SMC faster R-CNN: Toward a scene-specialized multi-object detector

Ala Mhalla, Thierry Chateau, Houda Maamatou, Sami Gazzah, Najoua Essoukri Ben Amara

SMC Faster R-CNN: Toward a Scene-Specialized Multi-Object Detector

Ala Mhalla[a,b], Thierry Chateau[b], Houda Maâmatou[a,b], Sami Gazzah[a], Najoua Essoukri Ben Amara[a]

[a]*LATIS ENISo, National Engineering School of Sousse, University of Sousse, Tunisia*
[b]*Institut Pascal, Clermont Auvergne University, France*

Abstract

Generally, the performance of a generic detector decreases significantly when it is tested on a specific scene due to the large variation between the source training dataset and the samples from the target scene. To solve this problem, we propose a new formalism of transfer learning based on the theory of a Sequential Monte Carlo (SMC) filter to automatically specialize a scene-specific Faster R-CNN detector. The suggested framework uses different strategies based on the SMC filter steps to approximate iteratively the target distribution as a set of samples in order to specialize the Faster R-CNN detector towards a target scene. Moreover, we put forward a likelihood function that combines spatio-temporal information extracted from the target video sequence and the confidence-score given by the output layer of the Faster R-CNN, to favor the selection of target samples associated with the right label. The effectiveness of the suggested framework is demonstrated through experiments on several public traffic datasets. Compared with the state-of-the-art specialization frameworks, the proposed framework presents encouraging results for both single and multi-traffic object detections.

Keywords: Transfer learning, Deep learning, Specialization, Faster R-CNN, Sequential Monte Carlo filter, Traffic object detection.

1. Introduction

Learning-based object detection algorithms have become an essential part for numerous video analysis applications, including security and intelligent transportation systems [1][2]. However, most detectors are learnt with generic annotated datasets that are sampled from a large number of situations to cover the maximum intra-class variability of the traffic objects. When applied on a specific scene, the distribution of objects captured by the camera, like the Closed-Circuit Television camera (CCTV camera), is only a small subset of the initial learning set, and the resulting generic detector is often limited. Therefore, the detector may fail to perform satisfactorily when tested on scenes that have data distributions different from the source training dataset [3][4].

This problem can be solved by transfer learning, referred to as cross-domain adaptation, which can specialize a generic detector to a target scene. A classical way of specializing a generic detector is to manually select positive and negative samples from the target scene to re-train a scene-specific one. This requires collecting labelled data in every new scene and training a new detector, which can be labor intensive. A typical solution to avoid these tasks is to automatically label samples from the target scene and to transfer only a set of useful target samples to re-train a scene-specific detector.

Most state-of-the-art researches have been recently made to iteratively develop a scene-specific detector, whose training process is aided by generic detectors for automatically collecting training samples from target scenes without manually labelling them [2][5][6][7]. Accordingly, we put forward a new formalization of transfer learning based on the theory of a Sequential Monte Carlo (SMC) filter [8] so as to automatically generate a specialized Faster R-CNN detector [9] for multi-traffic object detection, enhancing perform better than the generic one.

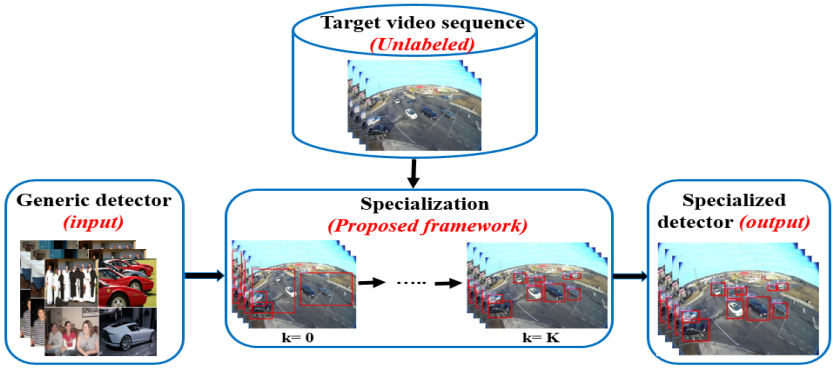
A global synoptic of our framework is illustrated in Figure 1.(a). We have a generic Faster R-CNN detector which is fine-tuned by a source labelled dataset with labeled information given in the form of traffic-object annotations. Given a target video sequence where labeled information is not available, an iterative process estimates both the set of target objects and the parameters of the specialized Faster R-CNN detector. This latter is automatically and iteratively trained and is called until a stopping criterion is reached. Then a final specialized Faster R-CNN detector is produced.

Our main contribution consists in putting forward a new transfer learning framework based on the formalism and the theory of the SMC filter for deep detector specialization. The aim of our formalization is to automatically label the target data, to favor the selection of the target samples associated with the right label and to fine-tune a scene specialized Faster R-CNN detector.

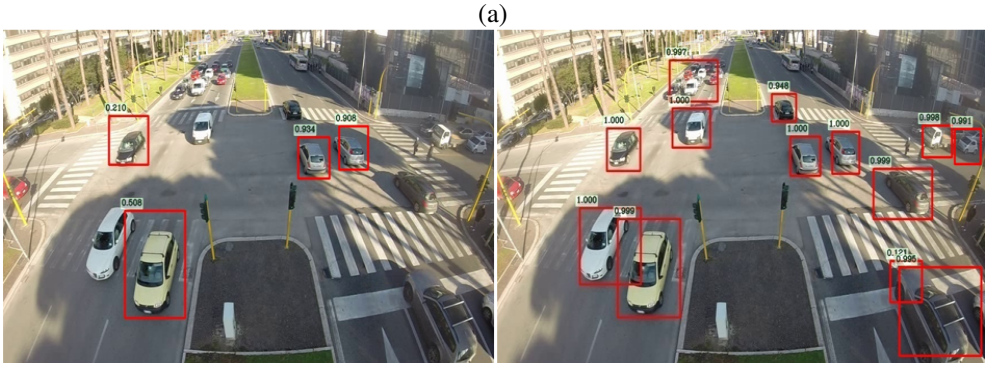
Although the use of the SMC filter for transfer learning is obviously not new, our work extends the SMC framework for deep detector and for multi-traffic object detection. Moreover, we propose new strategies for transfer learning inspired from the three steps of the SMC filter :

(1) Strategy of bounding box proposals: In order to use

Email addresses: mhallaala@gmail.com (Ala Mhalla), thchateau@gmail.com (Thierry Chateau), maamatou.houda@gmail.com (Houda Maâmatou), sami.gazzah@gmail.com (Sami Gazzah), najoua.benamara@eniso.rnu.tn (Najoua Essoukri Ben Amara)

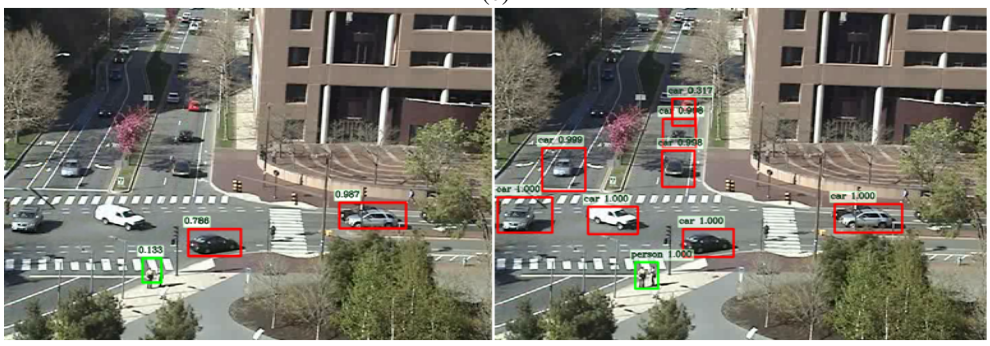


(a)



(b)

(c)



(d)

Figure 1: (a) General synoptic of the proposed framework. The input of the framework is a generic Faster R-CNN detector fine-tuned on a generic dataset, then given a target video sequence without any label information, an iterative process automatically estimates both the set of target objects and the parameters to specialize the Faster R-CNN deep detector; (b) and (c) improvement of specialized scene-specific detector over generic detector for single-class and (d) multi-class object detection (left images are generic Faster R-CNN detections and right images are specialized Faster R-CNN ones).

target samples for training a scene-specialized detector, the first strategy of the algorithm is to propose bounding boxes of traffic-object candidates by adapting the architecture of the Faster R-CNN deep network for only traffic-object detection. This strategy gives a set of suggestions composed by traffic proposals predicted by the output layers of the Faster R-CNN.

(2) Strategy of verification: We suggest a verification strategy to correctly select unlabeled samples from a target scene. This strategy utilizes a combination between the confidence-scores returned by the output layer of the Faster R-CNN and the visual context cues extracted from the target video sequence, in order to favor the selection of positive samples from a target scene and to reduce the risk of introducing wrong labelled examples in the training dataset.

