Neurocomputing 398 (2020) 71–82

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Weakly supervised easy-to-hard learning for object detection in image sequences

Hongkai Yu[a], Dazhou Guo[b], Zhipeng Yan[c], Lan Fu[d], Jeff Simmons[e], Craig P. Przybyla[e], Song Wang[d,*]

[a] Department of Electrical Engineering and Computer Science, Cleveland State University, United States
[b] Bethesda Research Lab, PAII Inc., United States
[c] Department of Computer Science, University of California, San Diego, United States
[d] Department of Computer Science & Engineering, University of South Carolina, United States
[e] Materials and Manufacturing Directorate, Air Force Research Laboratory, United States

ARTICLE INFO

Article history:
Received 4 August 2019
Revised 1 January 2020
Accepted 15 February 2020
Available online 24 February 2020

Communicated by Dr. Shen Wei

Keywords:
Weakly supervised
Easy-to-hard learning
Spatio-temporal consistency

ABSTRACT

Object detection is an important research problem in computer vision. Convolutional Neural Networks (CNN) based deep learning models could be used for this problem, but it would require a large number of manual annotated objects for training or fine-tuning. Unfortunately, fine-grained manually annotated objects are not available in many cases. Usually, it is possible to obtain imperfect initialized detections by some weak object detectors using some weak supervisions like the prior knowledge of shape, size or motion. In some real-world applications, objects have little inter-occlusions and split/merge difficulties, so the spatio-temporal consistency in object tracking are well preserved in the image sequences/videos. Starting from the imperfect initialization, this paper proposes a new easy-to-hard learning method to incrementally improve the object detection in image sequences/videos by an unsupervised spatio-temporal analysis which involves more complex examples that are hard for object detection for next-iteration training. The proposed method does not require manual annotations, but uses weak supervisions and spatio-temporal consistency in tracking to simulate the supervisions in the CNN training. Experimental results on three different tasks show significant improvements over the initialized detections by the weak object detectors.

1. Introduction

Object detection is extremely important in computer vision and pattern recognition communities, which has wide applications such as object tracking [1], traffic data collection [2], interested people discovery [3], multi-label classification [4] and so on. Currently, the popular and effective way is to train a Convolutional Neural Networks (CNN) based models like Faster R-CNN [5,6], YOLO [7] to detect the interested objects. However, these CNN based deep learning models require extensive labeled annotations for training, which are not available for many custom data in some real-world applications. For example, to detect a special kind of lesion or tumor from the cluttered backgrounds in medical image sequences, there are no available manual annotations to train or fine-tune the CNN based models. Another example is large-scale fiber detection in material images [8,9], each single image contains approximately 600 objects/fibers. Manually annotating so many fibers for multiple images is tedious and with very high working load.

In many real-world applications, an initialized weak object detector can be easily obtained using some weak supervisions like the prior knowledge of object shape, size or motion. For example, ellipse shape information can be used to detect the fibers in material images [8,9], and the motion information can be used to detect the moving vehicles on traffic images [10]. Meanwhile, in many cases, objects have little inter-occlusions and split/merge difficulties, so multi-target object tracking is reliable and accurate in these cases [11] and the spatio-temporal consistency in object tracking are well preserved. Our objective is to discover a new method to learn a more robust CNN based detection model from the initialized weak object detector with the assistance of spatio-temporal consistency.

This paper proposes a new weakly supervised learning method for object detection in image sequences using an easy-to-hard learning manner. This easy-to-hard learning strategy is inspired

* Corresponding author.
E-mail addresses: hongkaiyu2012@gmail.com (H. Yu), songwang@cec.sc.edu (S. Wang).

https://doi.org/10.1016/j.neucom.2020.02.075
0925-2312/

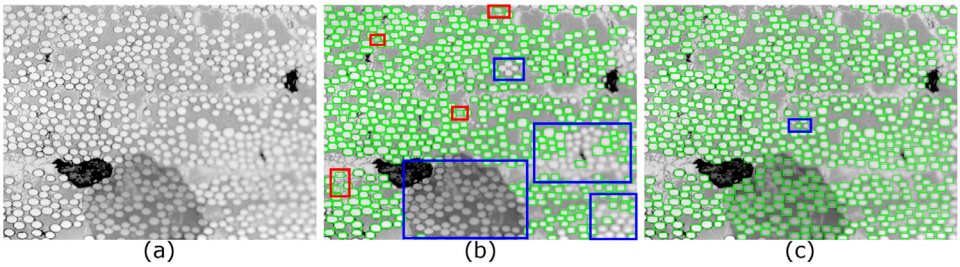


Fig. 1. Comparison of the state-of-the-art method of [9] and the proposed method: (a) one microscopic material image, (b) fiber detections by [9], and (c) fiber detections by the proposed method. Detected fibers are marked by green bounding boxes. Red and blue boxes highlight the false positive and false negative errors respectively. Note that [9] fails where the image quality is poor, i.e., blurred and stained regions. Ideally, the fibers and backgrounds in clear regions are simple examples, while the fibers and backgrounds in blurred and stained regions are complex examples.

by Curriculum Learning [12]. The idea is to firstly train a CNN based detector on simple examples and then incrementally involve more complex examples to train the CNN based detector. In this paper, we define the objects/background regions that are easily distinguished by the current object detector as the simple samples, while the objects/background regions that are hardly distinguished by the current object detector are the complex samples. As we discussed, in many cases, the spatio-temporal consistency in object tracking is very effective, so we develop an unsupervised spatio-temporal analysis to remove false positive detections and make up for false negative detections, which actually refines the training examples that are hard for current object detector. Removing false positive detections means to prune the current examples with poor spatio-temporal consistency and simultaneously add more complex background examples, while making up for false negative detections indicates involving more complex object examples. Alternately running CNN training and the unsupervised spatio-temporal analysis by several iterations is to add more complex examples to the simple examples so that we could train a more robust CNN based object detector. The proposed method is similar to the Baby-Steps Curriculum Learning [13], which believes that simple examples in the training set should not be discarded, instead, the complexity of the training data should be increased in a multiple-iteration training manner. Similarly, the proposed method uses the above mentioned unsupervised spatio-temporal analysis for adding more complex examples in the current iteration for the next-iteration CNN training.

