MULTI-TARGET DETECTION IN CCTV FOOTAGE FOR TRACKING APPLICATIONS USING DEEP LEARNING TECHNIQUES

A. Dimou[⋆†] *P. Medentzidou*[†] *F. Álvarez García*[⋆] *P. Daras*[†], *Senior Member IEEE*

[⋆] Universidad Politécnica de Madrid (GATV), Spain
[†] Information Technologies Institute, Centre for Research and Technology Hellas, Greece

ABSTRACT

Real-world CCTV footage often poses increased challenges in object tracking due to Pan-Tilt-Zoom operations, low camera quality and diverse working environments. Most relevant challenges are moving background, motion blur and severe scale changes. Convolutional neural networks, which offer state-of-the-art performance in object detection, are increasingly utilized to pursue a more efficient tracking scheme. In this work, the use of heterogeneous training data and data augmentation is explored to improve their detection rate in challenging CCTV scenes. Moreover, it is proposed to use the objects' spatial transformation parameters to automatically model and predict the evolution of intrinsic camera parameters and accordingly tune the detector for better performance. The proposed approaches are tested on publicly available datasets and real-world CCTV videos.

***Index Terms*—** CCTV, motion blur, PTZ, R-CNN, spatial transformer, RNN

1. INTRODUCTION

Object tracking has attracted substantial attention in the research community due to its value in practical applications and especially smart video surveillance solutions. Despite the progress made in recent years, object tracking methods are not robust enough for real-world content from CCTV cameras. In addition to poor and changing illumination, occlusions and cluttered scenes that pose tracking challenges, CCTV footage also suffers from motion blur and large affine transformations due to Pan-Tilt-Zoom (PTZ) operations [1].

Most of the existing detection methods are focused on building a robust object appearance model, working on handcrafted feature representation and classifier construction. However, most of these classifiers are limited by their shallow structures while object appearance variations are complex and time-varying [2]. Recent advances in deep learning have led to a new generation of object detection and localization methodologies that outperform traditional methods. They rely on automatically learning discriminative features via a multi-layer convolutional neural network, thus, alleviating the need for handcrafted features. Each layer is composed of different types of neurons featuring convolutional operations, non-linear filtering and spatial pooling. End-to-end training is used to automatically learn hierarchical and object-specific feature representations.

Object tracking with deep learning techniques, however, has attracted considerably less attention in the past, partly due to the lack of sufficient training data. Li et al. [3] incorporated a convolutional neural network (CNN) to visual tracking with multiple image cues as inputs. In [4] an ensemble of deep networks has been combined with an online boosting method. In [5], a single-target online learning tracker is proposed to alleviate blurring. Another line of research exploits auxiliary data to train offline a deep network, and then transfers knowledge to object tracking. Fan et al. [6] proposed learning a specific feature extractor with CNNs from an offline training set. In [7] a deep learning tracking method is proposed that uses stacked denoising autoencoder to learn the generic features from a large number of auxiliary images. Recently, Wang et al. [8] employed a two-layer CNN to learn hierarchical features from auxiliary data, which models complicated motion transformations and appearance variations. In [9], a deep learning architecture learns the most discriminative features via a CNN exploiting both the ground truth appearance information and the image observations obtained online.

In this work, a multiple-object detection framework for tracking by detection applications that confronts the challenges of real-world CCTV videos is proposed. It is based on a state-of-the-art detection and localization object framework [10] that is trained offline to facilitate a tracking-by-detection paradigm. A number of techniques for the augmentation of the training data are examined to streamline the performance of the detector. Moreover, a methodology to dynamically control the detector configuration using estimations of the intrinsic parameters of the camera is proposed. A Recurrent Neural Network (RNN) is employed to model the spatial transformation [11] of the objects due to the camera perspective and PTZ operations of the camera. The RNN is used to predict the affine transformation of the objects and dynamically modify the parameters of the detector. Experimental validation of the proposed concept is performed on real CCTV videos. The proposed framework is applicable to any type of objects but experiments will focus on pedestrians.

The rest of the paper is organized as follows. An exploration of training data augmentation for object detection is given in Section 2. Section 3 introduces the procedure to model and predict the object transformation for dynamic parametrization of the detector. The experimental results are given in Section 4 and conclusions are drawn in 5.