(3) Strategy of sampling: We suggest a sampling strategy that collects useful samples from target datasets according to their weights importance, reflecting the likelihood that they belong to the target distribution. The main role of this strategy is to build the specialized dataset with samples produced by the strategy of verification. To do this, we use the Importance Sampling (IR) algorithm inspired from the theory of the SMC filter [10]. This algorithm transforms the weight on a number of repetitions, through repeating the samples associated to a high weight by numerous ones and repeating the samples associated to a low weight by few ones. This strategy makes the suggested framework applicable to specialize any detector and avoids the distortion of the specialized dataset, while selecting training samples according to the importance of their weights without modifying the training function.

Another contribution is to make a comparative evaluation of the proposed framework to the state-of-the-art specialization frameworks on several public datasets and with new more challenging annotations.

The rest of the paper is organized as follows. Section 2 reviews the existing work performed in this field and provides a discussion about the advantages of our work over the state-of-the-art specialization frameworks. After that, a detailed description of our approach are provided in section 3. The experiments and results are described in section 4. Finally, the conclusion is given in section 5.

2. Related Work

2.1. State-of-the-art scene specialization frameworks

In this subsection, we are interested in the related specialization frameworks that suggest to automatically specialize scene-specific detectors or classifiers towards a target scene.

In the recent years, transfer learning has attracted a lot of research groups in developing state-of-the-art theories and new applications in several domains like object detection and recognition [5][6][7][11]. Transfer learning aims to address the problem when the distribution of the training data from the source domain is different from that of the target one.

According to the state-of-the-art theories, transfer learning approaches suggest to use the available annotated data and knowledge acquired through some previous tasks relative to source domains so as to improve a learning system of a target task in a target domain.

Generally, three categories of transfer learning methods, related to the proposed framework, were described in [5]. The first one would change the parameters of a source learning model to improve its accuracy in a target domain [12][13]. The second category would decrease the variation between the source and target distributions to adapt a detector to the target domain [14][15]. The third one would automatically choose the training samples that could provide a better detector or classifier for the target task [5][11]. In this paper, we focus on the third category which utilizes an automatic labeler to select data from the target domain.

Much of the state-of-the-art research has used an iterative self-training process to specialize a generic detector to a target scene. An ideal framework can apply a generic detector on some frames in a target scene, score each detection using some heuristics and then include the most confident positive and negative detections to the original dataset for retraining [16][17][18]. Rosenberg *et al.* [17] opted for a self-training framework based on background subtraction to label scene samples. Only the samples with high confidence scores were added in a new training dataset from one iteration to another. Contrarily, there was a risk of introducing a wrong labelled example in the training dataset, which may degrade the framework performance over iterations. In addition, Wang *et al.* [6] utilized different contextual cues such as visual appearances of objects, motion of pedestrian, model of road, size and location to select positive and negative samples from the target scene and to add the last ones in the training dataset for retraining. This approach proved to be sensitive to the risk of drifting and it can be applied only onto a particular classifier.

Moreover, some solutions collected the training source dataset with new samples extracted from the target scene, which increased the time of training and the size of the dataset during iterations [13][15]. Others were limited only to the use of samples extracted from the target domain [19][11], which caused the loss of useful samples stored in the source dataset. Htike *et al.* [7] presented an approach that used only target samples labeled by a background subtraction algorithm and verified by the tracklet method to train a specific detector. In the same vein, Mao and Yin [11] used tracklet chains to automatically label target information. They associated the proposal samples predicted by an appearance-object detector into tracklets and they propagated labels to uncertain tracklets based on a comparison between their features and those of labeled tracklets. This framework used many manual parameters and several thresholding rules for every target scene, which can affect the specialization performance.

Other solutions were proposed in [5][20][16], which collected new samples from the target scene and the source dataset. Maamatou *et al.* [5] suggested a transfer learning method based on the SMC filter to iteratively build a new specialized dataset that was used to train a new specialized pedestrian detector. This produced dataset consisted of both source and target samples that were utilized to estimate the

unknown target distribution. Our proposed framework is inspired from this latter.

Addressing this problem with deep learning has recently gained a growing attention. Some deep models have been investigated in the unsupervised and transfer learning challenge [21]. Transfer learning using deep models has been turned out to be effective in some challenges [22][23] like traffic-object detection [24][20], emotion recognition [25] and sentiment analysis [26]. In order to take advantage of these types of detectors, several transfer learning methods have been proposed to specialize a Convolutional Neuronal Network (CNN) detector by fine-tuning an ImageNet-pre-trained model with a small target dataset. Li *et al.* [20] suggested adapting a generic CNN vehicle detector to a target scene by appropriating the shared filters between source and target data and updating the non-shared filters. In contrary to [20][27], which needed several manual labeling of data in the target scene, Zeng *et al.* [24] proposed to use Wang's approach [6] to select target samples and utilized these latter as an input to their CNN deep model to re-weight samples from target and source domains without manually labeling data from the target scene.

In this paper, we use a recent deep model, the Faster R-CNN [9], thanks to its efficiency and robust performance in general object detection and we specialize it with a new formalism of transfer learning based on the theory of the SMC filter [8] for multi-traffic object detection.

The Faster R-CNN was put forward in [9] to accurately detect general objects in pictures. It achieved a state-of-the-art 73.2 mean average precision on the PASCAL VOC 2007 dataset. It was composed of two modules: The first module is a Region Proposal Network (RPN) that provided a set of rectangular object proposals from an input image. The second module was the Fast R-CNN deep model [28] which took as inputs this set of object proposals and then used them for classification. The entire system was a single, unified network for object detection.

The suggested framework presented in this paper proposes some improvements over the related specialization frameworks. These improvements will be described in the next subsection.

2.2. *Literature analysis and framework proposition*

This section provides a discussion about the advantages of our work over the state-of-the-art scene specialization frameworks and the main difference between the SMC framework proposed by Maamatou *et al.* [5] and the suggested one.

Most of the specialization frameworks cited above are based on hard-thresholding rules and are very sensitive to the risk of drifting during iterations, or they are applied only to particular classifiers or few detectors like the HOG-SVM. In fact, several frameworks are limited only for mono-traffic object detection, or they need many iterations for the convergence of the specialization process.

Differently from the existing work, we put forward an iterative process based on the formalism of the SMC filter to specialize the Faster R-CNN deep detector for multi-traffic object detection. Accordingly, our proposed framework allows reducing the risk of drifting by using efficient strategies during iterations and it can be used to specialize any deep detector like the Fast R-CNN [28] and the R-CNN [29]. Furthermore, this framework may be applied using several strategies on each step of the SMC filter. Particularly, we cite some advantages of the suggested framework:

- We propose a likelihood function based on an efficient strategy of verification. This latter is used to favor the selection of samples associated to the right label from a target scene, to decrease the risk of drifting the detector over iterations by reducing the introduction of mislabeled examples in the training dataset.

- The suggested framework automatically specializes a generic detector to a target scene. This framework iteratively estimates the unknown target distribution as a specialized dataset by selecting only relevant samples from the target dataset. These samples are selected to re-train a specialized detector that increases the detection accuracy in the target scene. Contrarily, several state-of-the-art frameworks have aimed to collect samples from both source and target datasets to improve accuracy by augmenting the training dataset. These frameworks have led to extend the size of the training dataset and to slightly decrease the performance of the detector during iterations.

- To permit training an accurate specialized detector with the same function as the generic one and avoiding the distortion of the specialized dataset, we suggest a sampling strategy which uses the IR algorithm to select the confidence samples relevant to their weight returned by the likelihood function. This makes our framework applicable to specialize any deep detector, while training the treating samples according to the importance of their weight without modifying the training function, as done by [6] [16].

- We derive a generic transfer learning framework in which many strategies can be integrated in the SMC steps.

Table 1 provides a comparison over the SMC framework proposed by Maamatou *et al.*[5] and our suggested one.

The advantages of our specialization framework over the SMC framework [5] are:

- In [5], for each iteration, they selected relevant samples from both source and target domains to create a specialized dataset. In contrast, our proposed framework selects only the relevant samples from target domains according to the importance of their weights to create a specialized dataset. This solution enables a faster learning of detector and leads to an increase in detection accuracy.