In this paper, we use the task of fiber detection from microscopy image sequences of continuous fiber reinforced composite materials to explain the proposed method. Recent studies [8,9] showed that the 3D fiber structures could be reconstructed by tracking the detected fibers through the 2D image sequence, which is modeled as a tracking-by-detection problem in computer vision. The tracking performance of tracking-by-detection algorithms is largely dependent on the detection accuracy, especially the Recall performance [14]. More accurate object detections could greatly improve tracking-by-detection algorithms. State-of-the-art fiber detection is currently achieved by [9], which detects ellipses by Hough transformation with the prior knowledge of fiber shape, followed by a bounding box fitting. This algorithm performs well in the material image regions whose quality is good, but it performs poorly in the material image regions whose quality is bad. In the present case, the images were degraded in local regions because of contaminants during the data collection, such as the blurred and stained regions in Fig. 1. In the degraded situations, the state-of-the-art fiber detection algorithm [9] fails to accurately detect the fibers, resulting in false positive or false negative errors as shown in Fig. 1. Experimental results on this fiber detection task show that the proposed method using [9] as initialization is able to greatly improve the fiber detection performance. In addition, we also conduct other experiments on vehicle detection and pedestrian detection tasks in image sequences. The initialized vehicle detection results are obtained by background subtraction with the weak supervision of vehicle motion information [10]. Experimental results on the vehicle detection task also show significant improvements from the initialized vehicle detection by the proposed method. We see the similar improvement by the proposed method on the pedestrian detection task. In the tasks of fiber and vehicle detections, object tracking can be well addressed because the objects have little inter-occlusions and split/merge difficulties. For example, the fiber tracking performance MOTA is as high as 99% [8,9] when the image sequence is high-continuity. The satisfactory spatio-temporal consistency in object tracking helps the proposed easy-to-hard learning method to achieve the successful performance boost. In the task of pedestrian detection, there are some inter-occlusions, feature changes and tracking difficulties, so the proposed method has some variances in algorithm convergence but we still see significant improvement by the proposed method.

The remainder of the paper is organized as follows. Section 2 briefly reviews the related work. Section 3 describes the proposed object detection method. Quantitative and qualitative evaluation results and discussions are presented in Section 4, followed by brief conclusions in Section 5.

2. Related work

Object detection: With the prior knowledge of object shape or size, some traditional methods have been proposed to detect interested objects like ellipse detection based fiber detection [8,9,15], HOG feature based human detection [16] and so on. With the prior knowledge of object motion, background subtraction can be used to detect moving vehicles [10]. However, these previous traditional methods using low-level features are not robust in the degraded images or complex situations. Recently, advanced CNN based methods [5–7,17–19] are quite effective for object detection and discovery, while the requirement of a large number of manual annotations for training CNN is expensive.

Learning without manually annotated labels: Machine learning and deep CNN models have shown advanced performances in many computer vision tasks [20–22] if enough training labels are available. In this paper, we focus on a different problem: CNN learning without manually annotated labels, which refers to CNN learning with prior knowledge or constraints but without manual annotations. With several prior knowledge or constraints, the powerful CNN can still be effective, approaching approximately close performance with the CNN trained with full supervision [23–29]. Due to the lack of manual annotations, different video or multiple images based properties are frequently utilized to simulate the human supervisions. Optical flow based motion information is

used to assist CNN for edge detection [23]. Assuming that adjacent video frames contain similar representation, feature learning is performed in unlabeled video data [24]. The fusion of multiple saliency maps is used to simulate human supervision to train CNN without manual annotations to improve unsupervised saliency detection [25]. Using the chronological order of frames as supervision, unsupervised deep representation learning is applied [26]. Given unlabeled videos, unsupervised object discovery is used to train a CNN for detecting objects in single images [27]. Assuming tracked patches have similar visual representation in deep feature space, unsupervised learning of visual representations is accomplished [28]. By emphasizing the inherent correlation among video frames, [29] proposed a co-attention siamese network for unsupervised video object segmentation. In the area of object detection, weak supervision was sometimes provided to avoid manually labeling the object locations, e.g., only using the image-level labels [30,31] to discover the object locations where the Multiple Instance Learning (MIL) based methods are effective.

Spatio-temporal consistency: Many tracking-by-detection methods have been proposed to track single or multiple detected objects [1,32]. In most cases, after obtaining the detected candidate objects in each image, an association or graph algorithm can be used to link the objects for tracking [33]. In [8,11], it shows that the tracking-by-detection task is not challenging if objects have little inter-occlusions and split/merge situations in high-continuity image sequences/videos. In other words, spatio-temporal consistency is well preserved for objects in these cases. Spatio-temporal consistency, as an important property in video processing, has many vision applications such as video object proposals [34], object instance search in videos [35], human segmentation [36], video object segmentation [37,38], video saliency [39] etc. As described in [40], tracking and detection can be jointly carried out in a supervised CNN based framework. In [41], object detection in videos can be improved by a proposed high quality object linking method based on the spatio-temporal consistency. In this paper, we use spatio-temporal consistency in object tracking to simulate the human-like supervisions in CNN training.

Easy-to-hard learning: In this paper, easy-to-hard learning is similar to the idea of Curriculum Learning. Bengio *et al* proposed Curriculum Learning [12] to show that machine learning model can be better learned in a meaningful order: from simple examples to complex examples. Baby-Steps Curriculum Learning [13] directly sorts the difficulties of the training examples and incrementally add more complex examples for training in next iteration. Curriculum Learning strategy is able to reach a global or reasonable local minima during optimizing the loss function in training machine learning models, so it is widely used in machine learning. [42] introduces a weakly supervised CNN learning method for semantic segmentation from simple images to complex images. [43] proposes a weakly supervised multiple instance learning method for within-image co-saliency detection from simple images to complex images.

Inspired by these researches, in order to train a robust CNN based object detector without manual annotations, we combine weak object detectors and spatio-temporal consistency in tracking to simulate the human-like supervision in an easy-to-hard learning way.

3. Proposed method

As mentioned above, we introduce the proposed method using the task of fiber detection in the material images that has wide applications in material science. The input is an material image sequence without any manual annotations. With some weakly supervised methods [9,44] using shape or size prior, the initial pseudo ground truth of fiber detections could be obtained. The powerful CNN based object detector is used as the base detector in our framework. Because there are little inter-fiber occlusions and split/merge difficulties in this fiber detection task, the great spatio-temporal consistency in fiber tracking is used to simulate the supervision to correct and refine the pseudo ground truth (reduce false positive and false negative detections). Refined pseudo ground truth with more complex examples would train a better CNN based object detector in next iteration, and the improved object detector would generate more accurate fiber detections so as to boost fiber tracking, while better fiber tracking would further correct and refine the pseudo ground truth. We expect that the CNN based object detector and tracking-by-detection algorithm could help each other. For a robust solution, the processes of CNN training/testing and fiber tracking are performed alternately in several iterations. The diagram of the framework of the proposed method is illustrated in Fig. 2. For the CNN based object detector, we use Faster R-CNN [6] (Region-based Convolutional Neural Networks) due to its outstanding detection performance. For the fiber tracking, we use Kalman filter based fiber tracking algorithm [9] because of its satisfactory performance in fiber tracking in high-continuity image sequences. We will introduce the details of each part in the following.