2. TRAINING DATA AUGMENTATION

CCTV videos often contain severely blurred objects due to low video quality and fast PTZ operations. Especially motion blur is a major challenge for object detection in CCTV content. In a single frame, motion blur is translated to degraded appearance information and reduced ability to accurately localize the position of an object. While de-blurring methodologies show good results [12], they have a high computational cost and they further degrade their appearance. Deep learning systems have been recently shown to achieve impressive performance in benchmark datasets for object detection. However, in challenging CCTV videos their performance deteriorates. In this section, the effect of training data selection in the detector's performance is explored.

Building on prior deep learning work, the object detection and localization framework Faster R-CNN [10] is employed. It combines the localization and detection tasks, while sharing convolutional layers to speedup the process. A ZF network model [13], pre-trained in ImageNet dataset [14], is selected for object detection. The model is fine-tuned to optimize the discriminative power of the features learned and therefore the detection accuracy. The strategy of fine-tuning has been widely used in deep learning greatly improving the performance of a CNN. It has been shown [15] that transfer learning, namely the use of unsupervised pre-training in a generic dataset, has significant value, offering a robust initialization of the network parameters.

Two training data augmentation approaches are examined to improve fine-tuning of the examined system: (a) the enrichment with object instances from heterogeneous sources and (b) the addition of blurred instances of the current object collection. In the former approach, annotated datasets featuring the examined object classes are utilized. An extended training set is created that contains samples from multiple datasets.

Despite the existence of several annotated datasets, their content is produced with quality measures that are superior to the conditions that a normal CCTV system will face. Therefore, the features learned by a deep learning object detector are often plagued by many missing detections, especially in action scenes. Following the latter approach, the training set is augmented with blurred instances to enhance the robustness to motion blur. A set of Gaussian kernels incorporating motion blur [16] has been created (Eq.1).

$$\mathbf{K} = \{k_{\theta,l} | \theta \in \Theta, l \in \mathbf{L}\} \tag{1}$$

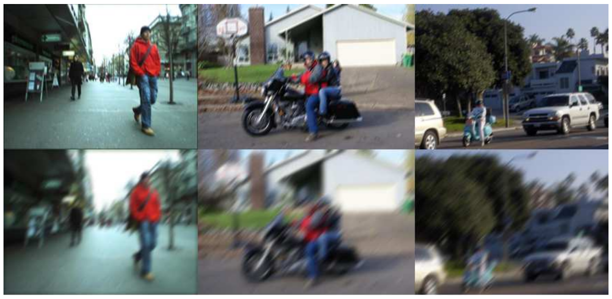

Fig. 1. Examples of motion blur effect on images [17][18].

As the kernel k is symmetric, the motion direction θ is randomly sampled from $\Theta = [0, \pi]$ and the magnitude is selected from $\mathbf{L} = [0, l_{max}]$, where l_{max} is a parameter. The original images I that form the training set are convolved with the motion-blur kernels.

$$\mathbf{I}_{bl} = \mathbf{I} \otimes \mathbf{k} \tag{2}$$

Fig. 1 shows examples of using kernels to generate blurred images with different parameters. The effect of data augmentation approaches is experimentally tested in Section 4.

3. DYNAMIC DETECTOR CONFIGURATION

CCTV cameras often have PTZ capabilities that are used by their operators to track suspicious activities in a scene. These camera operations constitute a serious challenge for object detection and tracking due to the implicit scale assumptions made. Object detection techniques have a predefined range of scales that are supported, to minimize detection errors. In this section it is proposed to dynamically adjust this scale range based on predictions of the tracked objects' size in the next frame.

The first step towards this approach is to have an accurate estimation of the detected object's scale and pose. Recently, a new module was proposed that applies a spatial transformation to a feature map during a single forward pass. The spatial transformer network (SPN) [11] can be used as a new type of layer in a feed-forward convolutional network. It learns an affine transformation of the input and uses bilinear interpolation to produce its output allowing it to zoom, rotate and skew the input. The transformation parameters (Eq.3) can be also exploited as a robust indication of the object's scale and pose.

$$A = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \tag{3}$$

An RNN is subsequently used to model and predict the evolution of the transformation matrices in the next frame. The transformation matrix of the detected objects is input to the recurrent network such that:

$$A_t = SPN(f_{conv}(I)) \tag{4}$$

$$h_t = f^{rnn}_{trans}(A_t, h_{t-1}) \tag{5}$$

where A is the transformation matrix from the current object, SPN is the spatial transformer module and h_{t-1} is the hidden state of the RNN model in the previous step. An affine transformation matrix A_{t+1} is produced at each time-step t from the hidden state of the RNN. The affine transformations predicted are conditioned on the previous transformations through the time dependency of the RNN.