- The specialized framework proposed in [5] was very sensitive to the risk of drifting because they used only a background subtraction algorithm to assign weights to the target samples. Indeed, several static objects or those

Table 1: Description of the difference between the work of Maamatou *et al.* [5] and our proposed one

	Maamatou *et al.* [5]	Our framework
Generic detector	HOG-SVM	Faster R-CNN
Transfer learning	Positive & negative samples	Positive samples
Specialized dataset	Source & target samples	Target samples
Output	Specialized SVM	SMC Faster R-CNN
Specialized process	SMC steps	SMC steps & fine-tuning step
Traffic objects	Pedestrian	Multi-traffic object

with similar background appearances were classified as negative samples, and mobile background objects were labeled as objects of interest. On the other hand, to avoid the distortion of the specialized dataset with mislabeled samples, we propose a likelihood function based on the verification strategy, which combines the confident-score given by the output layer of the Faster R-CNN network with spatial-temporal cues in order to attribute confidence weights to target samples.

- The work of Maamatou *et al.*[5] was limited for only single-traffic object detection, but our proposed one is extended for multi-traffic objects like cars, pedestrians, buses, motorbikes...
- Differently from the work in [5], we put forward new strategies for transfer learning inspired from the three steps of the SMC filter to specialize the Faster R-CNN deep detector.
- It is important to say that we need only two iterations for the convergence of our specialization process, whereas the framework suggested in [5] required at least 4 iterations for this convergence.
- The proposed approach in [5] was limited to specialize the SVM classifier, in contrary, our framework is applicable to specialize some deep detector like the Fast R-CNN [28], the Faster R-CNN [9] and the R-CNN [29].

3. Proposed specialization framework

In this section, we present the proposed framework for specializing the Faster R-CNN model to a target scene based on SMC filter steps. Figure 2 shows the block diagram representation corresponding to one iteration of our suggested SMC Faster R-CNN. First, a generic Faster R-CNN network $(\mathcal{R}_0, \mathcal{F}_0)$ is fine-tuned on a generic dataset (eg: PASCAL VOC). Given the videos taken by a stationary camera in target scenes, at a first iteration ($k = 1$), the generic detector $(\mathcal{R}_0, \mathcal{F}_0)$ is applied in the prediction step by using the strategy of bounding box proposals to suggest a set of traffic-object proposals in each individual image. Then an update step based on the likelihood function is used to favor the selection of the positive samples from a target scene by associating weight to each proposal sample returned by the prediction step. By utilizing the sampling strategy, the sampling step determines which samples should be included in the specialized dataset according to their weights. A new specialized detector $(\mathcal{R}_k, \mathcal{F}_k)$ is trained by using the training strategy in the fine-tuning step. This specialized one will become the input of the prediction step in the next iteration. The scene-specific detector is automatically and iteratively trained and is called until reaching a stopping criterion, for example a fixed number of iterations. When the number of iterations is reached, a final specialized detector $(\mathcal{R}_K, \mathcal{F}_K)$ will be generated.

In what follows, we first describe the specialization of the Faster R-CNN model based on the theory of the SMC filter.

3.1. *Faster R-CNN specialization based on SMC filter*

Given a source dataset, from which a generic Faster R-CNN detector can be trained from this source dataset, and a video sequence of a target scene, then a specialized Faster R-CNN detector will be generated. This latter is the output of the distribution approximation provided by the formalism of the SMC filter and the fine-tuning step. To do this, let us define:

- $\mathcal{I}_t \doteq \{\mathbf{I}^{(i)}\}_{i=1}^{I_i}$ is a set of unlabelled images extracted uniformly from a video sequence of a target scene.
- $\mathcal{D}_k \doteq \{\mathbf{x}_k^{(n)}\}_{n=1}^{N_k}$ is a specialized dataset at iteration k, where $\mathbf{x}_k^{(n)}$ is a target object sample to be detected in each target image of the set $\{\mathbf{I}^{(i)}\}_{i=1}^{I}$. This sample is defined by: $\mathbf{x}_k^{(n)} \doteq \{\mathbf{p}_k^{(n)}, y_k^{(n)}, s_k^{(n)}\}$ where $\mathbf{p}_k^{(n)} \doteq \{u_k^{(n)}, v_k^{(n)}, w_k^{(n)}, h_k^{(n)}\}$ is the position of an object, with $(u_k^{(n)}, v_k^{(n)})$ being the upper left coordinates of the object bounding box and $(w_k^{(n)}, h_k^{(n)})$ being the width and the height of the object bounding box, $y_k^{(n)}$ is the object class label and $s_k^{(n)}$ is an associated score.
- $\{\mathbf{x}^{(n)}\}_{n=1}^{N} = \Theta(\{\mathbf{I}^{(i)}\}_{i=1}^{I_i}; \mathcal{R}, \mathcal{F})$ is a function that applies the Faster R-CNN detector using the RPN network model $\mathcal{R}$ for the localization task and the Fast R-CNN network model $\mathcal{F}$ for detection. For both localization and detection, a set of candidate objects with associated scores is provided.
- $\{\tilde{\mathcal{R}}, \tilde{\mathcal{F}}\} = f(\{\mathbf{I}^{(i)}\}_{i=1}^{I_i}, \{\mathbf{x}^{(n)}\}_{n=1}^{N}; \mathcal{R}, \mathcal{F})$ is a fine-tuning function that returns the new parameters $\tilde{\mathcal{R}}$ and $\tilde{\mathcal{F}}$ of the Faster R-CNN network. The fine-tuning is performed from the Faster R-CNN network with initial $\mathcal{R}$ parameters for the RPN and initial $\mathcal{F}$ parameters for the Fast R-CNN, utilizing a training dataset given by the set of images $\{\mathbf{I}^{(i)}\}_{i=1}^{I_i}$ and the associated objects $\{\mathbf{x}^{(n)}\}_{n=1}^{N}$

