Self-learning Scene-specific Pedestrian Detectors using a Progressive Latent Model

Qixiang Ye[1]†, Tianliang Zhang[1], Wei Ke[1]†, Qiang Qiu[2], Jie Chen[3], Guillermo Sapiro[2], and Baochang Zhang[4]
[1]EECE, University of Chinese Academy of Sciences.
[2]ECE, Duke University. [3]University of Oulu, Finland. [4]ASEE, Beihang University.
qxye@ucas.ac.cn; kewei11@mails.ucas.ac.cn

Abstract

In this paper, a self-learning approach is proposed towards solving scene-specific pedestrian detection problem without any human' annotation involved. The self-learning approach is deployed as progressive steps of object discovery, object enforcement, and label propagation. In the learning procedure, object locations in each frame are treated as latent variables that are solved with a progressive latent model (PLM). Compared with conventional latent models, the proposed PLM incorporates a spatial regularization term to reduce ambiguities in object proposals and to enforce object localization, and also a graph-based label propagation to discover harder instances in adjacent frames. With the difference of convex (DC) objective functions, PLM can be efficiently optimized with a concave-convex programming and thus guaranteeing the stability of self-learning. Extensive experiments demonstrate that even without annotation the proposed self-learning approach outperforms weakly supervised learning approaches, while achieving comparable performance with transfer learning and fully supervised approaches.

1. Introduction

With widespread use of surveillance cameras, the need for automatically detecting objects, e.g., pedestrians, has significantly increased. Recent methods [9, 13, 18, 27] have achieved encouraging progress for detecting objects in images. However, their performance in video scenes is limited for the following reasons: 1) Supervised learning of detectors for different scenes requires repeated human effort; 2) Offline-trained detectors usually degrade with changes in the scene or camera; 3) Scene specific cues including object resolution, occlusions, and background structures are not incorporated into the detectors [29].

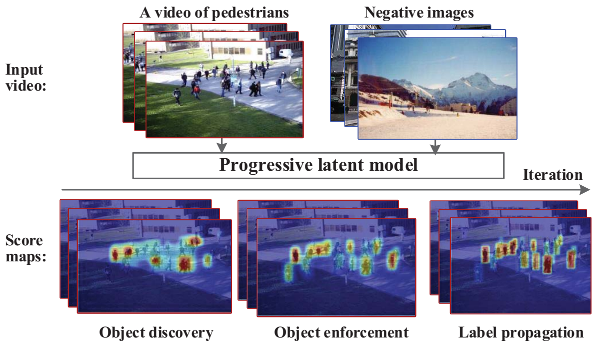


Figure 1. **Proposed self-learning framework**. Given a video where pedestrians are dominant moving objects, self-learning progressively constructs a scene-specific detector using object discovery, object enforcement, and label propagation procedures.

Learning scene-specific detectors, which aims at modeling objects in video scenes by incorporating scene-specific discriminative information, has been increasingly investigated [19, 25, 31]. To learn scene-specific detectors with less human supervision, transfer learning and semi-supervised learning are commonly used [19, 25, 31]. Transfer learning adapts pre-trained detectors to new specific domains, reduces annotation requirements and improves detector performance [35, 36, 37]. Semi-supervised learning saves human annotation effort by initially training detectors with a few annotated examples, and incrementally improving the detectors by extending the sample domains [11, 25, 41]. However, transfer learning is challenged when the object appearance in the target domains has significant differences with that in the source domains; while semi-supervised models might drift away from the intended aims given noisy or unrelated samples [25]. Most importantly, both methods require partial object-level annotations, and therefore, do not fully eliminate human supervision.

As a promising direction, recent unsupervised video object discovery techniques [23, 26, 39] had been significantly improved, which are supposed to break the bottleneck

†Corresponding Authors

of the self-taught learning in practical applications. This paper discusses the possibility of self-learning pedestrian detectors in specific and dynamically changing scenes, e.g., a city square, to build a pedestrian detection system in a fully unsupervised manner, given video sequences where pedestrians are the dominant moving objects and additional negative images randomly collected from the Web, Fig. 1. The problem of self-learning is decomposed into three main components: object discovery, object enforcement, and label propagation. Object discovery is implemented with a latent SVM method [43], which outputs coarse models and annotations by minimizing frame-level classification error. Object enhancement targets at enforcing object localization and reducing ambiguity, i.e., discriminate object parts with the objects themselves, by leveraging spatial regularization objective. Label propagation optimizes a graph-based objective function to gradually discover harder-positive instances in frames. It also enables the self-learning framework to find complex sample domains, e.g., a manifold space comprising multi-posture and multi-view objects [42]. The three procedures are formulated in a progressive latent model (PLM) with difference of convex (DC) objective functions, which are efficiently optimized with concave-convex programming in a progressive manner.

The main contributions of this paper consist of: (1) A self-learning pedestrian detection framework, which is deployed as iterative procedures of object discovery, object enforcement and label propagation, posing a new direction in the field of (unsupervised) object detection; (2) A progressive latent model (PLM), which uses spatial-temporal regularization to reduce ambiguity of samples, as well as addressing the stability of self-learning; and (3) Extensive experiments on PETS2009, Towncenter, PNN-Parking-Lot2/Pizza, CUHK Square, and 24-Hours datasets verify the performance of the proposed approach.

2. Related Works

Pedestrian detection using supervised methods has been extensively investigated [4, 10, 21, 32, 42, 45]. This work, however, is more related to scene-specific detection using transfer learning, online learning, weakly supervised learning, and unsupervised object discovery.