3.1. Initialization

To start the proposed method, we need initial pseudo ground truth of fiber detections, which can be accomplished by some weakly supervised methods. We provide two ways to create the initial pseudo ground truth in this paper.

The first one uses the EdgeBox [44] algorithm to detect contour-representative object proposals, including both fiber and non-fiber regions. To reduce the detection errors, we use size prior to delete many false positives. We first compute the mean size/area a of many detected fibers (bounding boxes) by [9] on one sample image of the input image sequence, and then prune the object proposals whose sizes are out of the range $[0.2a, 2a]$ in the input image sequence, followed by a Non-Maximum Suppression (NMS). The NMS threshold is set to 0.1 due to the highly overlapped proposal regions by EdgeBox. The second one is the state-of-the-art fiber detection algorithm [9], shortly written as EMMPMH in later statement. EMMPMH first applied the EMMPM segmentation algorithm [45], which is a Markov Random Fields based unsupervised algorithm for image segmentation, to segment the material images into fiber and non-fiber regions and then utilized a Hough transform based ellipse fitting algorithm [46] to detect the fibers. The minimum bounding boxes of each detected ellipse were taken as the fiber detection result. Both EdgeBox and EMMPMH algorithms are weakly supervised image processing methods using either size or shape prior to detect fibers. Due to the lack of more supervisions, EdgeBox and EMMPMH algorithms have some false positives and many false negatives, as shown in Fig. 3. Because the EMMPMH algorithm utilizes more meaningful shape prior, so it generates better initialization results than the EdgeBox algorithm. These two methods are displayed as initializations for the pseudo ground truth of fiber detections, where the proposed method is able to achieve better fiber detection and tracking starting from either one. Ideally, the fibers and backgrounds in clear regions are simple examples, while the fibers and backgrounds in blurred and stained regions are complex examples. Most of simple examples can be distinguished by the initialized weak fiber detectors, but most of complex examples are hard for them to discover.

3.2. Faster R-CNN for fiber detection

Recently, Faster R-CNN advanced many object detection related tasks [47,48] with stable performance and efficient computation,

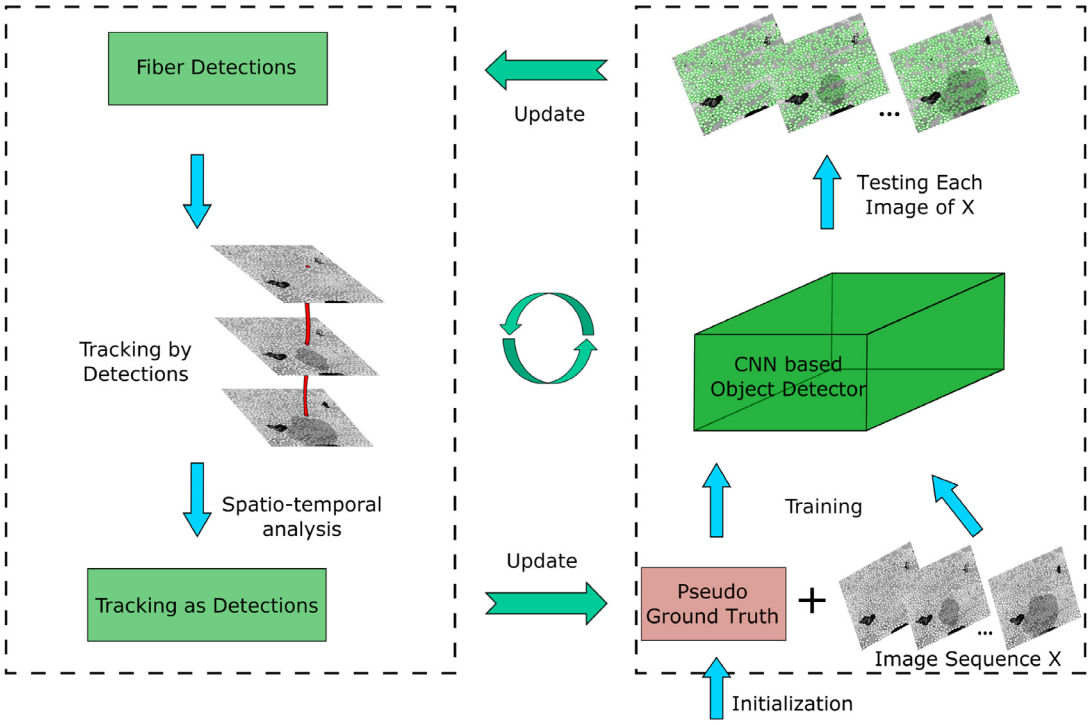


Fig. 2. The framework of the proposed easy-to-hard learning method for fiber detection. Left part: fiber tracking, Right part: CNN training/testing. The input is an image sequence without any manual annotations. More complex examples will be involved for next-iteration CNN training by a specially designed spatio-temporal analysis.

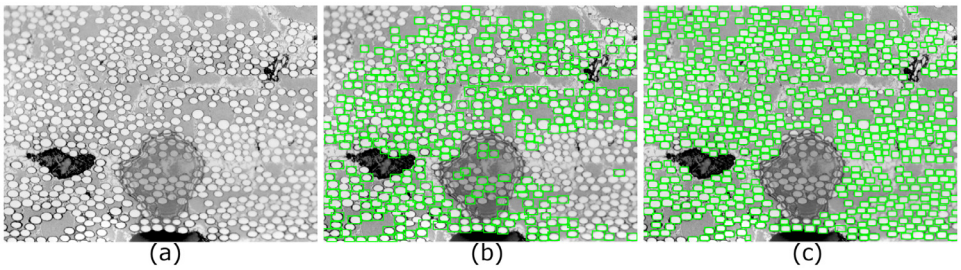


Fig. 3. Initialization for the pseudo ground truth of fiber detections shown as green bounding boxes. (a) original image, (b) initialization by EdgeBox [44] with many missed detections, and (c) initialization by EMMPMH [9] with some missed detections.

so it is utilized as the CNN based object detector in the proposed method. Given the pseudo ground truth, Faster R-CNN could be trained for fiber detection. Faster R-CNN is an end-to-end detection network towards real-time object detection. In particular, it is composed of two modules. The first module is a deep convolutional network that proposes regions, named as a Region Proposal Network (RPN). The second module is the Fast R-CNN detector [5] that uses the proposed regions for object detection and classification. Since the RPN shares full-image convolutional features with the detection network, the computation cost of region proposals is low. RPN serves as the 'attention' mechanism which tells the network where to look.

