Specialization of a Generic Pedestrian Detector to a Specific Traffic Scene by the Sequential Monte-Carlo Filter and the Faster R-CNN

Ala Mhalla[1,2], Thierry Chateau[2], Sami Gazzah[1] and Najoua Essoukri Ben Amara[1]
[1]*LATIS ENISo, University of Sousse, Sousse, Tunisia*
[2]*Institut Pascal, Blaise Pascal University, Clermont Ferrand, France*

Keywords: Transfer Learning, Deep Learning, Faster R-CNN, Sequential Monte Carlo Filter (SMC), Pedestrian Detection.

Abstract: The performance of a generic pedestrian detector decreases significantly when it is applied to a specific scene due to the large variation between the source dataset used to train the generic detector and samples in the target scene. In this paper, we suggest a new approach to automatically specialize a scene-specific pedestrian detector starting with a generic detector in video surveillance without further manually labeling any samples under a novel transfer learning framework. The main idea is to consider a deep detector as a function that generates realizations from the probability distribution of the pedestrian to be detected in the target. Our contribution is to approximate this target probability distribution with a set of samples and an associated specialized deep detector estimated in a sequential Monte Carlo filter framework. The effectiveness of the proposed framework is demonstrated through experiments on two public surveillance datasets. Compared with a generic pedestrian detector and the state-of-the-art methods, our proposed framework presents encouraging results.

1 INTRODUCTION

Pedestrian detection is always a complicated and uncertain area in many important applications of computer vision due to several factors like different poses, illumination variations, image resolution and camera angle. However, most pedestrian detectors are learnt with generic labeled datasets that are sampled from a large number of situations to cover the maximum variability of the pedestrian. When tested on a specific scene, the performance of such detector drops significantly due to the large variations between the target scene and the source training dataset. We can solve this problem by transfer learning, which can specialize a generic pedestrian detector to a specific scene. Much of the state-of-the-art research have been made to develop scene-specific pedestrian detectors, whose training process is aided by generic detectors for automatically collecting training samples from target scenes (Maâmatou et al., 2016), (Wang et al., 2014).

The aim of this paper is to automatically generate a specialized pedestrian detector trained for a target video, hence performing better than the generic detector.

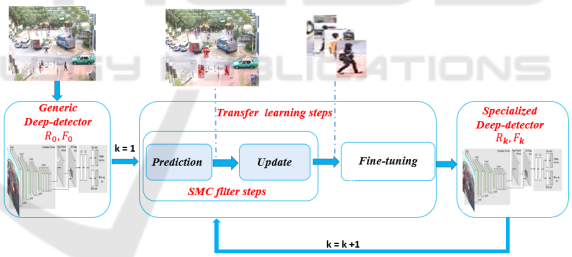


Figure 1: Block diagram of the proposed approach.

1.1 Motivations and Contributions

We propose a novel framework which starts from a generic pedestrian detector to train a scene-specific detector automatically. Our main motivations and contributions are summarized below.

- Transfer learning method to specialize generic pedestrian detector into a target video scene
- A likelihood function that used to select the positive samples and remove the negative ones from a specific scene
- Evaluating our proposed approach to the state of the art on two public datasets

The rest of the paper is organized as follows. The

Mhalla A., Chateau T., Gazzah S. and Essoukri Ben Amara N.
Specialization of a Generic Pedestrian Detector to a Specific Traffic Scene by the Sequential Monte-Carlo Filter and the Faster R-CNN.
DOI: 10.5220/0006097900170023
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 17-23
ISBN: 978-989-758-225-7

related literature is reviewed in section 2, followed by a detailed description of our approach in section 3. We describe the experiments and results in section 4 and finally conclude in section 5.

2 RELATED WORK

In this section, firstly, we describe the related works for transfer learning and deep learning, then we present the Faster R-CNN deep network (Ren et al., 2015) and the SMC scene specialization algorithm (Maâmatou et al., 2016) which are two main references for our paper.