Transfer learning: The motivation behind transfer learning is that contexts and object distributions in target domains might be leveraged to improve the performance of pre-trained detectors in source domains. Researchers have explored context cues [35, 37], confidence propagation [37, 44], and virtual-real world adaptation [33] to realize smooth transfer. Gaussian process regression [40] and super-pixel region clustering [29] have been explored to select “safe” samples in target domains. Large margin embedding [22] and transductive multi-view embedding [15] have been explored to expand detector horizons. Researchers have also been using domain adaptation to construct a self-learning-camera [16].

Transfer learning can obviously reduce human annotations. Nevertheless, it suffers from the concept gap problem, i.e., the major differences of object appearance, viewpoint, and illumination between source and target domains. When the gap is significant, the adaptation of pre-trained models becomes non-smooth or infeasible. By contrast, self-learning initializes and improves detectors in the same scenes, naturally avoiding the concept gap problem.

Online/semi-supervised learning: Online learning and semi-supervised learning improves scene-specific detectors by taking advantage of the continuous incoming data stream from the target domains. Classical detection-by-tracking (DBT) [1, 24] initializes the system using offline trained detectors and leverages temporal cues to extend sample domains and cancel detection errors. Tracking-Learning-Detection (TLD) [20] initializes the system with a single sample, and uses tracking and online learning to boost detectors. Despite the popularity of DBT and TLD approaches, recent studies [25] demonstrated that the simple combination of detection with tracking might introduce poor detectors because the errors from both detection and tracking could be amplified in a coupled system. A P-N expert [20] is used in TLD to control precision and recall rates, guaranteeing the learning stability as a linear dynamic system. The learning stability of our approach can also be guaranteed as the difference of convex (DC) objective functions of PLM converge at each learning iteration.

Weakly supervised learning: The inputs of WSL are image/video level tags (object category), and the algorithm discovers objects when learning detectors [23, 30]. A general assumption behind WSL is that objects of the same category are from a potential cluster while the backgrounds are diverse. Under such an assumption, clustering [8, 34], tracking [23], boosting [38], region matching [6], graph labeling [30], and multi-instance learning [7, 28] are used to find the correspondence of objects, depress the backgrounds and learn detectors.

WSL alternates between sample labeling and detector learning in a way similar to Expectation Maximization optimization. Due to the missing annotations, however, this optimization is non-convex and therefore prone to getting stuck in a local minimum and outputting wrong labelings [3]. Cinbis *et al.* [7] use a multi-fold splitting of the training set while Bilen *et al.* [3] use convex clustering to prevent getting stuck to wrong labels. This work alleviates the local optima problem with a more reasonable way by introducing regularization terms about domain knowledge, i.e., intra-frame hard-negative mining and inter-frame similarity propagation.

Unsupervised video object discovery: An early approach developed in [38] learns scene-specific object detector by online boosting of part detectors, but it requires general seed detectors learned offline. Recent research [23, 39] formulates unsupervised video object discovery as a combination of two complementary steps: discovery and tracking. The first step establishes correspondences between prominent regions across video frames, and the second step associates successive similar object regions within the same video. Xiao *et al.* [39] propose a fully unsupervised video object proposal approach which first discovers a set of easy-to-group instances by clustering and then updates its appearance model to gradually detect harder instances by the initial detector and temporal consistency. This unsupervised approach can automatically generate object proposals, but cannot output precise detections.

3. Proposed Self-learning Framework

In the supervised object detection setting, the locations of training samples would simply be given, while in self-learning, the annotations of object locations are not available. The primary objective of self-learning is guiding the missing annotations to a solution that disentangles object samples from noisy object proposals, as shown in Fig. 2.

3.1. Progressive Latent Model

Modeling: The self-learning framework is decomposed into three basic procedures: object discovery, object enhancement, and label propagation. Given a set of object proposals that have salient object-like appearance and motion, Fig. 2a and Fig. 2b, the object discovery step aims to find object windows from video frames that best discriminates positive video frames from the negative images. The object enhancement discovers hard negatives that help reducing falsely localized object parts, as well as improving object localization. The label propagation step mines harder instances of the corresponding object and throughout the entire video, Fig. 2c and Fig. 2d. The three procedures iterate until an error rate based stability criteria is met.

Let $x \in \mathcal{X}$ denotes a video frame or a negative image, $y \in \mathcal{Y}, \mathcal{Y} = \{0,1\}$ are labels denoting if x contains a pedestrian object. $y = 1$ indicates that there is at least one pedestrian in the frame while $y = 0$ indicates a frame without pedestrian object or a negative image. The self-learning is formulated with a multi-objective function that targets at jointly determining the latent object h and a latent model β in a progressive optimized procedure,

$$\begin{aligned}\{h^*,\beta^*\} &= \min_{\beta,h}\mathcal{F}_{(\mathcal{X},\mathcal{Y})}(\beta,h)\\ &= \min_{\beta,h}\mathcal{F}_l(\beta,h) - \lambda\mathcal{F}_s(\beta) + \gamma\mathcal{F}_g(\beta,h),\end{aligned} \tag{1}$$

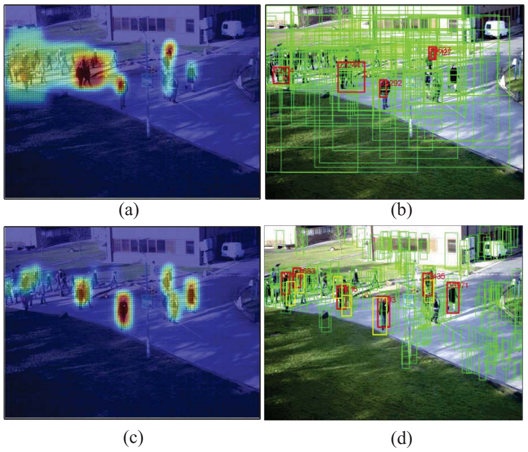


Figure 2. Object discovery from noisy proposals. (a) The score map in the first learning iteration and (b) candidate objects (red boxes) discovered. (c) The score map and in the fifth learning iteration. (d) Candidate objects (red boxes) and hard negatives (yellow boxes). (Best viewed in color.)

where $\mathcal{F}_l(\beta,h)$, $\mathcal{F}_s(\beta)$ and $\mathcal{F}_g(\beta,h)$ [*], as defined below, are the objectives for object discovery, spatial regularization and score propagation respectively. λ and γ are regularization factors.