To handle different scale and aspect ratios of objects, Faster R-CNN introduces anchors of different scales and aspect ratios in a sliding window manner. In the proposed method, we use 5 scales (32^2, 64^2, 128^2, 256^2, 512^2 pixels) and 3 aspect ratios (1: 1, 1: 2, and 2: 1). Following [6], the anchors whose Intersection-over-Union (IoU) overlap with a pseudo ground-truth box is higher than 0.7 or smaller than 0.3 are set as positive and negative samples respectively during training RPN. The loss function L of Faster R-CNN contains two components:

$$L = L_{cls} + \lambda L_{reg}, \tag{1}$$

where L_{cls} is the normalized *classification* loss and L_{reg} is the normalized *regression* loss with a balance weight λ, setting $\lambda = 1$ as in [5]. L_{cls} is a log loss over two classes (fiber v.s. non-fiber) and L_{reg} is the smooth L1 loss over bounding box locations [5]. Same as [6], we sample 256 anchors (positive and negative) for one image during training RPN (first module). For training Fast R-CNN (second module), we fix the IoU threshold for NMS as 0.7 so as to generate about 2000 proposal regions per image. The VGG network [49] is used as the base convolutional layers to extract deep features. The whole Faster R-CNN is a unified network that can be trained end-to-end by back propagation and stochastic gradient descent.

3.3. Tracking by detections

Given the large-scale detected fibers by Faster R-CNN, we model this problem as a tracking-by-detection problem in the image sequence. Since Kalman filter has been proven as a reliable model for large-scale fiber tracking in high-continuity image sequence [8,9], we apply Kalman filter to track each fiber by recursively performing prediction, association and correction along the image sequence.

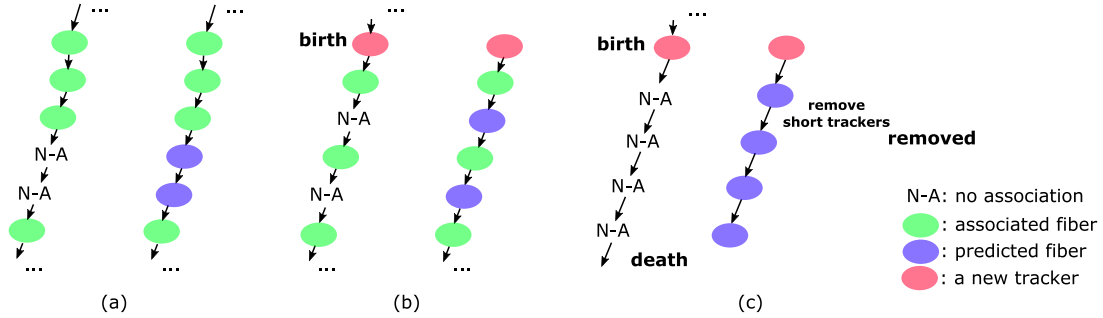


Fig. 4. The unsupervised strategy of spatio-temporal analysis in fiber tracking. (a) making up for false-negative detections by predictions (added), (b) tracking birth by a true positive detection (saved), and (c) tracking birth and death by a false-positive detection (removed).

For fiber detection, we define tracking state $\mathbf{s} = (x_1, y_1, v_{x_1}, v_{y_1}, x_2, y_2, v_{x_2}, v_{y_2})^T$ to denote the tracked fiber in the 2D images, where the first half is for the top-left point of the fiber's bounding box and the last half is for the bottom-right point of the fiber's bounding box. $(x, y)^T$ is the location and $(v_x, v_y)^T$ is the velocity in horizontal and vertical directions. Other reasonable definitions to represent the bounding-box-level object might be also acceptable, e.g., using location and velocity of object center, and the height and width of object. We set up a Kalman filter to track each fiber and assume that each fiber is highly smooth in 3D space with a constant velocity, which means the tracking state is evolved linearly from image to image. The prediction and correction steps are the same as those in the traditional Kalman filter. During the association step, we use the Hungarian algorithm [50] for a minimum-total-distance bipartite matching between the centers of predicted fibers (bounding boxes) and detected fibers (bounding boxes). Following [9], the number of predictions and detections are usually different, so dummy nodes are introduced into the Hungarian algorithm and the distance to a dummy node is set to 100 pixels in our experiments.

3.4. Tracking as detections

After fiber tracking, if we ignore the tracking identities and just take tracked bounding boxes as detected fibers, updated fiber detection could be obtained on each image of the image sequence. This procedure is all right except the tracking errors. Since fiber tracking is not perfect, it is possible that tracking errors might introduce some new detection errors, in terms of false positives and false negatives. In this section, we will introduce an unsupervised strategy of spatio-temporal analysis to reduce some tracking errors during fiber tracking.

Spatio-temporal analysis: Two issues are considered in fiber tracking. 1). To include new coming-in fibers and avoid tracking drifts, tracking birth and death are performed on each image. We start a new Kalman filter to track a detected fiber that is not associated with any predictions of current Kalman filters. We stop a Kalman filter that is tracking a fiber if it has moved out of image boundary or its prediction has not been associated for any detections for continuous α images. Note that the un-associated fibers use their predictions as their tracked locations for continuous α images, which makes up for some false negative detections. 2). Several false-positive detections are tracked for α images due to the above tracking birth and death, so after tracking, we prune the tracked fibers whose trajectories are shorter or equal than β, followed by a NMS on each image. We set $\alpha = 5$ and $\beta = 5$ in our experiments. NMS threshold for EMMPMH initialized tracking is set to 0.7 and for EdgeBox initialized tracking is set to 0.1.

The unsupervised strategy of spatio-temporal analysis is shown in Fig. 4. For the unsupervised strategy of spatio-temporal analysis, we assume that 1) some of missed detections could be added back by tracking predictions, 2) true positive detections are highly possible to be tracked for more than β images, 3) some of false positive detections might be not associated for continuous α images. This assumption is reasonable when the spatio-temporal consistency is good in the image sequences.

