the candidate windows proposal strategy while sacrificing little in terms of accuracy.

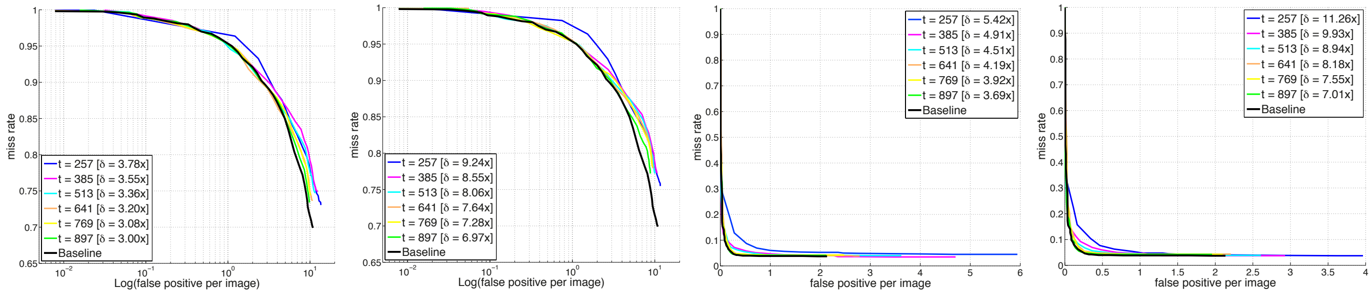
D. Comparison with the state-of-the-art

In table I we show the performance obtained by our method with respect to the state-of-the-art person detectors. Except the DPM detector [14] that uses a SVM classifier with a part-based model and is therefore slower, the rest of the strategies employ a soft-cascade architecture. On the Oxford dataset our strategies are very competitive with the ACF detector [2] in terms of miss-rate but we obtain a savings of 12.73× using both proposed strategies. Also, the miss-rate reached with any of our proposed strategies is less than the baseline because the number of detected false positive is reduced. On the PETS dataset with our candidate windows proposal strategy we obtain the lowest miss-rate value with respect to all the other detectors with a savings of 3.37×. These results prove the effectiveness of the proposed strategies.



(a) Grid 2×2 (b) Grid 4×4 (c) Grid 6×6

Fig. 6. ROC curves using candidate window proposal on Oxford sequence for a range of γ and grid sizes 2×2 (a), 4×4 (b), 6×6 (c).



(a) $\gamma = 0.25$ (4×) (b) $\gamma = 0.0625$ (16×) (c) $\gamma = 0.25$ (4×) (d) $\gamma = 0.0625$ (16×)

Fig. 7. ROC curves for both strategies, with a grid size of 2×2 and $\gamma \in \{0.25, 0.0625\}$, on both the Oxford (a-b) and PETS(c-d) sequences.

	Oxford		PETS	
Detectors	Miss-rate(%)	Savings(δ)	Miss-rate(%)	Savings(δ)
DPM [14]	80	-	34	-
ACF [2]	97	1×	51	1×
Baseline	99	1×	9	1×
Linear Cascade App.	98.1	1.17×	10.4	1.28×
Candidate Windows Pro.	98.9	11.09×	7.4	3.37×
With both	98.5	12.73×	11.4	4.19×

TABLE I. COMPARISON WITH THE MAIN PERSON DETECTORS OF THE STATE-OF-THE-ART. THE MISS-RATE SHOWN IS RELATIVE TO 10^{-1} FPPI.

VI. DISCUSSION

In this work we proposed two strategies to reduce the computational complexity of a multi-scale pedestrian detector. Both strategies are unsupervised, based only on region of support information measured on a training set of unlabeled images. Our first strategy linearly approximates soft cascades so that only a fraction of stages must be evaluated to obtain an (approximate) score. The second strategy builds a generative model for candidate window proposal to reduce the number of windows evaluated. The proposed techniques require only a training set of images and a soft cascade classifier.

Our experiments demonstrate that both techniques are effective at increasing the efficiency of detection while sacrificing little in terms of accuracy. Linear cascade approximation yields modest improvement in efficiency due to the fact that the evaluation of negative windows dominates the total computation time. Candidate window proposal instead yields significant gains since it reduces the total number of candidate detection windows considered.

REFERENCES

[1] P. Dollár, Z. Tu, P. Perona, and S. Belongie, “Integral channel features,” in *BMVC*, 2009.

[2] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” *PAMI*, 2014.

[3] R. Benenson, M. Mathias, R. Timofte, and L. J. V. Gool, “Pedestrian detection at 100 frames per second.” in *CVPR*, 2012.

[4] N. J. Butko and J. R. Movellan, “Optimal scanning for faster object detection,” in *In Proc. CVPR*, 2009, pp. 2751–2758.

[5] G. Gualdi, A. Prati, and R. Cucchiara, “Multi-stage sampling with boosting cascades for pedestrian detection in images and videos.” in *ECCV (6)*, 2010.

[6] P. Dollár, R. Appel, and W. Kienzle, “Crosstalk cascades for frame-rate pedestrian detection,” in *ECCV*, 2012.

[7] A. Del Bimbo, G. Lisanti, I. Masi, and F. Pernici, “Person detection using temporal and geometric context with a pan tilt zoom camera,” in *Proc. of ICPR*, Istanbul, Turkey, 2010.

[8] D. Hoiem, A. A. Efros, and M. Hebert, “Putting objects in perspective,” in *Proc. of CVPR*, 2006.

[9] P. Viola and M. Jones, “Robust real-time object detection,” in *International Journal of Computer Vision*, 2001.

[10] P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” *Int. J. Comput. Vision*, pp. 153–161, 2005.

[11] L. Bourdev and J. Brandt, “Robust object detection via soft cascade,” in *Proc. CVPR*, 2005.

[12] B. Benfold and I. Reid, “Stable multi-target tracking in real-time surveillance video,” in *CVPR*, June 2011, pp. 3457–3464.

[13] “Pets 2009 benchmark data, dataset s2: People tracking,” 2009.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *PAMI*, 2010.