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 domain. Over the past decades, numerous methods have been suggested for transfer learning in pedestrian detection (Maâmatou et al., 2016), (Wang et al., 2014). Recently, addressing this problem with deep neuronal networks has gained an increased attention. Some deep models have been investigated in the unsupervised and transfer learning challenge (Guyon et al., 2011). Transfer learning using deep models has been turned out to be effective in some challenges (Mesnil et al., 2012), (Goodfellow et al., 2012) like traffic-object detection (Zeng et al., 2014), (Li et al., 2015) and sentiment analysis (Glorot et al., 2011).

In addition, deep learning has led to great performance on a variety problems of computer vision like vehicle detection (Li et al., 2015), action recognition (Will Y. Zou, 2011), face recognition (Huang et al., 2012), (Taigman et al., 2014) and image classification (Duan et al., 2009). Among various types of deep neural networks, convolutional neuronal networks (Jia et al., 2014), (LeCun et al., 1998) have been proved to make great successes in machine learning and computer vision applications.

In this paper, we propose a new framework based on SMC filter to specialize the recent deep detector, the Faster R-CNN (Ren et al., 2015) for pedestrian detection.

The Faster R-CNN (Ren et al., 2015) has been put forward to accurately detect general objects in pictures. It has achieved a state-of-the-art 73.2 mean average precision on the PASCAL VOC 2007 dataset (Everingham et al., 2010). By using both region-proposal network for localization task and detector network together into one large network.

The Faster R-CNN 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, and the second one was the Fast R-CNN network which took as inputs this set of object proposals and then used them for detection. The entire system was a single, unified network for object detection.

In this paper, the proposed method is developed based on the SMC framewoek (Maâmatou et al., 2016) due to its superb efficiency and performance in traffic object detection. Maamatou *et al.* (Maâmatou et al., 2016) put forward a transductive transfer learning method based on an SMC filter to iteratively build a new specialized dataset that was used to train a new specialized pedestrian detector. This new produced dataset is composed of both source and target samples that estimated the unknown target distribution. The specialization algorithm was applied on a HOG-SVM detector. The general framework presented in this paper is inspired from this work. We propose a transfer learning framework based on the SMC filter to specialize the Faster R-CNN detector to a target scene. The specialization framework presented in this paper proposes various differences over the SMC framework proposed in (Maâmatou et al., 2016), we cite mainly: we remove the sampling step of the SMC filter and keeping only the prediction and the update steps. The aim of this improvement is to optimize the specialization chain.

In this section, we purport an approach based on the SMC filter to automatically specialize the Faster R-CNN deep model to a target scene for pedestrian detection. The block diagram of our suggested specialized Faster R-CNN is given in Fig.1. At the first iteration, we fine-tune an public ImageNet-pre-trained model (VGG 16) (Simonyan and Zisserman, 2014) to the Pascal VOC dataset to create a generic pedestrian detector. This latter is utilized in the first step of the SMC ”Prediction” to suggest samples from the target scene and then we apply the likelihood function in the update step to correctly select weight samples from a specific scene and determine the relevant ones for the specialization process. A new specialized detector is fine-tuned by the specialized dataset in the fine-tuning step and it will become the input of the prediction step in the next iteration. The prediction, update and fine-tuning steps are called until a stopping criterion is reached, for example a fixed number of iterations.

In what follows, we first describe the adaptation of the two SMC steps with the Faster R-CNN model, and then deal with the fine-tuning step.

3 PROPOSED APPROACH

In this section, we present our method for specializing a pedestrian detector to a target scene.

In this section, we present the general framework for specializing a pedestrian detector to a target scene based on the Faster R-CNN model and the SMC filter.

Let us define:

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

We assume that the specialized dataset $\mathcal{D}_k$ approximates the probability distribution $p(\mathbf{x}_k|\mathbf{z}_k)$ of the pedestrian given by the target video sequence (1):

$$p(\mathbf{x}_k|\mathbf{z}_k) \approx \{\mathbf{x}_k^{(n)}\}_{n=1}^{N_k} \tag{1}$$

where $\mathbf{x}_k$ is a state vector that defines the target pedestrian class and $\mathbf{z}_k$ is an observation vector provided by the target video sequence (ie: visual spatio-temporal information).