Easy-to-hard learning: We intend to learn a robust object detector in several iterations starting from simple examples and then adding more complex examples for CNN training. The current-iteration detections have false positive and false negative errors, which are considered as the complex examples in current iteration. In the proposed spatio-temporal analysis, removing false positive errors means to prune the current examples with poor spatio-temporal consistency and simultaneously add more complex background examples, while making up for false negative errors indicates adding more complex object examples. This easy-to-hard learning process is similar as Curriculum Learning [12] and Baby-Steps Curriculum Learning [13]. The detailed steps of the proposed algorithm are summarized in Algorithm 1. In our experiment, we see that some missed complex fibers in blurred and stained regions can be added by tracking predictions and some false positives (actually backgrounds) can be removed by the proposed spatio-temporal analysis over iterations. With more complex examples in blurred and stained regions involved in next-iteration CNN training, the proposed method always converged in 3 to 4 iterations in our experiment. Besides the easy-to-hard learning, three strategies contribute to the convergence: 1) We normally choose the initialization method with good Precision to start the proposed method (e.g., $\geq$ 92% for fiber detection, $>$ 99%

Algorithm 1 Weakly Supervised Easy-to-hard Learning for Object Detection.

Input: a sequence of images without any manual annotations, denoted as **X**.
Output: a well trained Faster R-CNN model.

1: Initialize the pseudo ground truth $\mathbf{G_p}$ of object detections using a weak object detector on each image of **X**.
2: **repeat**
3: Train a Faster R-CNN from scratch using $\mathbf{G_p}$ if previous Faster R-CNN model is not available. Otherwise, fine-tune the previous Faster R-CNN model using $\mathbf{G_p}$.
4: Apply the trained Faster R-CNN on each image of **X** to detect objects as **D** and save it.
5: Track detected objects **D** on **X** as inSection 3.3 and save it.
6: Take the tracked objects as detections with the spatio-temporal analysis as in Section 3.4.
7: Update the pseudo ground truth $\mathbf{G_p}$ which contains more complex examples.
8: Save the current Faster R-CNN model.
9: **until** convergence or maximum iterations reached

for vehicle detection and $>$ 76% for pedestrian detection), which means that the pseudo ground truth does not have many false positives; 2) Not all the pseudo ground truths are used for training Faster R-CNN, instead we sample 256 anchors (128 positives and 128 negatives) on each image during training Faster R-CNN same as the default setting in [6]; 3) Only the detected objects with high confidence by Faster R-CNN (e.g., $>$ 0.8 for fiber detection) is saved, so low-confidence pseudo ground truths are removed before the following tracking. Based on the three strategies, many not-so-sure pseudo ground truths are not used during Faster R-CNN training. Our easy-to-hard learning method makes sure that more complex samples are added in the next-iteration training, leading to improved detection. After convergence, a well trained Faster R-CNN model is built and it can be applied to directly detect objects on a new image without performing object tracking.

4. Experiments

In the experiment, we firstly apply the proposed method to detect and track large-scale fibers from S200, an amorphous SiNC matrix reinforced by continuous Nicalon fibers, using a recently publicized fiber dataset [9]. The microscopic images were collected by the RoboMet.3D automated serial sectioning instrument [51]. It takes about 15 minutes to grind for one slice. Given a material sample of S200, RoboMet.3D cross-sections the sample by mechanical polishing with dense inter-slice distance 1 μm, and each slice was then imaged with an optical microscope. We choose three datasets from the public fiber dataset [9], denoted as 'Set1', 'Set2' and 'Set3', to evaluate the proposed method. Set1 is a 90-slice image sequence and 40% of images contain certain-level degraded situations such as shadow and blur, as shown in Fig. 1(a). Set2 is 50-slice image sequence and 30% of images have certain-level degraded situations. Set3 is a set of 99 disordered single images and 15% of images have certain-level degraded situations. The size of each image is 1292 $\times$ 968 and each image contains about 600 fibers. On the collected Set1, Set2 and Set3, we manually annotate the bounding boxes of fibers on each image for detection evaluation only.

The proposed method described in Algorithm 1 does not need any manual annotations. We run Algorithm 1 (with tracking) on the image sequence Set1 without manual annotations, and obtain a well trained Faster R-CNN model as $\mathbf{M}^{\mathbf{1}}_{\mathbf{EMMPMH}}$ using EMMPMH [9] initialization and another well trained Faster R-CNN model as $\mathbf{M}^{\mathbf{1}}_{\mathbf{EdgeBox}}$ using EdgeBox [44] initialization. Running Algorithm 1 (with tracking) on Set2 without manual annotations, we could obtain a well trained Faster R-CNN model as $\mathbf{M}^{\mathbf{2}}_{\mathbf{EMMPMH}}$ using EMMPMH initialization and another well trained Faster R-CNN model as $\mathbf{M}^{\mathbf{2}}_{\mathbf{EdgeBox}}$ using EdgeBox initialization. The well trained Faster R-CNN models on one dataset are then respectively applied to detect the large-scale fibers on each single image on another two datasets without tracking.

In our experiment, the maximum iteration in Algorithm 1 is set to 4. Within each iteration, we train Faster R-CNN for 10 epochs. The learning rate is 0.001 and the batch size is 2 images during training. For Algorithm 1, we try two kinds of initializations for pseudo ground truth: EMMPMH and EdgeBox. We denote the proposed Algorithm 1 using the initialization EMMPMH as 'Proposed-EMMPMH' and the initialization EdgeBox as 'Proposed-EdgeBox'. After obtaining the well trained Faster R-CNN model **M**, we denote directly applying the well trained model on single images (without tracking) to detect large-scale fibers as 'Proposed-**M**'. We use MXNet to implement the code of Faster R-CNN framework. With a GeForce GTX 1080Ti GPU and a 12-core CPU, it takes about half an hour to run one iteration (Faster R-CNN training plus large-scale fiber tracking) in Algorithm 1 with Set1 (a 90-slice image sequence) as input and only takes about 0.2 seconds to detect the large-scale fibers on one material image when testing the trained Faster R-CNN model.

Five metrics are used to evaluate the fiber detection performance on Set1, Set2 and Set3: Precision, Recall, F-measure, Number of False Positives per image (N_{fp} per image), and Number of False Negatives per image (N_{fn} per image). For all the methods, we use a uniform threshold of 0.5 for the IoU between the predicted bounding box and ground truth. Because each image contains large-scale fibers (about 600), percentage results might be not representative enough to display errors. Therefore, we also show N_{fp} per image and N_{fn} per image to illustrate the detection errors. Higher (Precision, Recall and F-measure) and lower (N_{fp} and N_{fn} per image) indicate the better detection performance. In our experiment, an ellipse detection algorithm ELSD [15] is used as another comparison method for fiber detection, together with the above mentioned EMMPMH and EdgeBox methods. All these three comparison methods are weakly supervised and do not need manual annotations. For ellipse detections by EMMPMH and ELSD, we take the minimum bounding boxes of each detected ellipse as their outputs.