Object Discovery: The object discovery procedure is implemented with a latent SVM (LSVM) model to choose object proposals that best discriminate positive frames from negative images,

$$\{y^*,h^*,\beta^*\} = \underset{y\in\mathcal{Y},h\in\mathcal{H},\beta}{\arg\max}\ \beta^T\cdot v\left(x,y,h\right), \tag{2}$$

where $v(x,y,h)$ denotes a normalized feature vector, i.e., HOG features. $\mathcal{H}$ denotes the set of object proposals, made up of proposals $\mathcal{H}_i, i = 1, ..., N$ from video frames. Basically, solving Eq. 2 produces a high score $\beta^T\cdot v(x,y,h)$ for each positive frame $(y = 1)$ and a low score for each negative image $(y = 0)$. Concretely, we learn the model β on a collection of video frames and negative images $X = \{(x_i,y_i), i = 1, ..., N\}$ with

$$\min_{\beta,h}\mathcal{F}_l(\beta,h) = \min_{\beta,h}\frac{1}{2}||\beta||^2 + \mathcal{C}\sum_{i=1}^{N} l(\beta,x_i,y_i,h), \tag{3}$$

where $\mathcal{C}$ is a regularization factor and l is a difference-convex loss function defined as

$$\begin{aligned}l(\beta,x_i,y_i,h) = &\max_{y,h}\left(\beta^T\cdot v(x_i,y,h) + \Delta(y_i,y)\right)\\ &-\max_{h}\beta^T\cdot v(x_i,y_i,h),\end{aligned} \tag{4}$$

where $\Delta(y_i,y) = 0$ if $y = y_i$, and 1 otherwise. Eqs. 3 and 4 target at choosing and discriminating the highest scoring

[*] $(\mathcal{X},\mathcal{Y})$ is omitted for short.

proposals h from the other configurations, defining a max-margin formulation to measure the mismatch between the image, label, and proposals.

Object Enforcement: The object discovery procedure aims at optimizing the image-level classification instead of the sample-level classification. Once the classification objective function is optimized, whether or not the sample-level classification is optimized, the learning procedure stops [43]. Considering that all positive images contain the object parts but none of negative images does, LSVM could falsely select object parts as positive samples since Eq. 3 is non-convex and is easy to get stuck to local minimum.

Motivated by the success of hard negative mining [17], we propose using spatial regularization to enforce the localization of objects and the model. Denoting by $\mathcal{H}_i$ object proposals in frame i and h' the hard negatives corresponding to an object h in a video frame, we define a function to maximize the distance between the potential object and its spatial neighbors,

$$\max_{\beta} \mathcal{F}_s(\beta)=\sum_{i=1}^{N}\sum_{\substack{h\in\mathcal{H}_i\\ h'\in\Omega_{\mathcal{H}_i,h}}} ||\beta^T\cdot\big(v(x_i,h)-v(x_i,h')\big)||^2, \tag{5}$$

where $\Omega_{\mathcal{H}_i,h}$ denote the spatial neighbors of h in $\mathcal{H}_i$. The spatial neighbors are high score object parts and surrounding image patches that have IoU (Intersection of Union) with h in the interval (0.0 0.25). Eq. 5 optimizes the model β using fixed h, and thus is a convex regularization function. Such a function enforces the latent model, yielding a consistent and significant boosts in object localization with a progressive learning procedure.

Label Propagation: The object discovery procedure outputs only one sample for each frame. To mine more positives and negatives, we propose using the inter-frame label propagation for incremental learning.

Suppose there are l labeled samples from previous learning iterations. We select $u = l \times (r - 1.0)$ high-scored proposals as unlabeled samples, where $r > 1.0$ is the learning rate, related to the expected density of pedestrians. Given labeled samples $\{h_i\}, i = 1, ..., l$, and unlabeled proposals $\{h_j\}, j = l, ..., l + u$, a kNN graph in the feature space is first constructed. The graph vertex defines the nearest neighbor vertices of samples. h_i and h_j are connected if one of them is among the others kNN [46]. The graph-based label propagation procedure is defined as $g(\beta, h_j) = \frac{\sum_{k=l}^{l} w_{jk} g(\beta,h_k)}{\sum_{k=l}^{l} w_{jk}}, j = l+1, ..., l+u$, where w_{ik} denotes the edge weight defined with a Gaussian Function on Euclidean distance between h_i and h_k. This is equivalent to a convex optimal problem [46],

$$\min_{g(\beta,h)} \mathcal{F}_g(\beta,h) = \min_{g(\beta,h)} \sum_{i=1}^{l}\sum_{j=l}^{l+u} w_{ij}\big(g(\beta,h_i)-g(\beta,h_j)\big)^2 \\ s.t. \quad g(\beta,h_i)=y_i, i=1,...,l, \tag{6}$$

where $g(\beta, h_j)$ is the propagated score of proposal h_j and y_i is the label of the frame/image that h_i belongs to.