The produced A_{t+1} is utilized to dynamically adjust the optimal scale range of the detector, in this work Faster R-CNN. To achieve that, we modify the scaling parameter s of the Faster R-CNN that controls the scale of the processed image, achieving better scale invariance.

4. EXPERIMENTAL RESULTS

4.1. Datasets

In this section, the experimental setup for the validation of the above concepts is being described. Given that pedestrians is the main object class of interest, the experiments will focus on the pedestrian detection without losing its generality. For this purpose a number of datasets have been selected to feature the experiments. VOC2007 [17] is used as a generic dataset for image classification with 20 annotated classes, including the class $person$. The ETH dataset [18] is also used to extend the fine-tuning dataset. It contains annotated pedestrians on a public road. Finally, a set of videos from the Metropolitan Police of London (MET) from the riots of 2011 have been also used for qualitative validation. Those videos have been offered for research purposes in the framework of the LASIE FP7 project and they are neither annotated nor publicly available.

4.2. Experiments

The first set of experiments refers to the exploration of training data augmentation strategies. The Faster R-CNN object detection framework is fine-tuned with different training sets to test their effectiveness. Training with VOC2007 (∼10000 object instances), labeled as [VOC], is used as a baseline for performance. The training set is infused with sequences from the ETH dataset, namely “Bahnhof” sequence (∼7500 object instances) labeled as [BAH] and “Sunny Day” (∼1900 object instances), labeled as [SUN]. The datasets are divided in training and testing set of equal size. 50% of the training set is used for validation purposes. The evaluation of the trained models is performed on separate testing sets that include VOC testing set and an ensemble of the VOC and ETH testing sets, respectively. The results are reported in Table 1. Experiments show that the detection accuracy seems to benefit from extra training samples, even on the original VOC testing set.

	VOC	VOC+BAH	VOC+SUN
[VOC]	59,44%	57,67%	59,49
[VOC+BAH]	60,66%	62,03%	61,23
[VOC+SUN]	59,79%	57,68%	63,45

Table 1. Average precision of models trained with VOC2007 and with an ensemble of VOC2007 and the ETH dataset on the respective testing sets.

Subsequently, the effect of augmenting the data with motion blur is examined. Training with the VOC2007 dataset is again used as baseline. The dataset is then augmented with blurred instances of the VOC2007 dataset, creating [VOC5] for $l = 5$ pixels and [VOC10] for $l = [5, 10]$ pixels motion blur. The trained models are tested on all testing sets, named $NoBlur$, $Blur5px$ and $Blur10px$ respectively. The results are depicted in Table 2.

	NoBlur	Blur5px	Blur10px
[VOC]	59,44%	31,71%	23,50%
[VOC5]	63,48%	62,63%	61,79%
[VOC10]	63,20%	62,31%	60,33%
[BAH]	70,70%	68,42%	59,08%
[BAH5]	70,66%	70,65%	69,23%
[BAH10]	72,52%	71,75%	71,34%

Table 2. Average precision of detection models fine-tuned with VOC2007 utilizing different magnitude of data augmentation on testing sets with increasing levels of blurring.

It is evident from the results that the performance of the detector is quickly deteriorating when even small amounts of motion blur are introduced. On the other hand, augmenting the training data with blurred examples is making the detector more robust, even on non-blurred data. However, when the dataset is dominated by blurred samples (VOC10) average precision declines slightly. Examples of the detection capabilities of each model on MET videos are depicted in Fig.2.

Another set of experiments is performed to validate the proposed dynamic configuration of the detector. A spatial transformer layer is added in the input of the ZF network and it is applied in the region proposed by [10]. The allowed transformations (Eq.3) are further constrained allowing only cropping, rotation and isotropic scaling to reduce training complexity by varying s, θ in Eq.6

$$A_\theta = \begin{bmatrix} s\cos\theta & s\sin\theta & 0 \\ -s\sin\theta & s\cos\theta & 0 \end{bmatrix} \tag{6}$$

The transformation matrix A_t is provided to an RNN. For the RNN we use the configuration in [19]. The RNN is initially trained with artificial data created to simulate zooming operations. The set includes 100 sequences of bounding box evolution with a length of 100 frames. A linear layer is applied to convert h_t into A_{t+1}.