We propose to estimate the current target distribution from the previous one with the sequential Bayesian filter (2):

$$p(\mathbf{x}_k|\mathbf{z}_{0:k}) = C.p(\mathbf{z}_k|\mathbf{x}_k) \int_{\mathbf{x}_{k-1}} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{0:k-1}) \tag{2}$$

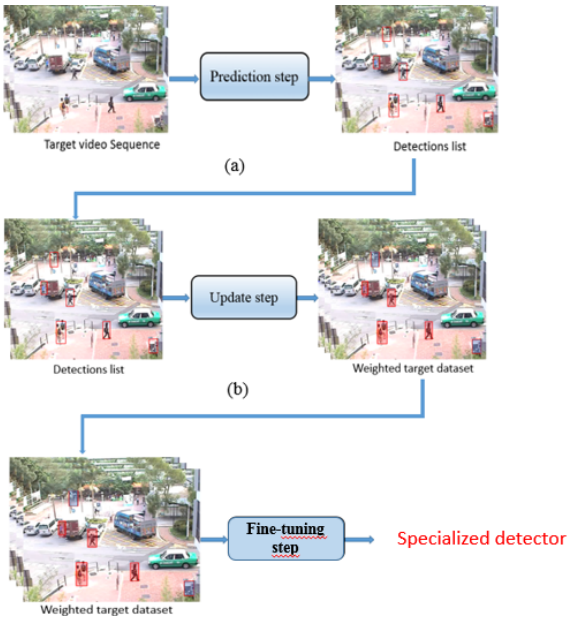


Figure 2: Specialization steps.

where C is a normalisation factor. We assume that this recursive framework converges toward the true target distribution. The resolution of eq. 2 is divided into two steps: prediction and update . These steps are similar to the popular particle filter framework, widely used to solve tracking problems in computer vision (Smith et al., 2013). The details of the two main steps of SMC filter are described in the following subsections.

3.1 Prediction Step

The prediction step consists in applying the Chapman-Kolmogorov equation (3):

$$p(\mathbf{x}_k|\mathbf{z}_{0:k-1}) = \int_{\mathbf{x}_{k-1}} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{0:k-1}) d\mathbf{x}_{k-1} \tag{3}$$

This step uses the system dynamics term $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ between two iterations to produce approximation (4):

$$p(\mathbf{x}_k|\mathbf{z}_{0:k-1}) \approx \{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k} \tag{4}$$

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 (see Fig.2.a) fine-tuned by $\{\mathbf{x}_k^{(n)}\}_{n=1}^{N_{k-1}}$ (the target set at iteration $k-1$):

$$\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k} = \Theta(\{I^{(i)}\}_{i=1}^{I_i}; \mathcal{R}_{k-1}, \mathcal{F}_{k-1}) \tag{5}$$

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

3.2 Update Step

A weight $\tilde{\pi}$ is estimated for each new target sample of the dataset $\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k}$ according to a likelihood function (6):

$$p(\mathbf{z}_k|\mathbf{x}_k = \tilde{\mathbf{x}}_k^n) \propto \tilde{\pi}_k^n \quad (6)$$

The likelihood function utilizes visual spatio-temporal information extracted from the target video sequences. The update step gives as an output a set of weighted target samples (see Fig.2.b), which will be referred to as "the weighted target dataset" hereafter (7):

$$\{(\tilde{\mathbf{x}}_k^{(n)}, \tilde{\pi}_k^{(n)})\}_{n=1}^{\tilde{N}_k} \quad (7)$$

where $\{\tilde{\mathbf{x}}_k^{(n)}, \tilde{\pi}_k^{(n)}\}$ represents a target sample and its associated weight.

Several likelihood functions can be proposed. More details are given in section 3.4

3.3 Fine-tuning Step

The fine-tuning step consists in training the RPN and the Fast R-CNN networks to generate a new specialized detector. The training of the RPN is inspired from (Ren et al., 2015). 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 Fig.3). The overlap between the boxes of the specialized dataset $\mathcal{D}_k$ and the bounding boxes is then computed. A proposal is designated as a positive training example if it overlaps with a specialized dataset $\mathcal{D}_k$ box having an Intersection-over-Union (IoU) greater than a predefined threshold, or if it is the bounding box that has the highest IoU with $\mathcal{D}_k$. A proposal is designated as a negative example if its maximum IoU with any box 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 objective. After training the RPN, these proposals are used to train the Fast R-CNN.