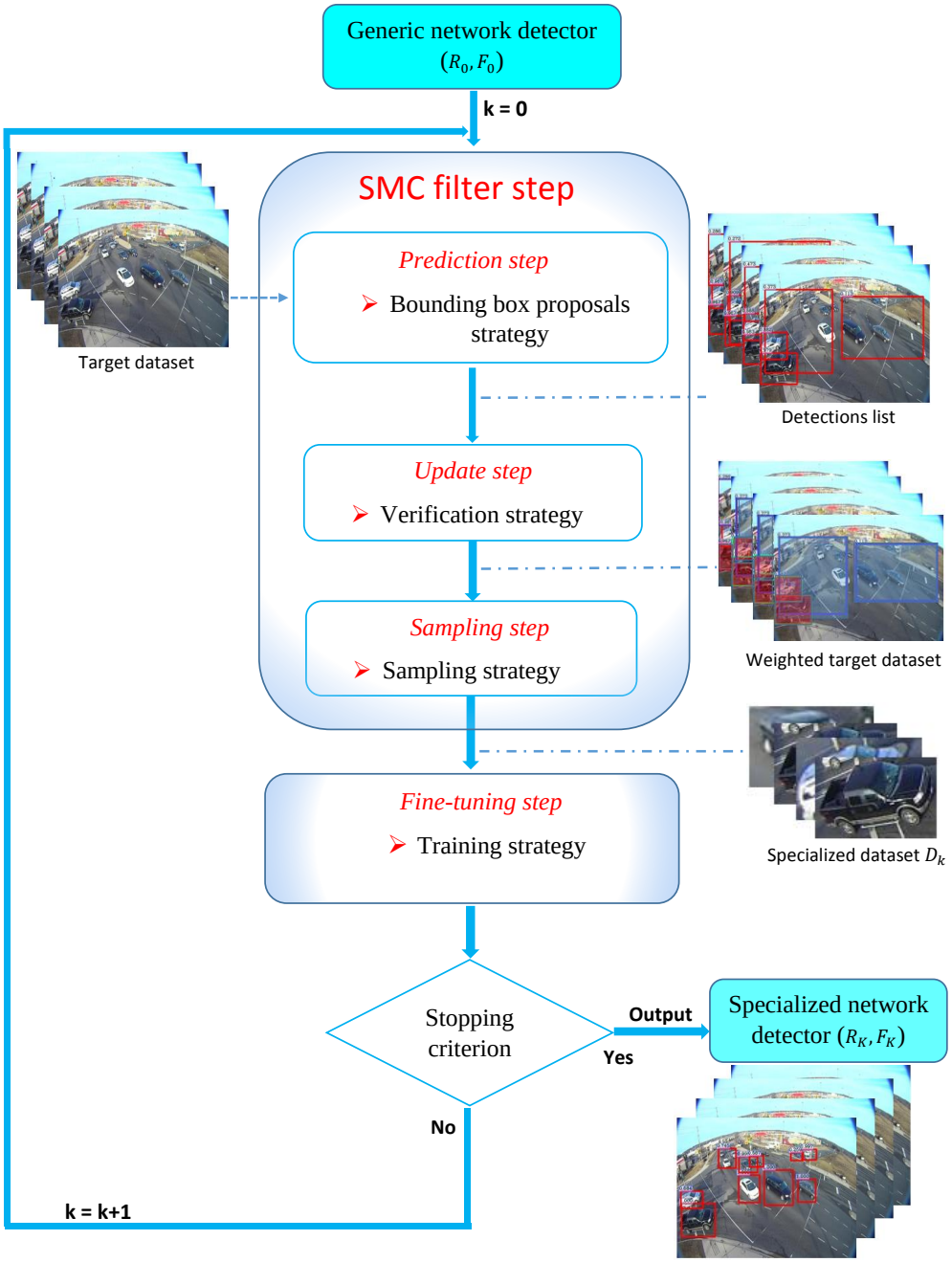


Figure 2: Block diagram of proposed approach: At the first iteration, our generic detector $(\mathcal{R}_0, \mathcal{F}_0)$ which is fine-tuned by the source dataset is utilized in the first prediction step by using bounding box proposals strategy to produce a list of traffic-object bounding boxes from the target scene, and then a update step based on the likelihood function is used to favor the selection of positive samples from a target scene. The sampling step determines which samples will be included in the specialized dataset by using the sampling strategy. A new specialized detector $(\mathcal{R}_k, \mathcal{F}_k)$ is fine-tuned by utilized training strategy in the fine-tuning step, which will become the input of the prediction step in the next iteration $k = k + 1$. A final specialized detector $(\mathcal{R}_K, \mathcal{F}_K)$ is called when a predefined number of iterations is reached. The red rectangles in the output image of update step mean that samples have a high weights attributed by our suggested likelihood function and a blue ones mean that samples have a low weights.

We define $\mathbf{x}_k$ to be a hidden random state vector associated to a joint distribution between labels and features of dataset samples at an iteration k and $\mathbf{z}_k$ a random measure vector associated to information extracted from the target scene (i.e. visual spatio-temporal information). Based on our assumption, the target distribution can be approximated by iteratively applying equation (1):

$$p(\mathbf{x}_k|\mathbf{z}_{0:k}) = C.p(\mathbf{z}_k|\mathbf{x}_k) \int_{\mathbf{x}_{k-1}} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{0:k-1}) d\mathbf{x}_{k-1} \quad (1)$$

where C is a normalisation factor: $C = 1/p(\mathbf{Z}_k|\mathbf{Z}_{0:k})$.

The SMC filter estimates the probability distribution $p(\mathbf{x}_k|\mathbf{z}_k)$ by a set of N particles (samples in this case), according to equation (2):

$$p(\mathbf{x}_k|\mathbf{z}_k) \approx \sum_{n=1}^{N} \pi_k^{(n)} \delta_{\mathbf{x}_k^{(n)}}(\mathbf{x}_k) \quad (2)$$

- δ represents the Dirac function (3):

$$\delta_{\mathbf{x}_k^{(n)}}(\mathbf{x}_k) = \begin{cases} 1 & \text{if} \quad \mathbf{x}_k = \mathbf{x}_k^{(n)} \\ 0 & \textit{otherwise} \end{cases} \quad (3)$$

- $\pi_k^{(n)} \in [0, 1]$ is the weight associated to sample n at iteration k and N is the number of target samples (4):

$$\pi_k^n = \frac{\pi_{k-1}^n \; p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{x}_k^n)}{\sum_{n=1}^{N} \pi_{k-1}^n \; p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{x}_k^n)} \quad (4)$$

  It is important to note that the sum of the weights of all the samples is equal to (5):

$$\sum_{n=1}^{N} \pi_k^{(n)} = 1 \quad (5)$$

All notations mentioned above are introduced in [8].

Therefore, the formalism of the SMC filter is used to approximate the unknown joint distribution of traffic objects by a set of samples that are initially unknown. We suppose that the iterative process selects relevant samples for the specialized dataset from one iteration to another, leading to converge to the right target distribution, and making the resulting Faster R-CNN detector more and more efficient.

The resolution of equation (1) is divided into three steps: prediction, update and sampling. These steps are similar to the popular particle filter framework, widely used to solve the tracking problems in computer vision [30][31]. The details of the three main steps are described in the following subsections.

3.1.1. Prediction step

The prediction step consists in applying the Chapman-Kolmogorov equation (6):

$$p(\mathbf{x}_k|\mathbf{z}_{0:k-1}) = \int_{\mathbf{x}_{k-1}} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{0:k-1}) d\mathbf{x}_{k-1} \quad (6)$$

Equation (6) uses the system dynamics term $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ between two iterations in order to suggest a specialized dataset $\mathcal{D}_k \doteq \{\mathbf{x}_k^{(n)}\}_{n=1}^{N_k}$ producing the approximation (7):

$$p(\mathbf{x}_k|\mathbf{z}_{0:k-1}) \approx \{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k} \quad (7)$$

We suggest to extract the proposal set $\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k}$ from the set of proposals produced by the Faster R-CNN fine-tuned by $\{\mathbf{x}_{k-1}^{(n)}\}_{n=1}^{N_{k-1}}$ (the target set at iteration $k-1$):

$$\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k} = \Theta(\{\mathbf{I}^{(i)}\}_{i=1}^{I_i}; \mathcal{R}_{k-1}, \mathcal{F}_{k-1}) \quad (8)$$

with a first iteration ($k = 1$) that uses an initial generic network $(\mathcal{R}_0, \mathcal{F}_0)$.

3.1.2. Update step

This step defines the likelihood term (9) by utilizing a likelihood function. This latter assigns a weight $\tilde{\pi}$ to each proposal sample $\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k}$ returned by the Faster R-CNN at the prediction step.

$$p(\mathbf{z}_k|\mathbf{x}_k = \tilde{\mathbf{x}}_k^n) \propto \tilde{\pi}_k^n \quad (9)$$

The likelihood function employs visual contextual cues extracted from the target video sequence and the confidence scores given by the output layer of the Faster R-CNN, to attribute a weight for each sample. More details about the likelihood function are given in section 3.2. The update step gives as an output a set of weighted target samples, which will be referred to as "the weighted target dataset" hereafter (10):

$$\{(\tilde{\mathbf{x}}_k^{(n)}, \tilde{\pi}_k^{(n)})\}_{n=1}^{\tilde{N}_k} \quad (10)$$

where $\{\tilde{\mathbf{x}}_k^{(n)}, \tilde{\pi}_k^{(n)}\}$ represents a target sample with its associated weight and $\tilde{N}_k$ is the number of weighted samples.

3.1.3. Sampling step

The aim of this last recursive-filter step is to build a new specialized dataset by deciding, according to the strategy of sampling (defined in the contribution), which samples will be included in the produced dataset $\mathcal{D}_k = \{\mathbf{x}_k^{(n)}\}_{n=1}^{N_k}$ at the iteration k from the weighted dataset $\{\tilde{\mathbf{x}}_k^{(n)}, \tilde{\pi}_k^{(n)}\}_{n=1}^{\tilde{N}_k}$. A sampling strategy is applied in order to generate a new unweighted dataset which has the same number of samples as the weighted one. To do this, we apply the IR algorithm, according to equation (11):

$$\mathcal{D}_k = \{\mathbf{x}_k^{(n)}\}_{n=1}^{N_k} = IR\left(\{\tilde{\mathbf{x}}_k^{(n)}, \tilde{\pi}_k^{(n)}\}_{n=1}^{\tilde{N}_k}\right) \quad (11)$$

This step generates a new set $\mathcal{D}_k$ by drawing samples according to the weight $\tilde{\pi}_k^{(n)}$

*3.2. **Likelihood function***

In order to choose the correct proposal, we put forward a likelihood function based on the verification strategy, which assigns a weight $\pi_k^{(n)}$ for each sample $\tilde{\mathbf{x}}_k^{(n)}$ returned by the prediction step. Our specifically designed likelihood function not only incorporates the confidence scores given by the output layer of the Faster R-CNN but also adds a spatial-temporal cues, to prioritize the selection of the correct samples and to reduce the risk of including wrong proposal samples in the specialized dataset.

Summarising the tests carried out on different databases, it is noticed that the generic Faster R-CNN is robust to generate true positive samples with a high score, and its selection of these ones will start to fail when the score of samples is lower than the score threshold α_k. For this reason, we keep the samples which have a confidence score greater than or equal to α_k and we propose an observation function f_L to assign a weight to each proposal sample that has a score lower than α_k, according to (12):

$$\pi_k^{(n)} = \begin{cases} s_k^{(n)} & \text{if} \quad s_k^{(n)} \geq \alpha_k \\ f_L(\tilde{\mathbf{x}}_k^{(n)}) & \text{if} \quad s_k^{(n)} < \alpha_k \end{cases} \tag{12}$$

Accordingly, we choose a dynamic threshold through iterations to avoid the problem of integrating negative samples into the specialized dataset. We are not limited to a fixed predefined threshold because the choice will be dynamic and will be related to the following equation (13):

$$\alpha_k = \begin{cases} \dfrac{\tilde{s}_k}{\tilde{s}_{k-1}}\,\alpha_{k-1} & \text{if} \quad k \neq 0 \\ \alpha_0 & \text{if} \quad k = 0 \end{cases} \tag{13}$$

where α_0 is the initial value of the score threshold (fixed to 0.5 for our experiments) and $\tilde{s}_k$ is the mean value of $s_k^{(n)}$ at iteration k :

Lower than α_k, the deep detector will start to fail and it will become unable to correctly select positive samples from a specific scene. To solve this problem, we propose an observation function f_L in order to favor the selection of positive samples using the information extracted from the target scene. This function is based on the visual spatio-temporal cue "Background extraction overlap score", to attribute a weight for each sample.