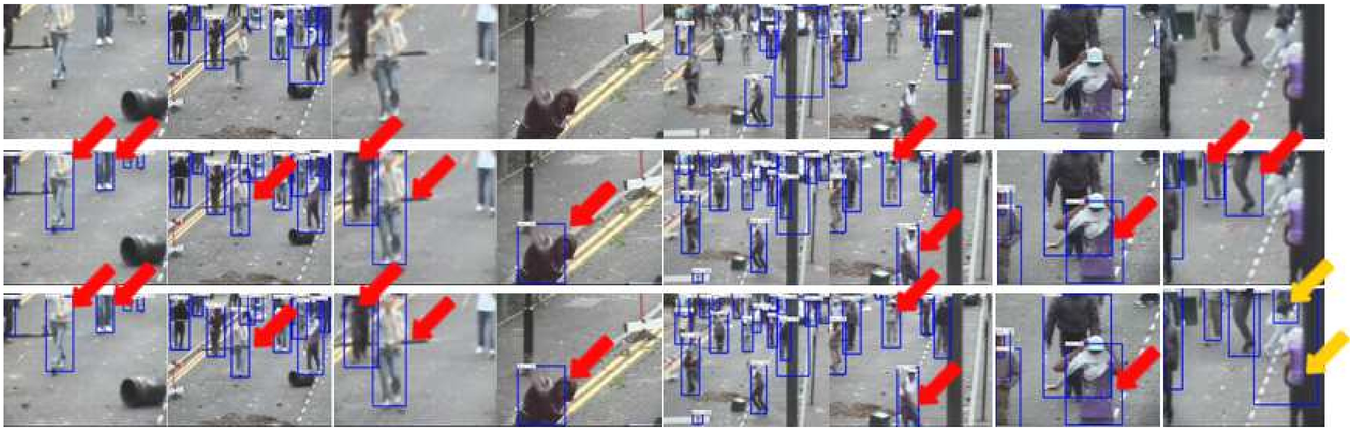

Fig. 2. Example detections on a MET CCTV video trained with ascending levels of blurrness. VOC, VOC5 and VOC10 are depicted in rows 1, 2, 3, respectively. Red arrows depict new detections with VOC5 and yellow new detections with VOC10.

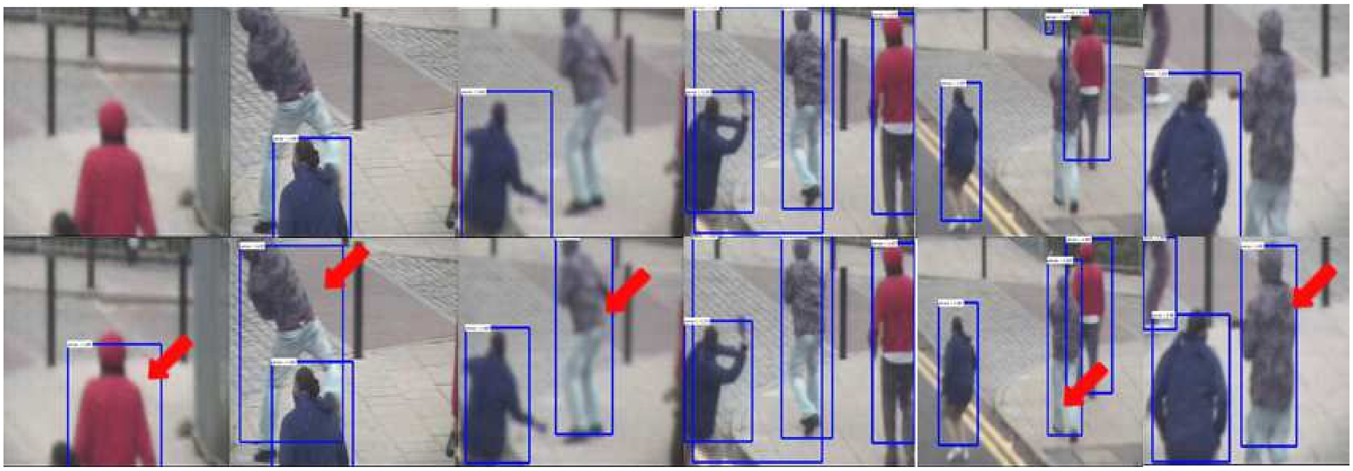

Fig. 3. Example detections on a MET CCTV video. In the first row the default scaling parameter is used ($s = 600$), while in the second the detector uses a dynamically modified scaling parameter. New detections are depicted with red arrows.