Progressive Optimization: In the learning procedure, the optimization of $F_s(\beta)$ (object enforcement) and $F_g(\beta, h)$ (label propagation) depends on the results of $F_l(\beta, h)$. Eq. 1 is thus a progressive model, where F_l , F_s and F_g are alternatively optimized. According to Eq. 4, $\mathcal{F}_l$ could be written as $A(x) - B(x)$ and $\mathcal{F}$ could be written as $A(x) - B(x) + C(x) - D(x)$. This means that the objective functions of Eq. 1 could be written as the difference of convex functions. This allows us to optimize it with a two-step Concave-Convex Procedure (CCCP) [43]. The first-step CCCP for $\mathcal{F}_l$ discovers potential pedestrian objects in frames and initializes the latent model, the second-step CCCP for $\gamma\mathcal{F}_g - \lambda\mathcal{F}_s$ performs object enforcement and label propagation. The two-steps CCCP progressively optimizes the PLM until the change of the estimated sample error rate is negligible. CCCP algorithms guarantee the optimization with difference of convex objective functions converges to a local minimum or saddle point [43]. Therefore, iterative usage of the two-steps CCCP algorithm and keeping the decreasing of the sample error rates (discussed in Sec. 3.3) can guarantee the stability of self-learning.

3.2. Self-learning a Detector

With the proposed PLM, a self-learning approach is implemented as described in Fig. 3. The proposal generation component localizes potential objects using objectness, motion, and appearance cues. The proposal ranking component chooses the high-ranked proposals as positive candidates, and low-ranked proposals as negatives. The proposal tracking component helps in finding proposals in successive video frames. The PLM identifies positives and hard negatives from given proposals. With mined positive samples, a DPM detector $f_\beta(h)$ is trained to perform pedestrian detection.

Given a video of static background, a motion score map is calculated for each video frame with a background modeling algorithm. On the motion score map, detection proposals (as shown in Fig. 2b) are extracted using the EdgeBoxes approach [47], according to which edge maps are computed first, and contours, i.e., edge groups, are obtained by aggregating high affinity edges. On the contours, the regions of high confidence are extracted as object proposals using a sliding window strategy in locations, scales, and aspect ratios. From the second iteration, with

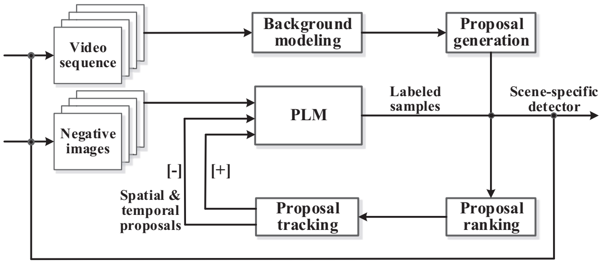


Figure 3. Block diagram of the proposed self-learning approach

an initialized detector, a sliding window strategy is used to generated object proposals, as shown in Fig. 2d. To extend the proposals in the temporal domain, a KLT tracking algorithm is employed to track and collect proposals from frame t to frame $t + \tau$, where τ is empirically set to 10. Before feeding these spatial-temporal proposals to the learning algorithm, their aspect ratios are normalized to the average aspect ratio. To prevent falsely choosing static backgrounds in videos of sparse pedestrians, the average background probability of a proposal is required to be larger than a threshold, empirically set to 0.20 in our experiments.

We propose using a combinatorial score, i.e., $f(h) = \alpha^T \cdot (f_\beta(h), f_m(h), f_o(h))$, to choose high-ranked proposals, where α^T is a ranking weight vector. $f_\beta(x)$, $f_m(h)$ and $f_o(h)$, respectively, are the detection, motion, and objectness scores. The motion score $f_m(h)$ of a proposal is defined as the averaged motion scores of all pixels in its image region. Objectness score $f_o(h)$ is defined by calculating contours in the proposal regions [47]. A larger score gives higher confidence that the proposal is an object. Detection score $f_\beta(h)$ is calculated from the second learning iteration, by the learned detector. From this iteration, the proposal region centers are set as root locations, around which we use sliding window to localize proposals.

In each learning iteration, the ranking weight vector α^T is updated using a zero-space regression method [5], which performs learning without using output values. It basically minimizes the regression error of all samples, as well as maximizing the distance from a hyperplane to the origin. This results in a weight vector which captures regions in the input sample space where the probability density of the data is found, and enables the proposal ranking to be adaptive.

3.3. Error Rate Discussion

PLM incorporates a label propagation procedure, which iteratively introduces new samples and updates the model. In this procedure, the primary problems to be solved are avoiding model drift and reducing the error rate. Eq. 6 implies that a larger γ value introduces more newly labeled samples, as well as a larger error rate ξ, and vice versa. The number of newly labeled samples u is determined to be an implicit function of γ, $u(\gamma)$. The value of γ needs to essentially guarantee that the error rate of newly labeled samples is smaller than that of existing samples, meaning the error rate of the training set is monotonically non-increased. It is also expected that there is a large γ, which implies that more samples could be labeled in each iteration. To decide the value of γ, an optimization objective function is defined:

$$\begin{aligned} &\max_{\gamma,\beta,y_j} \gamma \\ &s.t. \quad \xi_{u(\gamma)} \leq \xi_l \\ &\approx \frac{1}{l+u(\gamma)} \sum_{j=1}^{l+u(\lambda)} (f_\beta(h_j) - \widetilde{y}_j) \leq \frac{1}{l} \sum_{i=1}^{l} (f_\beta(h_i) - \widetilde{y}_i), \end{aligned} \tag{7}$$

where l and $u(\gamma)$, respectively, denote the numbers of labeled samples in previous iterations and unlabeled samples in current iteration.