In a traffic scene, it is rare for a traffic object to stay fixed for a long time, and a good detection occurs on a foreground blob; whereas, false positive background detections give some Region of Interests (RoIs) that appear over time at the same location and with the same size.

To assign a weight for each sample, we calculate an overlap_score λ_o (equation 14) that compares the RoI associated to one sample with the output of a binary foreground extraction algorithm.

$$\lambda_o \doteq \frac{2(RoI\_AR \times FG\_AR)}{RoI\_AR + FG\_AR} \tag{14}$$

where RoI_AR is the area in pixels of the considered RoI and FG_AR is the foreground area at the RoI position (see Figure 3).

The background subtraction algorithm used in the proposed observation function is adopted from [32] and was called the "$BackgroundSubtractorMOG2$" algorithm. This latter is a Gaussian mixture-based background / Foreground segmentation algorithm.

One important property of this algorithm is that it chooses the appropriate number of Gaussian distribution for each pixel. It provides better adaptability to illumination changes. In our work, to ameliorate the result generated by the background subtraction algorithm mentioned above, we put forward some improvements such that:

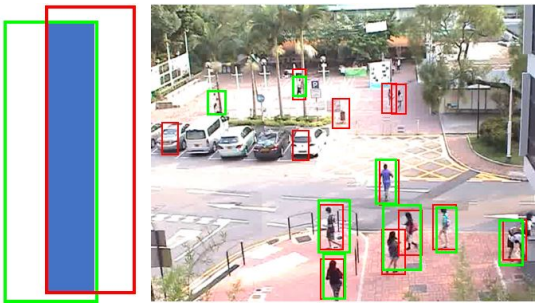


Figure 3: The red rectangle presents the area in pixels of the considered RoI, the green rectangle is the foreground area, and the rectangle filled in blue is the area of intersection.

- We apply several morphological filtering operations like erosion and dilation to the result of this algorithm so as to remove unwanted noise.
- We remove the blobs which have a surface area less than 100 pixels.

The observation function (Algorithm 1) will assign a high weight to a positive proposition if it has an overlap_score λ_o that exceeds a fixed threshold α_p, which is determined empirically.

Algorithm 1 Observation function

Input: Set $\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k}$ with associated RoI position $\{\mathbf{p}_k^{(i)}\}_{i=1}^{\tilde{N}_k}$ into the target video-sequence
Target video sequence $\mathcal{I}_t$
α_p: overlap threshold

Output: Set $\{\tilde{\pi}_k^{(i)}\}_{i=1}^{\tilde{N}_k}$ of weights associated to samples

for $i = 1$ to $\tilde{N}_k$ **do**
  $\tilde{\pi}_k^{(i)} \leftarrow 0$
  `/* Visual contextual cue computation */`
  $\lambda_o = \dfrac{2(RoI\_AR \times FG\_AR)}{RoI\_AR + FG\_AR}$
  `/* Weight assignment */`
  if $(\lambda_o \geqslant \alpha_p)$ **then**
    $\tilde{\pi}_k^{(i)} \leftarrow \lambda_o$
  end if
end for

Considering the likelihood function, the favoring of sample associated to the right label becomes efficient and easier.

3.3. *Fine-tuning step*

In the proposed framework, the aim of the fine-tuning step is to specialize the RPN and the Fast R-CNN deep networks to a specific scene. Accordingly, we use the target detection boxes included in the specialized dataset $\mathcal{D}_k$ and the RPN fine-tuning process mentioned in [9].

To do this, we use a sliding window approach to generate k bounding boxes for each position on the feature map produced

by the last convolutional layer, where each bounding box is centered on the sliding window and is associated with an aspect ratio and a scale (see Figure 4). The intersection-over-Union (IoU) overlap between each box of the specialized dataset $\mathcal{D}_k$ and the bounding boxes is then computed. A bounding box is designated as a positive training example if it has an IoU overlap greater than a predefined threshold with any $\mathcal{D}_k$ box, or if it is the bounding box that has the highest IoU with a $\mathcal{D}_k$ box. A proposal is designated as a negative example to a non-positive bounding box if its maximum IoU ratio with all boxes of the specialized dataset $\mathcal{D}_k$ is less than another predefined threshold. The bounding boxes that are neither positive nor negative do not contribute to the training.

Note that, the RPN fine-tuning process mentioned above does not consider that there might exist multiple copies (maximum twice) of the target detection box in the specialized dataset $\mathcal{D}_k$ because the main objective of using the IR algorithm proposed in the sampling strategy is not to increase the size of the database with samples which have high weights but to decrease the risk of distorting the specialized dataset $\mathcal{D}_k$ with wrong labelled examples because it is possible that the weighted target dataset contains wrong samples classified as traffic objects because their $\lambda_o >= \alpha_p$.

After training the RPN, these proposals are used to train the Fast R-CNN. Figure 4 illustrates the training strategy of the RPN fully-convolutional network.

Algorithm 2 SMC Faster R-CNN

Input: Generic network $(\mathcal{R}_0, \mathcal{F}_0)$
Number of iterations: K
Number of target samples $\tilde{N}_k$
Unweighted target dataset: $\tilde{\mathcal{W}}_k$
Target video sequence $\mathcal{I}_t$
Output: Specialized network $(\mathcal{R}_K, \mathcal{F}_K)$
Specialized dataset $\mathcal{D}_K$

for k=1,...,K **do**
/* Prediction step */
$\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k} = \Theta(\{\mathbf{I}^{(i)}\}_{i=1}^{I_i}; \mathcal{R}_{k-1}, \mathcal{F}_{k-1})$
/* Update step */
$\tilde{\mathcal{W}}_k = \{\tilde{\mathbf{x}}_k^{(n)}, \tilde{\pi}_k^{(n)}\}_{n=1}^{\tilde{N}_k}$
/* Sampling step */
for $n = 1$ to $\tilde{N}_k$ **do**
Draw a sample: $\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k}$, according to the weight $\tilde{\pi}_k^{(n)}$
end for
/* Fine-tuning step */
$\{\mathcal{R}_k, \mathcal{F}_k\} = f(\mathcal{I}_t, \mathcal{D}_k; \mathcal{R}_{k-1}, \mathcal{F}_{k-1})$
end for
$\{\mathcal{R}_K, \mathcal{F}_K\} = (\mathcal{R}_k, \mathcal{F}_k)$

Therefore, a new specialized RPN network and the Fast R-CNN one are generated being fine-tuning with the specialized dataset. These networks will become the input of the prediction step in the next iteration and will generate new object proposals (bounding boxes) in the target scene.

$$\{\mathcal{R}_k, \mathcal{F}_k\} = f(\mathcal{I}_t, \mathcal{D}_k; \mathcal{R}_{k-1}, \mathcal{F}_{k-1}) \quad (15)$$

The suggested SMC Faster R-CNN framework is summarized in Algorithm 2.

4. Experimental results

This section presents the experiments that have been achieved in order to compare the SMC Faster R-CNN with the relevant frameworks on several public and private datasets for single and multi-traffic object detection.

4.1. Implementation details

We describe the implementation details of the SMC Faster R-CNN algorithm. We use the pre-trained VGG16 model [33] to initialize the Faster R-CNN network, which is used in most recent state-of-the-art approaches [28][29].

Both RPN and Fast R-CNN are fine-tuned end-to-end by back-propagation and stochastic gradient descent [34] with a weight decay of 0.0005 and a momentum of 0.9. We use the alternating training algorithm [9] for Faster R-CNN training from one iteration to another. The Faster R-CNN is fine-tuned on a NVIDIA GeForce GTX TITAN X GPU with a 12GB memory.

Following multiple experiments, we chose 9 as the number of bounding boxes (3 aspect ratios [2:1, 1:1, 1:2] and 3 scales [128^2, 256^2, 512^2]) generated on each position of the sliding windows. We also chose 0.7 as the threshold of the IoU to select the positive samples and 0.3 for the negatives to build the training dataset.

The parameter K (number of iterations of the SMC process) is fixed to $K = 2$. Figure 6 shows that the specialization converges after two iterations for both car and pedestrian detection applied on the MIT Traffic dataset (introduced in the next section).