Therefore, a new specialized RPN network and the Fast R-CNN network are generated being fine-tuning with the specialized dataset. These networks generate new pedestrians (bounding boxes) in the target scene, according to (8):

$$\{\mathcal{R}_k, \mathcal{F}_k\} = f(I_t, \mathcal{D}_k; \mathcal{R}_{k-1}, \mathcal{F}_{k-1}) \quad (8)$$

The proposed specialized Faster R-CNN framework is summarized in Algorithm 1.

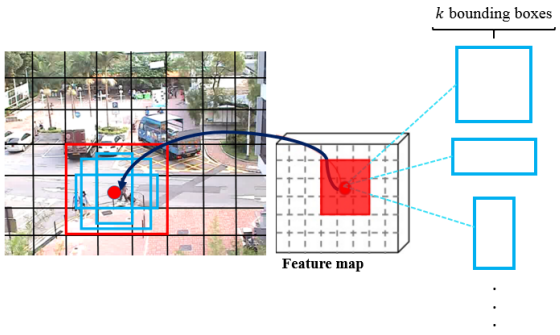


Figure 3: Description of fine-tuning step for RPN.

Algorithm 1: Transfer learning for specialization.

Input: Generic network $(\mathcal{R}_0, \mathcal{F}_0)$
Number of iterations: K
Likelihood function: $p(\mathbf{z}|\mathbf{x})$
Target video sequence : $\{I^{(i)}\}_{i=1}^{I}$
Output: Final Specialized Faster R-CNN $(\mathcal{R}_K, \mathcal{F}_K)$

for k=1,...,K **do**
/* Prediction step */
$\{\tilde{\mathbf{x}}_k^{(n)}\}_{n=1}^{\tilde{N}_k} = \Theta(\{I^{(i)}\}_{i=1}^{I_i}; \mathcal{R}_{k-1}, \mathcal{F}_{k-1})$
/* Update step */
$\tilde{\pi}_k^{(n)} = p(\mathbf{z}_k|\mathbf{x}_k = \tilde{\mathbf{x}}_k^{(n)})$
/* Fine-tuning step */
$\{\mathcal{R}_k, \mathcal{F}_k\} = f(I_t, \mathcal{D}_k; \mathcal{R}_{k-1}, \mathcal{F}_{k-1})$
end for

3.4 Likelihood Function

We put forward a likelihood function adapted from (Maâmatou et al., 2016) that assigns a weight $\pi_k^{(n)}$ for each sample $\tilde{\mathbf{x}}_k^{(n)}$ returned by the prediction step, according to (9):

$$\pi_k^{(n)} = \begin{cases} s_k^{(n)} & \text{if } s_k^{(n)} \geq \alpha_k \\ f_L(\tilde{\mathbf{x}}_k^{(n)}) & \text{if } s_k^{(n)} < \alpha_k \end{cases} \quad (9)$$

Where f_L is an observation function proposed by Maamatou *et al.*in (Maâmatou et al., 2016) and modified to be adapted to our work. It assumes that the pedestrians to be detected in the target video sequence have to be moving objects. The observation function uses visual contextual cues and prior information extracted from the target video sequence like background subtraction to select the positive samples. The output of this function is a set of weighted target samples that approximates the posterior probability function.

α_k is a score threshold that has to be fixed to 0.7 for our experiments and $s_k^{(n)}$ is the score generated by the Faster R-CNN for each sample $\tilde{\mathbf{x}}_k^{(n)}$.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of the proposed specialization framework with the relevant methods on several popular datasets including MIT (Wang et al., 2009) and CUHK (Wang et al., 2012) datasets.

4.1 Datasets

The PASCAL VOC 2007 dataset (Everingham et al., 2010) was utilized to learn the generic Faster. To do this, we used 2,008 annotated pedestrians.

The evaluation was achieved on two public datasets.