4.1. Experimental results on fiber detection

After the convergence of running Algorithm 1 on Set1 and Set2 respectively (with tracking), we evaluate the large-scale fiber detection performance on the Set1 and Set2. Meantime, a well trained Faster R-CNN model **M** is obtained after the convergence of running Algorithm 1.

Using the well trained Faster R-CNN model **M**, we directly apply it to detect the large-scale fibers on another two datasets (without tracking). The performance on Set1, Set2 and Set3 is summarized in Table 1. We can see that Proposed method using EMMPMH as initialization achieves the best performance in most cases with high Precision, Recall and F-measure and low N_{fp} per image and N_{fn} per image. The Proposed method using EdgeBox as initialization achieves second best performance and comparable or better performance than the state-of-the-art algorithm EMMPMH. Even without using any manual annotations, the proposed method could achieve nearly 99% F-measure for large-scale fiber detection on three datasets, which fully demonstrates the accuracy and effectiveness of the proposed method. The fiber detection example is shown in Fig. 5.

On Set1, we also run the proposed method using the pseudo ground truth by combining the results of EdgeBox and EMMPMH. Since the results of these two methods have many overlapped detections, we applied NMS on the combined result with an IoU threshold 0.1 to delete the overlaps. The combined initialization gets 91.7% F-measure, where the proposed method initialized by the combined detection finally achieves 96.8% F-measure, compared to 96.0% F-measure using EdgeBox as initialization and 98.6% F-measure using EMMPMH as initialization.

4.2. Improvement and convergence for fiber detection

In this section, we will show the improvement from initialization and the algorithm convergence. We treat the initialization result as baseline denoted as iteration 0, and then show the fiber-detection performance change from iteration 0 to iteration 4 in Fig. 6. We can see that the proposed method significantly improve the fiber-detection performance by both the initialization EMMPMH and EdgeBox. From iteration 0 to 4 using either initialization, the proposed method could incrementally boost the Precision, Recall and F-measure, and simultaneously reduce the N_{fp} per image and N_{fn} per image. We can see that the proposed method converges in 3 to 4 iterations.

In addition, Fig. 7 shows an illustration of the changes of the pseudo ground truth $\mathbf{G_p}$ by Tracking and the corresponding

Table 1
Large-scale fiber detection performance. Set1 and Set2 are two image sequences and Set3 is a set of disordered single images. Note that each image contains about 600 fibers and all the methods in this table do not use manual annotations. Taking EdgeBox as an example, Proposed-EdgeBox means running the proposed method using EdgeBox initialization on the corresponding image sequence, and Proposed-$\mathbf{M}^1_{\text{EdgeBox}}$ means directly testing the model trained on Set1 by the proposed method using EdgeBox initialization, and Proposed-$\mathbf{M}^2_{\text{EdgeBox}}$ means directly testing the model trained on Set2 by the proposed method using EdgeBox initialization. We use the same way to define the methods that use the EMMPMH initialization.

Set1	with tracking?	Precision	Recall	F-measure	N_{fp} per image	N_{fn} per image
EdgeBox [44]	no	93.0%	54.3%	68.6%	26.8	303.3
ELSD [15]	no	93.4%	92.5%	93.0%	43.1	49.3
EMMPMH [9]	no	96.9%	91.7%	94.2%	19.3	54.7
Proposed-$\mathbf{M}^2_{\text{EdgeBox}}$	no	99.0%	97.3%	98.2%	6.1	17.5
Proposed-$\mathbf{M}^2_{\text{EMMPMH}}$	no	**99.3%**	96.1%	97.7%	**3.9**	28.3
Proposed-EdgeBox	yes	99.0%	93.2%	96.0%	6.0	45.1
Proposed-EMMPMH	yes	99.1%	**98.1%**	**98.6%**	5.2	**12.2**
Set2	with tracking?	Precision	Recall	F-measure	N_{fp} per image	N_{fn} per image
EdgeBox [44]	no	92.0%	64.8%	76.0%	32.2	203.6
ELSD [15]	no	91.0%	90.6%	90.8%	51.3	53.8
EMMPMH [9]	no	97.7%	95.1%	96.4%	12.5	28.2
Proposed-$\mathbf{M}^1_{\text{EdgeBox}}$	no	98.8%	93.8%	96.2%	6.4	35.4
Proposed-$\mathbf{M}^1_{\text{EMMPMH}}$	no	99.4%	**98.5%**	**98.9%**	3.1	**8.6**
Proposed-EdgeBox	yes	99.3%	98.4%	**98.9%**	3.9	**8.6**
Proposed-EMMPMH	yes	**99.5%**	98.1%	98.8%	**2.4**	10.6
Set3	with tracking?	Precision	Recall	F-measure	N_{fp} per image	N_{fn} per image
EdgeBox [44]	no	93.4%	56.0%	70.1%	24.2	270.6
ELSD [15]	no	93.6%	91.8%	92.7%	38.4	50.5
EMMPMH [9]	no	97.2%	97.7%	97.4%	17.2	14.1
Proposed-$\mathbf{M}^1_{\text{EdgeBox}}$	no	99.1%	93.5%	96.2%	4.7	39.9
Proposed-$\mathbf{M}^1_{\text{EMMPMH}}$	no	99.3%	98.5%	98.9%	3.8	9.0
Proposed-$\mathbf{M}^2_{\text{EdgeBox}}$	no	99.0%	**99.0%**	**99.0%**	5.8	**6.0**
Proposed-$\mathbf{M}^2_{\text{EMMPMH}}$	no	**99.5%**	98.5%	**99.0%**	**2.5**	8.7

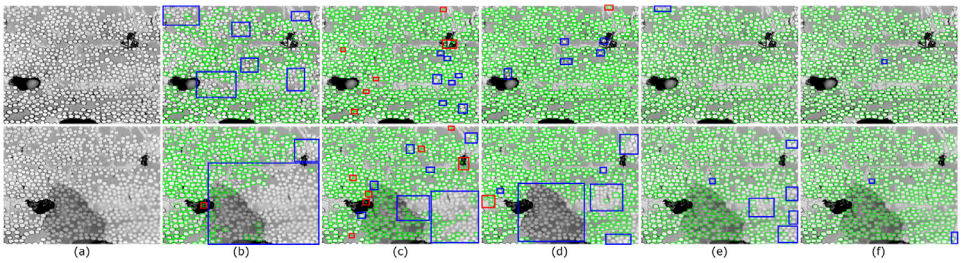


Fig. 5. Illustration of large-scale fiber detection. (a) two sample images (clear and degraded), and detections by (b) EdgeBox, (c) ELSD, (d) EMMPMH, (e) Proposed-$\mathbf{M}^1_{\text{EdgeBox}}$ and (f) Proposed-$\mathbf{M}^1_{\text{EMMPMH}}$. Fibers are detected as green bounding boxes. Red and blue boxes highlight the false positive and false negative errors respectively.

detection $\mathbf{D}$ by Faster R-CNN in each iteration of the proposed method on one example image of Set1. We can clearly see that more complex samples in stained and blurred regions have been added in $\mathbf{G_p}$ in later iterations, which indicates the improved quality of $\mathbf{G_p}$, leading to improved detection $\mathbf{D}$.