4.2. Datasets

The PASCAL VOC 2007 dataset [35] was utilized to learn the generic Faster R-CNN. This dataset consists of about 5,011 trainval images and 4,952 test ones over 20 object categories. In our experiments, we use only 713 annotated cars, 2,008 pedestrian, 186 buses and 245 for motorbikes, to fine-tune the generic Faster R-CNN. The evaluation is achieved on three target datasets (two public ones and a private one):

- **CUHK Square dataset [16]:** This is a video sequence of road traffic which lasts 60 minutes. 352 images are utilized for specialization, uniformly extracted from the first half of the video. 100 images are used for the test, extracted from the latest 30 minutes. Annotations were provided by Wang [16] for pedestrian detection (called **CUHK_WP** after). However, we notice that some annotation errors are made in the public ground truth and we suggest a new annotation (called **CUHK_MP** after) (see Figure 8.a).
- **MIT Traffic dataset [4]:** This is a 90-minute video. We use 420 images from the first 45 minutes for specialization. 100 images are uniformly sampled from the last 45 minutes for the test. Annotations are available for

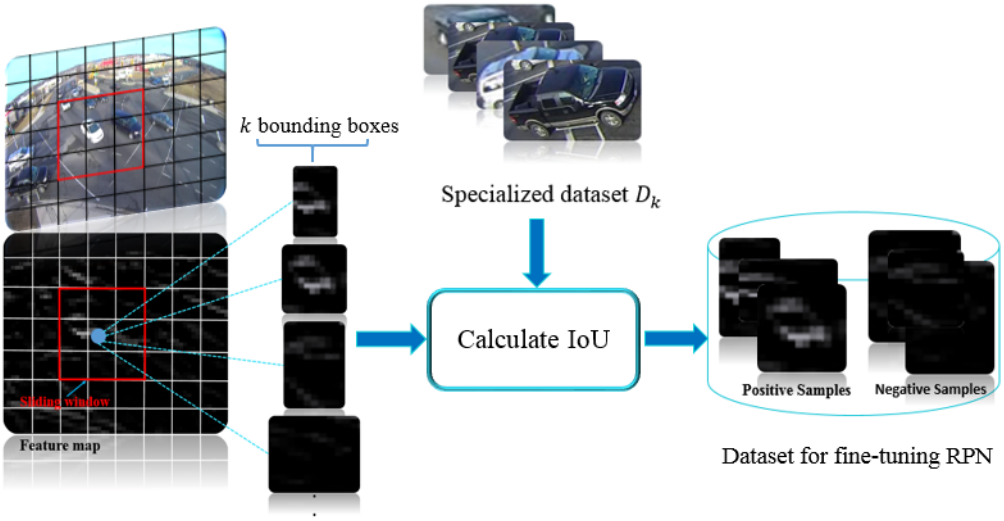


Figure 4: Description of training strategy for the RPN fully-convolutional network

pedestrians [4] (called **MIT_WP**) and cars [20] (called **MIT_LV**). Since some annotation errors are present, we propose new annotations (called **MIT_MV**) (see Figure 8.b).

- **Logiroad Traffic dataset:** This is a private video sequence of road traffic which lasts 20 minutes. We utilize 600 images for specialization, extracted uniformly from the first 15 minutes of the video. 100 images are used for the test, extracted from the latest 5 minutes. Annotations are available for vehicles (called **Logiroad_MV**).

4.3. Descriptions of experiments

Evaluation is performed in terms of recall False Positives Per Image (FPPI) curves. The PASCAL 50 percent overlap criterion [35] was utilized to give a score for the detection bounding boxes. The SMC Faster R-CNN framework is compared with several state-of-the-art frameworks:

- Generic Faster R-CNN: It is a detector fine-tuned on the generic dataset. This is the baseline for our comparison.
- Maamatou (2016) [5]: An SMC framework was applied to specialize a generic HOG-SVM classifier to a particular video sequence for traffic object detection.
- Xudong Li (2015) [20]: A deep learning domain adaptation framework was proposed for vehicle detection with manually annotated data from the target scene. Unlike other methods, the latter was not totally automatic and requires some manual annotations.
- Mao (2015) [11]: A framework was suggested to automatically train scene-specific pedestrian detectors based on tracklets.
- Htike (2014) [7]: A non-iterative domain adaptation framework was used to adapt a pedestrian detector to video scenes.
- Zeng (2014) [24]: A deep learning domain adaptation framework was proposed to automatically select training samples from target scenes without manual labelling for pedestrian detection.
- Wang (2014) [6]: A specific-scene detector was trained on only relevant samples collected from both source and target datasets.
- Nair (2004) [36]: An iterative self-training framework for detector adaptation was opted for using a background subtraction algorithm.

4.4. Results and analysis for single-traffic object

Given each dataset and its annotation, we present the ROC curves (Figure 5) of the generic Faster R-CNN, the SMC Faster R-CNN and the available state-of-the-art frameworks. The ROC curves present the comparison between the true detection rate and the false positive detection rate per image. Furthermore, we give two comparative synthetic tables: one for pedestrian detection (cf. Table 2) and the other for vehicle detection (cf. Table 3). In addition, on the last line of both tables, the improvement between the generic Faster R-CNN and the SMC Faster R-CNN is given.

- **Comparison with generic detector:** Figure 5 shows that the specialized Faster R-CNN detector significantly outperforms the generic one on all public and private datasets with several annotations. The median improvement is 51%.
- **Comparison with state-of-the-art:** According to the ROC curves at the top of Figure 5, for the CUHK pedestrian detection, the SMC Faster R-CNN outperforms all other state-of-the-art frameworks. Besides, the detection rate achieved with our proposed annotations on **CUHK_MP** is nearly 90% for 0.5 FPPI. However, despite of the wrong annotations given by Wang (left curve

in the top of Figure 5), the SMC Faster R-CNN also exceeds the six other specialized detectors of Nair (2004), Wang (2014), Zeng (2014), Htike (2014), Mao (2015) and Maamatou (2016) respectively by 24%, 45%, 53%, 49%, 58% and 62%.

For the MIT pedestrian detection (**MIT_WP** in Table 2), the specialized deep detector proposed by Zeng (2014) exceeds the SMC Faster R-CNN detector for an 0.5 FPPI, which is less than 0.9.

Despite the wrong annotations given by Li *et al.* [20], Figure 5 (right curve in the middle) shows that for the MIT car detection (**MIT_LV**), the proposed SMC Faster R-CNN clearly outperforms the specialized CNN detector proposed by Li (2015) which trained with manual data labeling from the target scene. According to Table 3, for the MIT and Logiroad car detection with the proposed annotations, the SMC Faster R-CNN is ranked first and exceeds the specialized detector suggested by Maamatou (2016).

One can notice that the generic Faster R-CNN, fine-tuned on the PASCAL VOC 2007 dataset, has a poor detection rate resulting in a limitation of the size of the specialized dataset.

Table 2: Comparison of detection rate for pedestrian with state of the art (at 0.5 FPPI)

Dataset / Approach	CUHK_WP	CUHK_MP	MIT_WP
Nair [36]	0.24	–	0.35
Wang [6]	0.45	–	0.42
Zeng [24]	0.53	–	0.58
Htike [7]	0.49	–	–
MAO [11]	0.58	–	–
Maamatou [5]	0.62	0.58	0.40
Generic Faster R-CNN [9]	0.60	0.69	0.07
SMC Faster R-CNN	**0.65**	**0.88**	**0.47**
Improvement / generic (%)	8%	28%	571%

Table 3: Comparison of detection rate for car with state of the art (at 1 FPPI)

Dataset / Approach	MIT_LV	MIT_MV	Logiroad_MV
Li [20]	0.77	–	–
Maamatou [5]	–	0.29	0.47
Generic Faster R-CNN [9]	0.68	0.38	0.40
SMC Faster R-CNN	**0.77**	**0.80**	**0.70**
Improvement / generic (%)	13%	110%	75%

- **Effect of likelihood function:** To show the effectiveness of our likelihood function, the ROC curves in Figure 7 show the comparison between using the likelihood function based only on confidence score predicted by the Faster R-CNN and our proposed one on two datasets.

  The red curves in Figure 7 present our proposed likelihood function based on the combination between the confidence score and the spatio-temporal cue, and the blue ones indicate the use of the confidence score only, which is given by the output layer of the Faster R-CNN. The results demonstrate that the proposed likelihood function based on using the verification strategy improves the detector performance and accelerates the convergence of the specialization process. Furthermore, we cannot say that this choice is the best because it is possible to ameliorate the suggested framework by proposing other strategies for the SMC steps. For example, we can improve the likelihood function with more complex visual cues like tracking, optical flow or contextual information to enhance the weighting of positive samples.