- **CUHK Square dataset (Wang et al., 2012):** This is a video sequence of road traffic which lasts 60 minutes. 352 images were used for specialization, uniformly extracted from the first half of the video. 100 images were utilized for the test, extracted from the latest 30 minutes. Annotations were provided by Wang (Wang et al., 2012) for pedestrian detection (called **CUHK** after). However, we have noticed that some annotation errors were made in the public ground truth and we have proposed a new annotation (called **CUHK_M** after).
- **MIT Traffic dataset (Wang et al., 2009):** This is a 90-minute video. We used 420 images from the first 45 minutes for specialization. 100 images were uniformly sampled from the last 45 minutes for the test. Annotations were available for pedestrians (Wang et al., 2009) (called **MIT**).

4.2 Implementation Details

We describe the implementation details of the SMC Faster R-CNN algorithm. The Faster R-CNN network consists of two networks, the RPN network for localization task and the Fast R-CNN network for detection task. The two networks are initialized with an ImageNet-pre-trained network (VGG16).

Following multiple experiments, we used the following parameters: 9 as the number 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, 0.7 as the threshold of 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 $K = 2$, the specialization converges after two iterations on two public datasets.

Table 1: Comparison of detection rate for pedestrian with state of the art (at 0.5 FPPI).

Approach	CUHK	CUHK_M	MIT
Nair (2004)	0.24	–	0.35
Wang (2014)	0.45	–	0.42
Htike (2014)	0.49	–	–
MAO (2015)	0.58	–	–
Maamatou (2016)	0.62	0.58	0.40
Generic Faster	0.60	0.69	0.07
Our approach	**0.63**	**0.80**	**0.44**
Improvement/gen	6%	22%	564%

4.3 Evaluated Algorithms

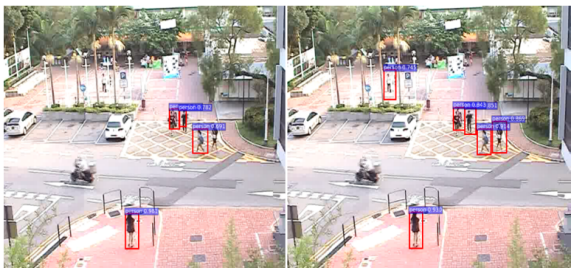
Evaluation was performed in terms of recall False Positives Per Image (FPPI) curves. The PASCAL 50 percent overlap criterion (Everingham et al., 2010) was used to give a score for the pedestrian bounding boxes. The SMC Faster R-CNN algorithm was compared with several state-of-the-art algorithms:

- Generic Faster R-CNN: It is a detector fine-tuned on the generic dataset. This is the baseline for our comparison.
- Maamatou (2016) (Maâmatou et al., 2016): An SMC framework was applied to specialize a generic HOG-SVM for pedestrian detection.
- Mao (2015) (Yunxiang Mao, 2015): an algorithm to automatically train scene-specific pedestrian detectors based on tracklets.
- Htike (2014) (Htike and Hogg, 2014): a non-iterative domain adaptation algorithm used to specialize a pedestrian detector to video scenes.
- Wang (2014) (Wang et al., 2014): a specific-scene detector trained on only relevant samples selected from both source and target datasets.
- Nair (2004) (Nair and Clark, 2004): an iterative self-training algorithm for detector adaptation using background subtraction.

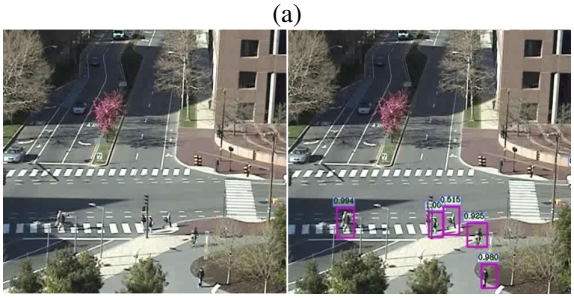
4.4 Results and Analysis

Given each dataset and its groud-truth, we give a comparative synthetic table for pedestrian detection (cf. Table 1). We compare the true detection rate for a constant false positive per image detection rate of several methods related to several datasets and annotations. Moreover, on the last line of table 1, the improvement between the generic Faster R-CNN pedestrian detector and the SMC Faster R-CNN is given.