The optimization of Eq. 7 guarantees that the estimated error rate of newly labeled samples $\xi_{u(\gamma)}$ is smaller than that of labeled samples ξ_l by finding a proper γ in each learning iteration. γ is optimized with a linear searching algorithm [12], which searches in the interval [0.0, 1.0] with step size 0.1 and updates $f_\beta(h_j)$ to $f_{\widetilde{\beta}}(h_j)$ at each step. Meanwhile, $\widetilde{y}_j$ is estimated with $\widetilde{y}_j = f_{\widetilde{\beta}}(\cdot)$, with which the error rate $\xi_{u(\gamma)}$ is calculated.

4. Experiments

4.1. Datasets and Performance Metrics

The proposed approach is evaluated on five real-world datasets (six sequences) captured with surveillance cameras. The datasets involve challenges from object occlusions, low resolution, and/or moving distractors.

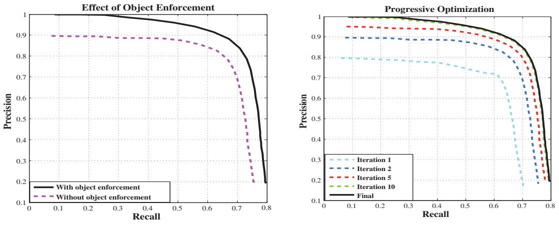
PETS2009 [14]: A crowded video sequence captured in a public space, with 720×576 resolution.

Towncenter [2]: A moderately crowded video sequence of a town center, with 1920×1080 resolution.

PNN-Parking-Lot2/Pizza [29]: Moderately crowded video sequences including groups of pedestrians walking in queues with complex motion and similar appearance, with 1920×1080 resolution. It is challenging due to the large amounts of pose variations and occlusions.

CUHK Square [37]: A 60-minutes long video of sparse pedestrians and other moving distractors, e.g., moving vehicles. The resolution of the video is 704×576. The resolution of pedestrian objects is much lower than those of other datasets. As the camera has an approximately 45-degree bird-view, objects have perspective deformation.

24Hours: A 24-hours long video of sparse/dense pedestrians, 24-hour illumination change and other moving distractors, e.g., moving vehicles, which allows to asses model



(a) (b)

Figure 4. Model effect.

drift. The resolution of the video is 704×576. 6000 frames were uniformly sampled from the long video for learning and 2600 frames for testing.

For all datasets except the 24Hours, half of the video frames are used for learning while the other annotated frames are used for testing. The proposed approach is evaluated and compared against the following supervised learning, transfer learning, and weakly supervised learning approaches.

Offline-DPM [13]: A DPM detector off-line trained on the PASCAL VOC person class.

Supervised-DPM: A supervised DPM detector trained with human annotated samples on specific scenes and additional negative samples mined from negative images.

Supervised-SLSV [19]: A state-of-the-art scene-specific pedestrian detector learned from virtual pedestrians whose appearance is simulated in the specific scene under consideration. Without public available source code, SLSV is only compared on the Towncenter dataset using author reported results.

Transfer-DPM [29]: A scene-specific detection approach based on transfer learning. Detections are originally obtained with a DPM detector off-line trained using PASCAL VOC person class and then improved using super-pixel based clustering and classification.

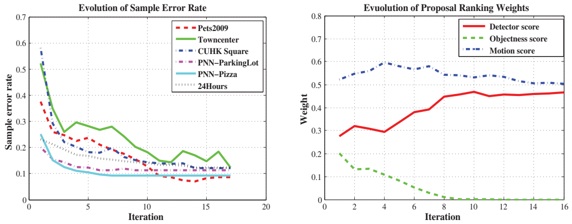
Transfer-SSPD [37]: A state-of-the-art scene-specific pedestrian detector with transfer learning.

Weakly-MIL [7]: A widely used weakly supervised approach based on multi-instance learning. A DPM learner is then learned from annotated positive samples.

4.2. Model Effect

In Fig. 4a and Fig. 4b, we respectively evaluate the effects of object enforcement and label propagation, showing that the PLM is more effective than the LSVM model.

Object enforcement: Considering that the objective function in Eq. 3 is non-convex, learning tends to get stuck into local minimum in the optimization procedure. By using the object enforcement procedure, Eq. 5, the performance of the learned detector significantly improved, Fig. 5a. The reason is that pedestrians are more precisely localised and most falsely detected object parts are depressed. Given the 0.7 recall rate, the precision improved more than 10% when using such a regularization term, which shows that the convex objective function does help the non-convex optimization to escape from poor local minimum.



(a) (b)

Figure 5. Learning stability. (a) Monotonical decrease of sample error rates. (b) Evolution of proposal ranking weights.