*4.5. **Results and analysis for multi-traffic object***

We evaluate the proposed approach for multi-traffic objects on two datasets, the MIT Traffic dataset and the Logiroad one using two evaluation criteria: namely the ROC curves and the confusion matrix (classical metrics for object detection).

For the MIT Traffic dataset, we select 2 classes {'pedestrian', 'car'} and 4 classes for the Logiroad Traffic dataset {'pedestrian', 'car', 'bus', 'motorbike'}.

The results are reported in Table 4. The SMC Faster R-CNN presents a median improvement of 89% related to the generic detector. Moreover, Tables 5 and 6 provide the associated similarity matrix. We show that some confusion may occur between motorbikes and cars or between buses and cars. Furthermore, these results illustrate that our framework has a robust performance for multi-traffic object detection. This indicates that it is useful to run our specialization algorithm whenever we have a new sequence and we want to automatically generate a much better deep detector than the generic one.

5. Conclusion and future work

We have put forward an efficient framework based on the formalism of the SMC filter to specialize the Faster R-CNN deep detector for multi-traffic object detection. This framework approximates the unknown target distribution by selecting relevant samples from target datasets. These samples are utilized to fine-tune a specialized deep detector in order to decrease the detection rate in the target scene. Given a generic detector and a target video sequence, this framework automatically provides a robust specialized detector. Moreover, the proposed framework allows reducing the risk of drifting by using efficient strategies during iterations and it can be used to specialize any deep detector. The extensive experiments have demonstrated that the suggested framework has produced a specialized detector that performs much better than the generic one for both single and multi-traffic object detections in different scenes. Furthermore, the results show that the framework outperforms the state-of-the-art specialization ones on several challenging datasets. Our future work will deal with an extension of the algorithm to improve the likelihood function by using a new strategy of verification based on more complex visual cues like tracking, optical flow, tracklets or contextual

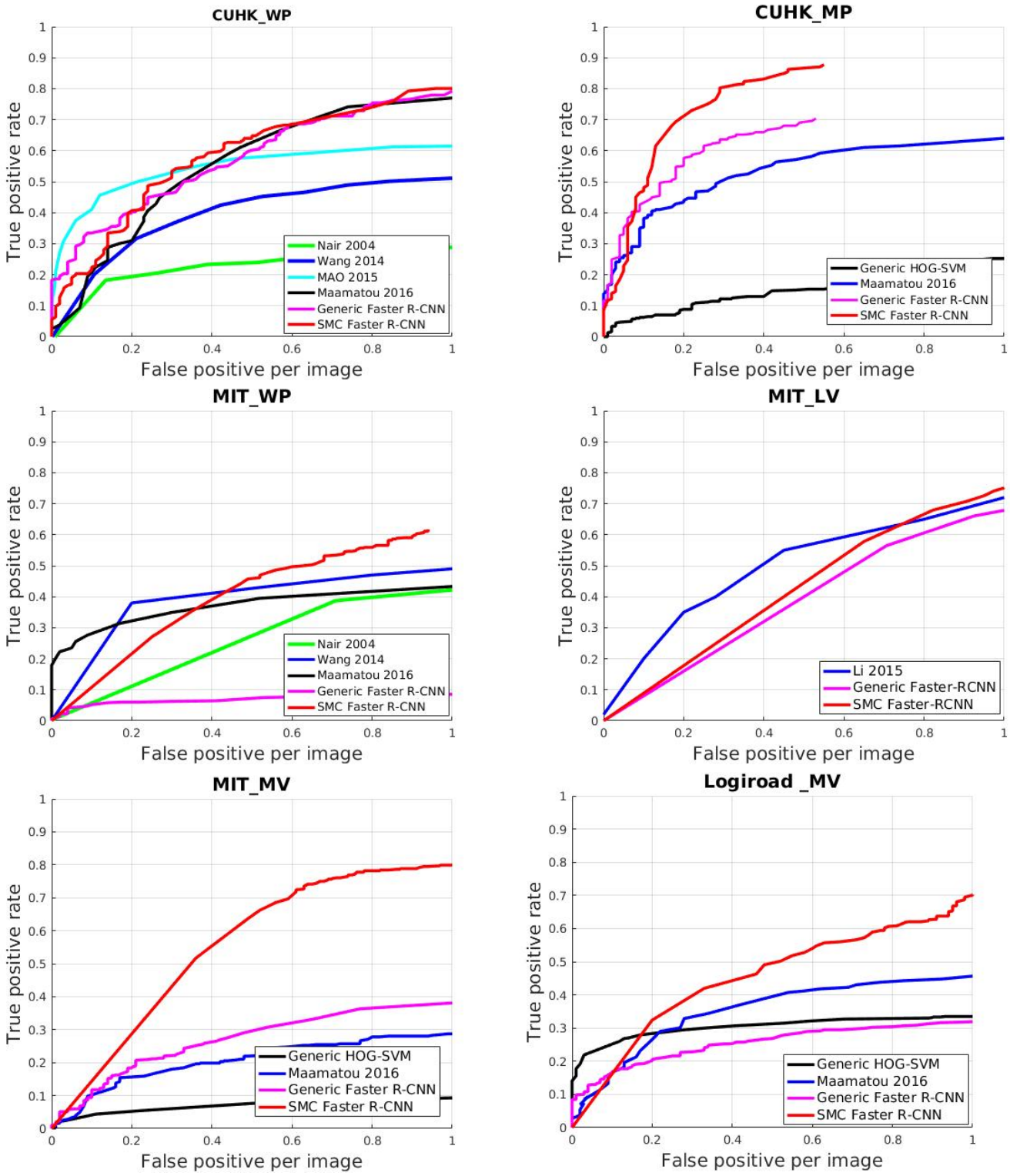


Figure 5: ROC curves for several datasets and annotations

Table 4: Detection rate for multi-traffic object detection with SMC Faster R-CNN (at 1 FPPI)

Dataset / Approach	Logiroad_Car	Logiroad_Person	Logiroad_Moto	MIT_Car	MIT_Person
Generic Faster R-CNN	0.28	0.24	0,065	0.32	0.05
SMC Faster R-CNN	**0.60**	**0.36**	**0.18**	**0.73**	**0.30**
Improvement/ generic(%)	114%	50%	176%	128%	500%

Table 5: Illustration of similarity matrix between traffic object categories on Logiroad Traffic dataset (diagonal row shows the accuracy to recognize traffic objects of its own class)

Predicted class / Actual class	Pedestrian	Car	Motorbike	Bus
Pedestrian	140/**97%**	12/1.5%	5/14%	0
Car	0	750/**96%**	1/3%	1/2.5
Motorbike	5/3%	12/1.5%	30/**83%**	1/2.5%
Bus	0	7/1%	0	38/**95%**
Total	145	781	36	40

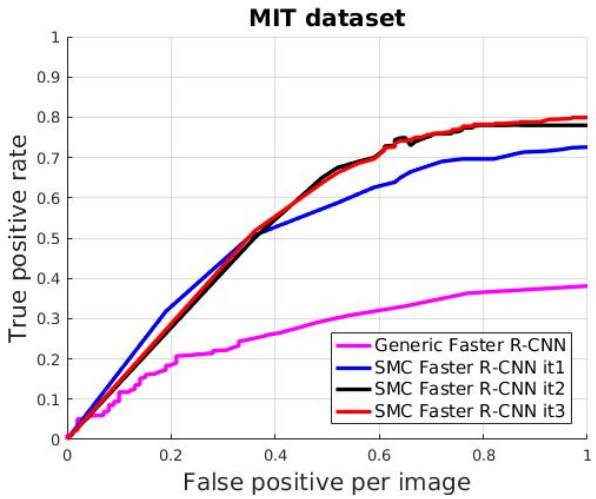


Figure 6: ROC curves for convergence of specialization process

Table 6: Illustration of similarity matrix between traffic object categories on MIT Traffic dataset

Predicted class / Actual class	Pedestrian	Car
Pedestrian	342/**99.7%**	7/1.6%
Car	1/0.3%	420/**98.4%**
Total	343	427

information and injecting some spatio-temporal information into the Faster R-CNN network.

Acknowledgment

This work is within the scope of a co-guardianship between the university of Sousse (Tunisia) and Clermont Auvergne University (France). It is sponsored by the Tunisian Ministry of Higher Education & Scientific Research and the French government research program "Investissements d'avenir" through the IMobS3 Laboratory of Excellence (ANR-10-LABX-16-01), by the European Union through the program Regional competitiveness and employment 2007-2013 (ERDF - Auvergne region), and by the Auvergne region.

References

References

[1] X. Pan, Y. Guo, A. Men, Traffic surveillance system for vehicle flow detection, in: Computer Modeling and Simulation, 2010. ICCMS'10. Second International Conference on, Vol. 1, IEEE, 2010, pp. 314–318.

[2] B. Benfold, I. Reid, Stable multi-target tracking in real-time surveillance video, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 3457–3464.