4.3. Experimental results on vehicle detection

The proposed method can be used for other object detection task in videos, such as vehicle detection from an aerial-view video. Different vehicles have more variances in feature space (color, texture and structure) than the fibers, so this task is more difficult than fiber detection. We collect a training set (a video of 1000 traffic images) from aerial view, and also a testing set (100 disordered single traffic images) from aerial view in the same location but different time. The size of each image is 1280 × 720 and each image contains about 20 to 50 vehicles. The traffic scene is focused on a two-way urban freeway as shown in Fig. 8. Like the fiber detection task in our training, we do not use any manual annotations, and the manual annotations are only used for performance evaluation. We keep the same parameter setting as that in fiber detection.

Because of the motion information of vehicles, a weak initialized vehicle detector can be easily obtained by background subtraction. Following [10], we compute the background image by averaging the three channels of the 1000 images of training set independently. The moving vehicles can be detected by subtracting each image with the computed background image, thresholding with an adaptive OTSU [52] threshold, removing small connected components and extracting the remaining connected components. Background subtraction based weak vehicle detector provides a good initialization but not perfect so that some vehicles might be not successfully detected, such as the vehicles with similar color and intensity with the road and the vehicles partially occluded by the trees or traffic signs. These vehicles hard for background subtraction based detection are considered as the complex examples for this task. More details for this task are shown in Fig. 8.

In this experiment, we use the background subtraction based weak vehicle detector to initialize the proposed method and train a robust vehicle detector on the training set without manual annotations. Then, we apply the trained robust vehicle detector to detect the vehicles on both the training set (a video of 1000 images) and the testing set (100 disordered single images). We report the evaluation results of each iteration of the proposed method in Table 2. We can clearly see the performance improvement by the proposed method starting with the baseline (initialized) weak vehicle detector. It is worth mentioning that the trained CNN model for vehicle detection on the video can be used to detect

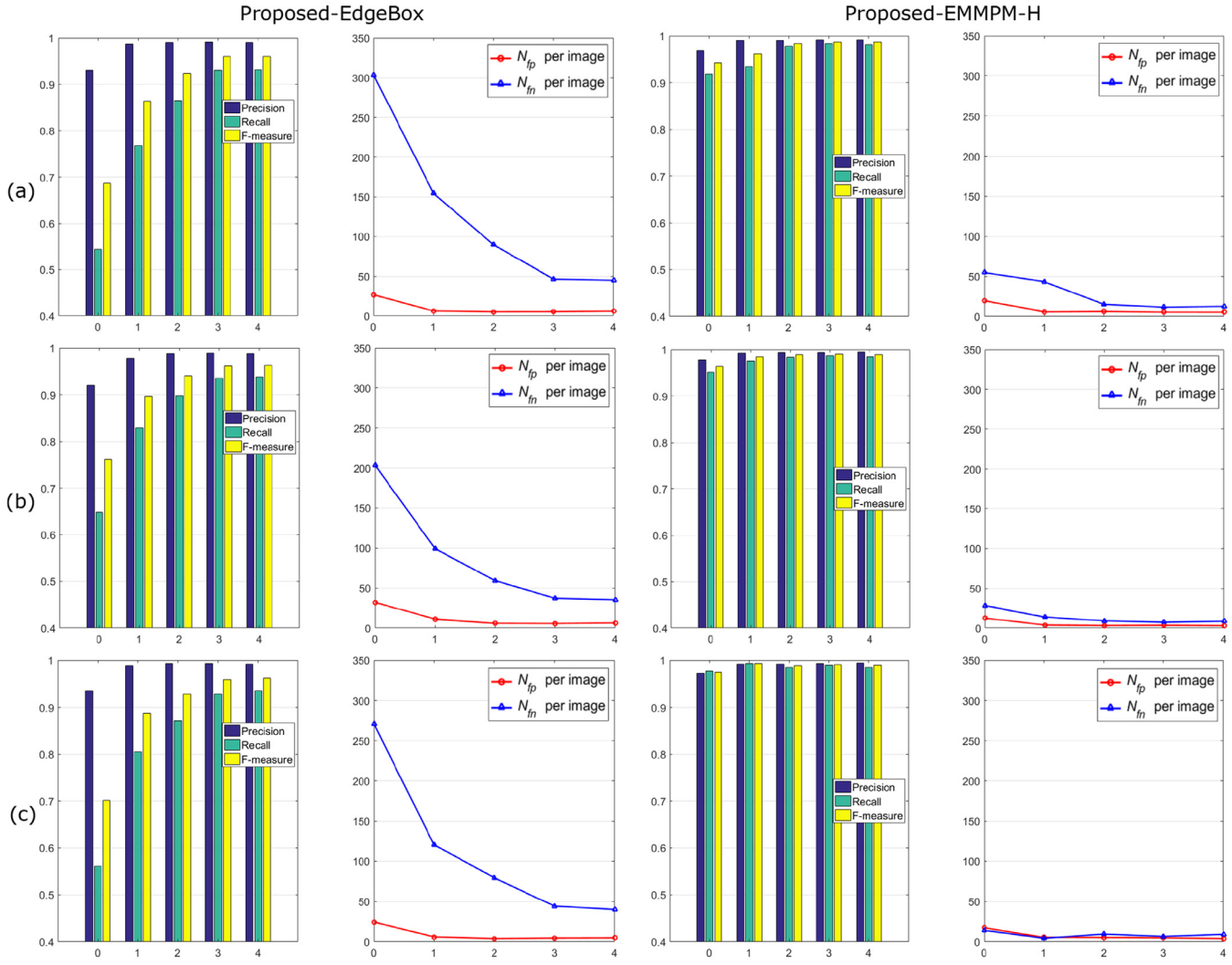
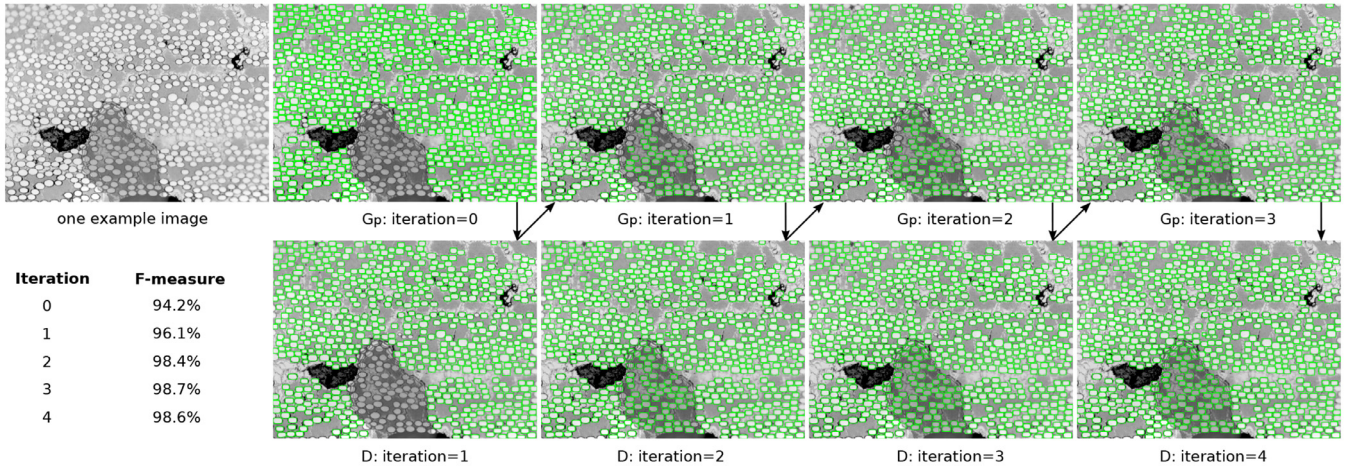