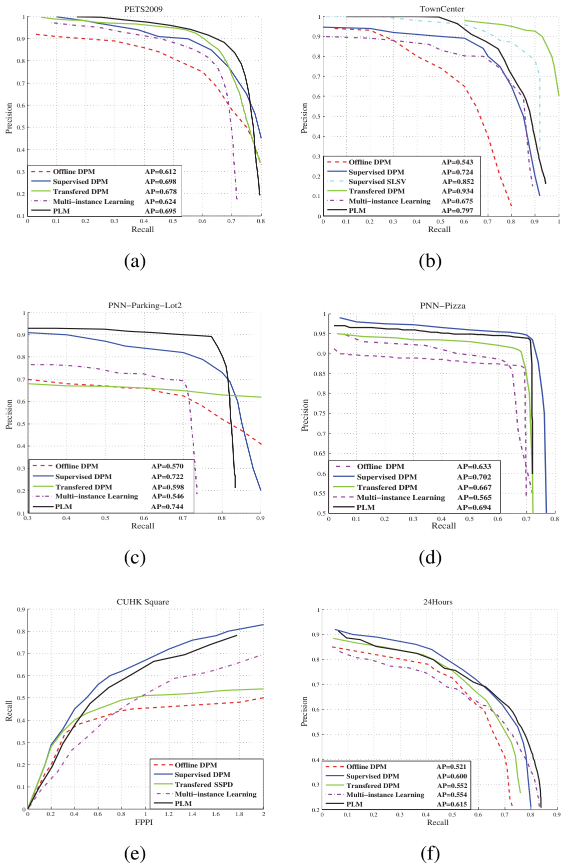
Table 1. Label propagation parameters on different datasets.

Dataset	PETS	Towncenter	PNN	CUHK	24Hours
γ	0.50	0.70	0.60	0.30	0.30

Label propagation: Combined with the proposal ranking strategy, label propagation can incrementally annotate pedestrian samples without supervision. Fig. 5b clearly shows that the detection model is iteratively improved, showing the effectiveness of the graph-prorogation based incremental learning. After tens of iterations of learning, no additional positives are labeled and the performance is observed to be stable.

Stability: Fig. 5a shows that the error rates of labeled training samples basically monotonically decreased, showing the stability of the proposed self-learning approach. Fig. 5b shows the evolution of proposal ranking weights in the learning procedure of the PETS2009 dataset. The weight for the objectness score quickly decays to zero, which implies that the objectness score is not as discriminative as the detection and the motion scores. The weight for the detection score keeps increasing in learning, which indicates that the detector is progressively improved. The weight for motion cue decreases to a value that is similar to the detection cue, which implies that the motion feature is also discriminative.

Tab. 4.2 shows the largest γ values for the four datasets. γ of the Towncenter dataset is the largest, while γ of the CUHK dataset is the smallest. Larger γ implies that the object proposals have fewer noises. The Towncenter dataset is a video with little illumination variance and few moving distracters, and therefore use a larger γ. The CUHK and 24Hours datasets have many moving distracters, so they need a smaller γ.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6. Performance of our approach and comparisons with weakly supervised, supervised, and transfer learning approaches. On five datasets the Precision-Recall metric is adopted to evaluate the approach and compare it with other approaches. On the CUHK dataset the FPPI-Recall metric is adopted, consistent with the state-of-the-art scene-specific detection approach [37].

4.3. Performance

The PR and FR curves in Fig. 7 show that our approach significantly outperforms the off-line learned DPM detector on all datasets. It also significantly outperforms the Weakly-MIL approach. On the PETS2009 and PNN-Parking-Lot2 datasets, our approach outperforms all of the compared approaches. On the CUHK dataset our approach significantly outperforms the scene-specific approach with transfer learning [37], which reports the state-of-the-art performance on this dataset. It is even comparable to the supervised learning approach (Supervised-DPM). On the Towncenter dataset, our approach outperforms the MIL approach as well. However, it shows lower performance than the fully supervised approach SLSV [19] and the transfer learning approach [29]. The reason could be that the pedestrians in that video scene are sparse, thus our approach could not label sufficient positive samples. It should be stressed once again that our proposed approach does not use any annotated training sample.

On the 24Hours dataset, the AP (average precision) of our approach is highest among all compared approaches, Fig. 7e. It is about 6% higher than the transfer learning method, validating our previous analysis: transfer learning suffers from the concept gap problem, e.g., adapt a model trained on day-time captured images to a video sequence of 24-hours illumination changes. By contrast, the proposed self-learning approach just applies the learned detectors from the same scenes, naturally avoiding the concept gap problem. More surprisingly, using additional motion cues, the proposed approach outperforms the fully supervised approaches in this dataset.

In Fig. 7, we use key frames in each row to illustrate the incremental learning procedure. It can be seen that the positive samples are incrementally labeled and noise samples are reduced. On the crowded PES2009 dataset and the PNN-Pizza dataset of significant occlusions our approach accurately labels samples, demonstrating that the learned detector has incorporated scene-specific discriminative information. On the Towncenter and CUHK datasets, although there exist moving distractors, e.g., bicycles and vehicles, the proposed approach correctly localize the pedestrians, demonstrating its robustness in noisy environments. In the 24Hours dataset, some video frames have dense pedestrians (daytime) but others have sparse pedestrians (at night). Learning from the early morning to the middle of the night, our approach could progressively improve its performance, without model drift. In the last column of Fig. 7, the detection results show that the learned scene-specific detectors are discriminative, showing robustness to occlusions, low resolution, and appearance variations. In Fig. 8, it can be seen that the self-learning approach is adaptive to view variance and 24-hours illumination changes, but transfer leaning suffers from those.

5. Conclusions

Supervised learning of detectors for all scenes requires significant human effort on sample annotation. Commonly used transfer learning and semi-supervised learning do not eliminate human supervision, as they require partial object-level annotations. We show that by leveraging extremely weakly annotated video data, it is possible to automatically learn customized pedestrian detectors for specific scenes. A new progressive latent model is proposed by incorporating discriminative and incremental functions. A self-learning approach is implemented by optimizing the model over spatio-temporal proposals. Experiments demonstrated that the self-learned detectors are comparable to supervised ones, taking a step towards self-learning cameras [16].