The transformation matrix A_{t+1} is used to predict possible severe scale changes of the objects in the next frame. The predicted scale is used to modify the scaling parameter of the detector. Initial experiments of the proposed method have been performed on the MET dataset and are depicted in Fig.3. The results show an improvement of the detection performance in challenging zooming conditions.

5. CONCLUSIONS

In this paper, methodologies to improve the efficiency of deep learning based multi-target detectors in challenging CCTV footage are proposed. The use of heterogeneous data and data augmentation with motion blur is explored for training detectors. Experimental results have shown that the detector benefits from both methodologies. Robust performance is reported in both original and blurred content, as well as challenging action scenes in CCTV videos. Moreover, a novel methodology to dynamically tune the detector parameters during intense PTZ operations is proposed. The spatial transformation of the objects, derived from a spatial transformer network, is used to train an RNN to predict the intrinsic camera properties in the next frame. The predicted parameters are used to tune the detector parameters, leading to more robust results. Initial experiments have shown that dynamic scaling significantly improves the performance of the detector compared to fixed scale operations.

6. ACKNOWLEDGMENTS

This work was supported by the European Commission under contract FP7-607480 LASIE. The authors would like to thank the London Metropolitan Police for providing CCTV footage for research purposes.

7. REFERENCES

[1] Anthony C Davies and Sergio A Velastin, "Progress in computational intelligence to support cctv surveillance systems," *International Journal of Computing*, vol. 4, no. 3, pp. 76–84, 2014.

[2] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah, "Visual tracking: An experimental survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1442–1468, 2014.

[3] Hanxi Li, Yi Li, and Fatih Porikli, *Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, November 1-5, 2014, Revised Selected Papers, Part V*, chapter Robust Online Visual Tracking with a Single Convolutional Neural Network, pp. 194–209, Springer International Publishing, Cham, 2015.

[4] Xiangzeng Zhou, Lei Xie, Peng Zhang, and Yanning Zhang, "An ensemble of deep neural networks for object tracking," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 843–847.

[5] J. Ding, Y. Huang, W. Liu, and K. Huang, "Severely blurred object tracking by learning deep image representations," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[6] Jialue Fan, Wei Xu, Ying Wu, and Yihong Gong, "Human tracking using convolutional neural networks," *Neural Networks, IEEE Transactions on*, vol. 21, no. 10, pp. 1610–1623, 2010.

[7] Naiyan Wang and Dit-Yan Yeung, "Learning a deep compact image representation for visual tracking," in *Advances in Neural Information Processing Systems*, 2013, pp. 809–817.

[8] Li Wang, Ting Liu, Gang Wang, Kap Luk Chan, and Qingxiong Yang, "Video tracking using learned hierarchical features," *Image Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 1424–1435, 2015.

[9] Yan Chen, Xiangnan Yang, Bineng Zhong, Shengnan Pan, Duansheng Chen, and Huizhen Zhang, "Cntracker: Online discriminative object tracking via deep convolutional neural network," *Applied Soft Computing*, vol. 38, pp. 1088–1098, 2016.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," *CoRR*, vol. abs/1506.02025, 2015.

[12] Jian-Feng Cai, Hui Ji, Chaoqiang Liu, and Zuowei Shen, "Framelet-based blind motion deblurring from a single image," *Image Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 562–572, 2012.

[13] Matthew D. Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," *CoRR*, vol. abs/1311.2901, 2013.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[15] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[16] Hailin Jin, P. Favaro, and R. Cipolla, "Visual tracking in the presence of motion blur," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, June 2005, vol. 2, pp. 18–25 vol. 2.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, June 2010.

[18] A. Ess, B. Leibe, K. Schindler, , and L. van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. June 2008, IEEE Press.

[19] Søren Kaae Sønderby, Casper Kaae Sønderby, Lars Maaløe, and Ole Winther, "Recurrent spatial transformer networks," *CoRR*, vol. abs/1509.05329, 2015.