[3] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 304–311.

[4] X. Wang, X. Ma, W. E. L. Grimson, Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models, PAMI (2009) 539–555.

[5] H. Maâmatou, T. Chateau, S. Gazzah, Y. Goyat, N. Essoukri Ben Amara, Transductive transfer learning to specialize a generic classifier towards a specific scene, in: VISAPP, 2016, pp. 411–422.

[6] X. Wang, M. Wang, W. Li, Scene-specific pedestrian detection for static video surveillance, PAMI (2014) 361–362.

[7] K. K. Htike, D. C. Hogg, Efficient non-iterative domain adaptation of pedestrian detectors to video scenes, in: 2014 22nd International Conference on Pattern Recognition (ICPR), IEEE, 2014, pp. 654–659.

[8] A. Smith, A. Doucet, N. de Freitas, N. Gordon, Sequential Monte Carlo methods in practice, Springer Science & Business Media, 2013.

[9] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[10] A. Doucet, N. De Freitas, N. Gordon, Sequential Monte Carlo Methods in Practice, Springer, 2001.

[11] Y. Mao, Z. Yin, Training a scene-specific pedestrian detector using tracklets, in: Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, IEEE, 2015, pp. 170–176.

[12] T. Tommasi, F. Orabona, B. Caputo, Learning categories from few examples with multi model knowledge transfer, IEEE transactions on pattern analysis and machine intelligence 36 (5) (2014) 928–941.

[13] Y. Aytar, A. Zisserman, Tabula rasa: Model transfer for object category detection, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2252–2259.

[14] S. J. Pan, I. W. Tsang, J. T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Transactions on Neural Networks 22 (2) (2011) 199–210.

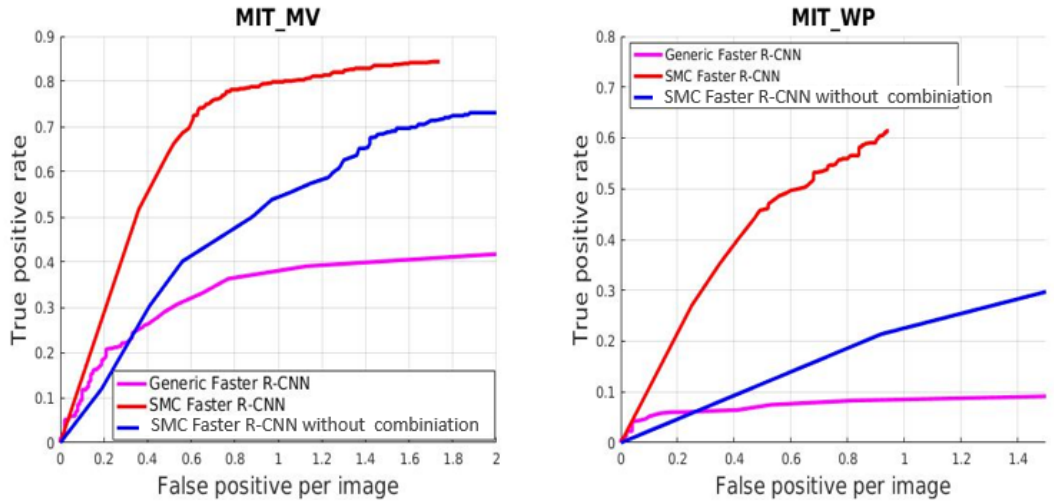
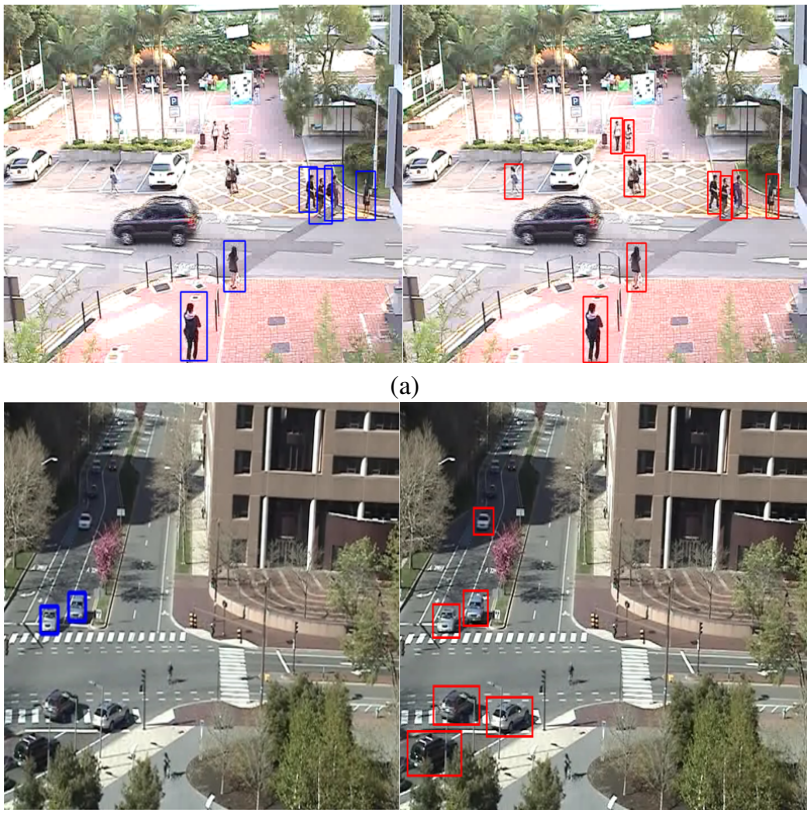


Figure 7: Effect of the likelihood function in our specialization framework on MIT Traffic dataset for both pedestrian and car detections.



(a)

(b)

Figure 8: Some annotations provided by Wang [6] ground truth on CUHK dataset ((a) left image), Li [20] ground truth on MIT Traffic dataset ((b) left image) and our annotations ((a), (b) right images). There are several missing objects in the baseline annotations: This is why we propose an updated version.

[15] B. Quanz, J. Huan, M. Mishra, Knowledge transfer with low-quality data: A feature extraction issue, IEEE Transactions on Knowledge and Data Engineering 24 (10) (2012) 1789–1802.
[16] M. Wang, W. Li, X. Wang, Transferring a generic pedestrian detector towards specific scenes, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 3274–3281.
[17] C. Rosenberg, M. Hebert, H. Schneiderman, Semi-supervised self-training of object detection models.
[18] A. Levin, P. Viola, Y. Freund, Unsupervised improvement of visual detectors using cotraining, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE, 2003, pp. 626–633.
[19] K. All, D. Hasler, F. Fleuret, Flowboostappearance learning from sparsely annotated video, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1433–1440.
[20] X. Li, M. Ye, M. Fu, P. Xu, T. Li, Domain adaption of vehicle detector based on convolutional neural networks, International Journal of Control, Automation and Systems 13 (4) (2015) 1020–1031.
[21] I. Guyon, G. Dror, V. Lemaire, G. Taylor, D. W. Aha, Unsupervised and transfer learning challenge, in: IJCNN, IEEE, 2011, pp. 793–800.
[22] G. Mesnil, Y. Dauphin, X. Glorot, S. Rifai, Y. Bengio, I. J. Goodfellow, E. Lavoie, X. Muller, G. Desjardins, D. Warde-Farley, et al., Unsupervised and transfer learning challenge: a deep learning approach., ICML Unsupervised and Transfer Learning (2012) 97–110.
[23] I. J. Goodfellow, A. Courville, Y. Bengio, Spike-and-slab sparse coding for unsupervised feature discovery, arXiv.
[24] X. Zeng, W. Ouyang, M. Wang, X. Wang, Deep learning of scene-specific classifier for pedestrian detection, in: ECCV, Springer, 2014, pp. 472–487.
[25] H.-W. Ng, V. D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 443–449.
[26] X. Glorot, A. Bordes, Y. Bengio, Domain adaptation for large-scale sentiment classification: A deep learning approach, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 513–520.
[27] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1717–1724.
[28] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[29] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[30] X. Mei, H. Ling, Robust visual tracking and vehicle classification via sparse representation, PAMI 33 (11) (2011) 2259–2272.
[31] I. Smal, W. Niessen, E. Meijering, Advanced particle filtering for multiple object tracking in dynamic fluorescence microscopy images, in: BIFNM, IEEE, 2007, pp. 1048–1051.
[32] Z. Zivkovic, F. Van Der Heijden, Efficient adaptive density estimation per image pixel for the task of background subtraction, Pattern recognition letters 27 (7) (2006) 773–780.
[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.
[34] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural computation 1 (4) (1989) 541–551.
[35] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2) (2010) 303–338.
[36] V. Nair, J. J. Clark, An unsupervised, online learning framework for moving object detection, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, Vol. 2, IEEE, 2004, pp. II–II.