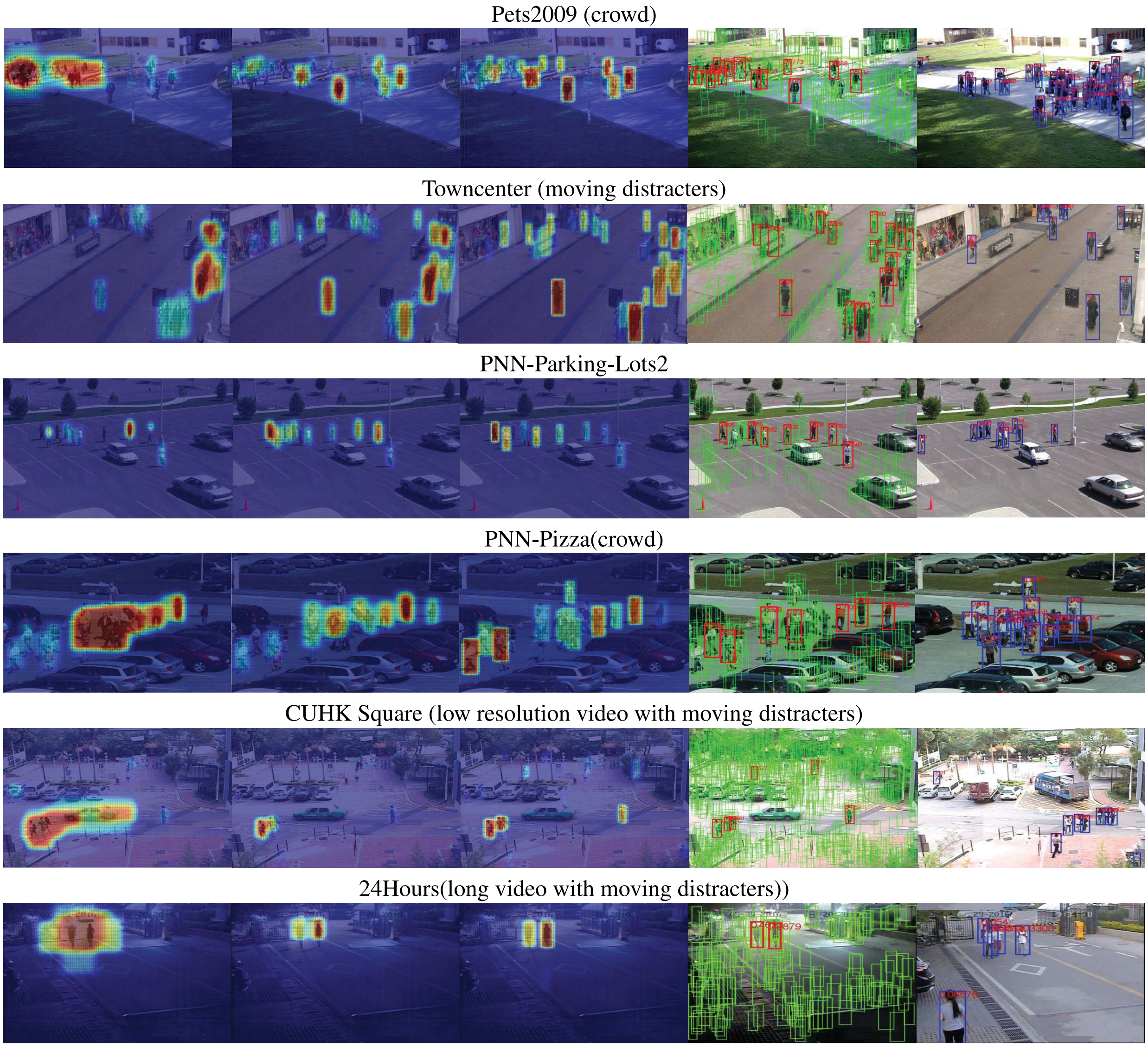
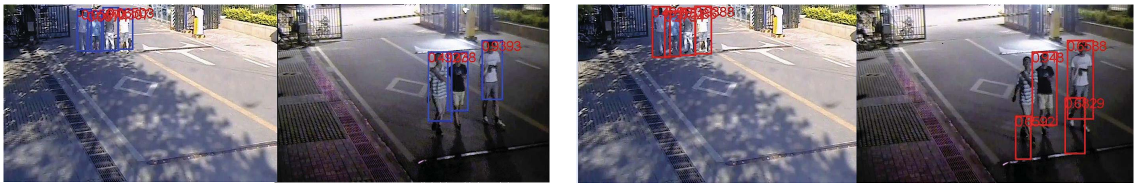

Pets2009 (crowd)
Towncenter (moving distracters)
PNN-Parking-Lots2
PNN-Pizza(crowd)
CUHK Square (low resolution video with moving distracters)
24Hours(long video with moving distracters))

Figure 7. Illustration of learning and detection. First three columns: score maps in the first, firth and tenth learning iterations, respectively. Fourth column: annotated positive samples (red boxes). Last column: detection examples in the test sets. (Best viewed in color)


Our proposed self-learning approach

The transfer-learning approach

Figure 8. Detection results on 24Hours dataset. The self-learning detection correctly detects all pedestrians from the daytime (left) and night (right), but transfer learning has missed and false detections.

Acknowledgement

This work is partially supported by the NSFC under Grant 61671427, Beijing Municipal Science and Technology Commission under Grant Z161100001616005, and NSF. Tekes and Infotech Oulu are gratefully acknowledged.