The SMC Faster R-CNN has a higher detection rate than the generic Faster R-CNN in all the



(a)



(b)

Figure 4: (a), (b): improvement of our proposed specialized framework (the left images shows the detection results for the generic detector and the right for the specialized one) in CUHK and MIT datasets.

experiments (Fig.4). The median improvement is 22%.

For the CUHK pedestrian detection, the SMC Faster R-CNN outperforms all the other state-of-the-art algorithms. Besides, the detection rate achieved with our annotations is nearly 80% for 0.5 FPPI.

For the MIT pedestrian detection, the SMC Faster R-CNN is ranked first.

5 CONCLUSIONS

In this paper, we have proposed a specialization framework for pedestrian detection based on the sequential Monte Carlo filter and the Faster R-CNN deep model. Given a generic pedestrian detector and a target video sequence, our method automatically provides a specialized detector. Moreover, the experimental results show that the algorithm outperforms the generic detector and the state-of-the-art specialization approaches on two challenging datasets. Our future works will deal with an extension of the algorithm to a multi-traffic object.

ACKNOWLEDGEMENTS

This work is within the scope of a co-guardianship between the university of Sousse (Tunisia) and Blaise Pascal University (France). It is sponsored by the Tunisian Ministry of Higher Education & Scientific Research.

REFERENCES

Duan, L., Tsang, I. W., Xu, D., and Maybank, S. J. (2009). Domain transfer svm for video concept detection. In *CVPR*, pages 1375–1381. IEEE.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *IJCV*.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520.

Goodfellow, I. J., Courville, A., and Bengio, Y. (2012). Spike-and-slab sparse coding for unsupervised feature discovery. *arXiv*.

Guyon, I., Dror, G., Lemaire, V., Taylor, G., and Aha, D. W. (2011). Unsupervised and transfer learning challenge. In *IJCNN*, pages 793–800. IEEE.

Htike, K. K. and Hogg, D. C. (2014). Efficient non-iterative domain adaptation of pedestrian detectors to video scenes. In *2014 22nd International Conference on Pattern Recognition (ICPR)*, pages 654–659. IEEE.

Huang, G. B., Lee, H., and Learned-Miller, E. (2012). Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525. IEEE.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACM*, pages 675–678. ACM.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pages 2278–2324.

Li, X., Ye, M., Fu, M., Xu, P., and Li, T. (2015). Domain adaption of vehicle detector based on convolutional neural networks. *International Journal of Control, Automation and Systems*, pages 1020–1031.

Maâmatou, H., Chateau, T., Gazzah, S., Goyat, Y., and Essoukri Ben Amara, N. (2016). Transductive transfer learning to specialize a generic classifier towards a specific scene. In *VISAPP*, pages 411–422.

Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I. J., Lavoie, E., Muller, X., Desjardins, G., Warde-Farley, D., et al. (2012). Unsupervised and transfer learning challenge: a deep learning approach. *ICML Unsupervised and Transfer Learning*, pages 97–110.

Nair, V. and Clark, J. J. (2004). An unsupervised, online learning framework for moving object detection. In *CVPR*, pages II–317. IEEE.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, A., Doucet, A., de Freitas, N., and Gordon, N. (2013). *Sequential Monte Carlo methods in practice*. Springer Science & Business Media.

Taigman, Y., Yang, M., Ranzato, M. A., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708.

Wang, M., Li, W., and Wang, X. (2012). Transferring a generic pedestrian detector towards specific scenes. In *CVPR*, pages 3274–3281. IEEE.

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

Wang, X., Wang, M., and Li, W. (2014). Scene-specific pedestrian detection for static video surveillance. *PAMI*, pages 361–362.

Will Y. Zou, Serena Y. Yeung, A. Y. N. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CSD*, pages 3361–3368.

Yunxiang Mao, Z. Y. (2015). Training a scene-specific pedestrian detector using tracklets. pages 170–176.

Zeng, X., Ouyang, W., Wang, M., and Wang, X. (2014). Deep learning of scene-specific classifier for pedestrian detection. In *ECCV*, pages 472–487. Springer.