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. *IEEE CVPR*, 2008. 2

[2] B. Benfold and I. D. Reid. Stable multi-target tracking in real-time surveillance video. *IEEE CVPR*, 2011. 5

[3] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. *IEEE CVPR*, 2015. 2

[4] Z. Cai, M. Saberian, X. Wang, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. *IEEE ICCV*, 2015. 2

[5] C. Chang and C. Lin. Libsvm:a library for support vector machines. *ACM Trans. Intell. Sys. and Tech.*, 2(3):27, 2011. 5

[6] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild Part-based matching with bottom-up region proposals. *IEEE CVPR*, 2015. 2

[7] R. G. Cinbis, J. J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell., DOI: 10.1109/TPAMI.2016.2535231*, 2016. 2, 6

[8] S. K. Divvala, A. Farhadiy, and C. Guestrin. Learning everything about anything Webkly-supervised visual concept learning. *IEEE CVPR*, 2015. 2

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(8):1532–1545, 2014. 1

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, 2012. 2

[11] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell. Semi-supervised domain adaptation with instance constraints. *IEEE CVPR*, 2013. 1

[12] K. Donald. Sorting and searching. *The Art of Computer Programming 3 (3rd ed.). Addison-Wesley.*, 1997. 5

[13] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 1, 6

[14] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. *Twelfth IEEE Int'l workshop on performance evaluation of tracking and surveillance*, 2009. 5

[15] Y. Fu, T. M. Hospedales, T. Xiang, Z. Y. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. *ECCV*, 2014. 2

[16] A. Gaidon, G. Zen, and J. A. R. Serrano. Self-learning camera autonomous adaptation of object detectors to unlabeled video streams. *CoRR*, 2014. 2, 7

[17] R. Girshick. Fast r-cnn. *IEEE ICCV*, 2015. 4

[18] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE CVPR*, 2014. 1

[19] H. Hattori, V. N. Boddeti, K. Kitani, and T. Kanade. Learning scene-specific pedestrian detectors without real data. *IEEE CVPR*, 2015. 1, 6, 7

[20] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1409–1422, 2012. 2

[21] W. Ke, Y. Zhang, P. Wei, Q. Ye, and J. Jiao. Pedestrian detection via pca filters based convolutional channel features. *IEEE ICASSP*, 2015. 2

[22] A. Kuznetsova, S. J. Hwang, B. Rosenhahn1, and L. Sigal. Expanding object detectors horizon Incremental learning framework for object detection in videos. *IEEE CVPR*, 2015. 2

[23] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid. Unsupervised object discovery and tracking in video collections. *IEEE ICCV*, 2015. 1, 2, 3

[24] Y. Mao and Z. Yin. Training a scene-specific pedestrian detector using tracklets. *IEEE WACV*, 2015. 2

[25] I. Misra, A. Shrivastava, and M. Hebert. Watch and learn:semi-supervised learning of object detectors from videos. *IEEE CVPR*, 2015. 1, 2

[26] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrined video. *IEEE ICCV*, 2013. 1

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 1

[28] W. Ren, K. Huang, D. Tao, and T. Tan. Weakly supervised large scale object localization with multiple instance learning and bag splitting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):405–416, 2016. 2

[29] G. Shu, A. Dehghan, and M. Shah. Improving an object detector and extracting regions using superpixels. *IEEE CVPR*, 2013. 1, 2, 5, 6, 7

[30] H. O. Song, R. B. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. *ICML*, 2014. 2

[31] S. Stalder, H. Grabner, and L. V. Gool. Exploring context to learn scene specific object detectors. *IEEE Workshop on PETS*, 2009. 1

[32] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. *IEEE ICCV*, 2015. 2

[33] D. Vazquez, A. M. Lopez, J. Mar?n, D. Ponsa, and D. Geronimo. Virtual and real world adaptation for pedestrian detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4):797–809, 2014. 2

[34] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. *ECCV*, 2014. 2

[35] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. *IEEE CVPR*, 2015. 1, 2

[36] X. Wang, G. Hua, and T. X. Han. Detection by detections: Non-parametric detector adaptation for a video. *IEEE CVPR*, 2012. 1

[37] X. Wang, M. Wang, and W. Li. Scene-specific pedestrian detection for static video surveillance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2):361–374, 2014. 1, 2, 5, 6, 7

[38] B. Wu and R. Nevatia. Improving part based object detection by unsupervised via online boosting. *IEEE CVPR*, 2007. 2, 3

[39] F. Xiao and Y. J. Lee. Track and segment: An iterative unsupervised approach for video object proposals. *IEEE CVPR*, 2016. 1, 3

[40] J. Xu, S. Ramos, D. Vazquez, and A. M. Lopez. Domain adaptation of deformable part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(12):2367–2380, 2014. 2

[41] Y. Yang, G. Shu, and M. Shah. Semi-supervised learning of feature hierarchies for object detection in a video. *IEEE CVPR*, 2008. 1

[42] Q. Ye, Z. Han, J. Jiao, and J. Liu. Human detection in images via piecewise linear support vector machines. *IEEE Transactions on Image Processing*, 22(2):778–789, 2013. 2

[43] C. J. Yu and T. Joachims. Learning structural svms with latent variables. *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009*, pages 1169–1176, 2009. 2, 4

[44] X. Zeng, W. Ouyang, M. Wang, and X. Wang. Deep learning of scene-specific classifier for pedestrian detection. *ECCV*, 2014. 2

[45] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection. *IEEE CVPR*, 2016. 2

[46] X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *MIT Press*, 2009. 4

[47] C. L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges. *ECCV*, 2014. 4, 5