Fig. 6. Illustration of the fiber-detection performance change of the proposed method in (a) Set1, (b) Set2, and (c) Set3 from iteration 0 to iteration 4. Left two columns display the proposed method using EdgeBox as initialization and right two columns show the proposed method using EMMPMH as initialization.



Iteration	F-measure
0	94.2%
1	96.1%
2	98.4%
3	98.7%
4	98.6%

Fig. 7. Illustration of the changes of the pseudo ground truth $\mathbf{G_p}$ by Tracking and the corresponding detection **D** by Faster R-CNN in each iteration of the proposed method. One example image of Set1 is used in this illustration. The detection performance change on all the images of Set1 in each iteration of the proposed method is also displayed in the bottom left. Iteration = 0 indicates the initialization by EMMPMH.

Fig. 8. Illustration of vehicle detection of the proposed method: (a) computed background image from the training set (a video), and vehicle detection by baseline/background subtraction (left) and the proposed method (right) on (b) one sample image of the training set, (c) another sample image of the training set and (d) one sample image of the testing set. Vehicles are detected as green bounding boxes. Red and blue boxes highlight the false positive and false negative errors respectively. Ideally, the vehicles with similar color and intensity with the road and the vehicles partially occluded by the trees or traffic signs are considered as complex examples.

Table 2
Performance of the proposed method in each iteration for vehicle detection on training set (a video of 1000 images) and testing set (100 disordered single images). 'Baseline' is the result of initialized detection by background subtraction, which can be considered as 'iteration = 0'.

Training Set	with tracking?	Precision	Recall	F-measure	N_{fp} per image	N_{fn} per image
Baseline	no	99.2%	86.7%	92.6%	0.2	4.3
iteration = 1	no	**99.8%**	90.6%	94.9%	**0.1**	3.1
iteration = 2	yes	99.4%	92.5%	95.8%	0.2	2.4
iteration = 3	yes	99.1%	93.1%	96.0%	0.3	2.2
iteration = 4	yes	98.3%	**96.0%**	**97.1%**	0.5	**1.3**
Baseline	no	99.3%	87.6%	93.1%	0.2	3.9
iteration = 1	no	**99.7%**	92.5%	96.0%	**0.1**	2.4
iteration = 2	no	99.1%	93.3%	96.1%	0.3	2.1
iteration = 3	no	99.3%	93.6%	96.3%	0.2	2.0
iteration = 4	no	98.8%	**95.5%**	**97.1%**	0.4	**1.4**

Table 3
Performance of the proposed method in each iteration for pedestrian detection on the public PETS09-S2L1 dataset [53] (a video of 795 images). 'Baseline' is the result of initialized detection by a pre-trained human detector [54], which can be considered as 'iteration = 0'.

PETS09-S2L1	with tracking?	Precision	Recall	F-measure	N_{fp} per image	N_{fn} per image
Baseline [54]	no	76.1%	**91.3%**	83.1%	1.6	**0.5**
iteration = 1	no	83.2%	90.5%	86.7%	1.0	**0.5**
iteration = 2	yes	86.9%	90.7%	**88.8%**	0.7	**0.5**
iteration = 3	yes	**88.4%**	88.8%	88.6%	**0.6**	0.6
iteration = 4	yes	88.3%	88.6%	88.5%	**0.6**	0.6

vehicles on a single image without any temporal information, as shown on the testing set.

In addition, we design an experiment to discover the effects of α and β in the proposed spatio-temporal analysis. Typically, we set $\alpha = \beta$. On the Training Set of vehicle detection, after iteration = 1, we try different settings: $\alpha = \beta = 3, 5, 7, 9, 11,$ and the standard deviation of the results under different settings is just 0.68%, which shows that the proposed method is not sensitive to the setting of α and β in that range. Spatio-temporal consistency for each object in the video is unknown, so it is unreasonable to set too big or too small values to α and β. We recommend to use some values between 3 and 11 for different tasks. In this paper, all the experiments use 5 for this hyper-parameter.

4.4. Experimental results on pedestrian detection

To further test the proposed method, we run it on the public image sequence PETS09-S2L1 of the 2D Multiple Object Tracking 2015 benchmark [53]. The public image sequence PETS09-S2L1 contains 795 images, where each image contains several pedestrians with the resolution of 576 × 768 pixels. The benchmark provides the public detection by a pre-trained human detector based on aggregated channel features (ACF) [54].

We use the provided public detection as the initialization, i.e., iteration = 0, to start the proposed method. The result is shown in Table 3. The proposed method can still significantly improve the initialized detection from 83.1% F-measure to 88.5% F-measure, while there are some variances in the algorithm convergence due to the complexity in pedestrian movement and tracking difficulties. Pedestrian detection is more complex than the fiber detection and vehicle detection. The dynamic model of pedestrian movement is more complicated, unknown and not very smooth, due to the occlusions and interactions. The deformation change of pedestrian shape sometimes can be large, because of the occlusions, view changes, hand and gait changes, etc. Therefore, tracking becomes difficult in the challenging situation for pedestrians. The

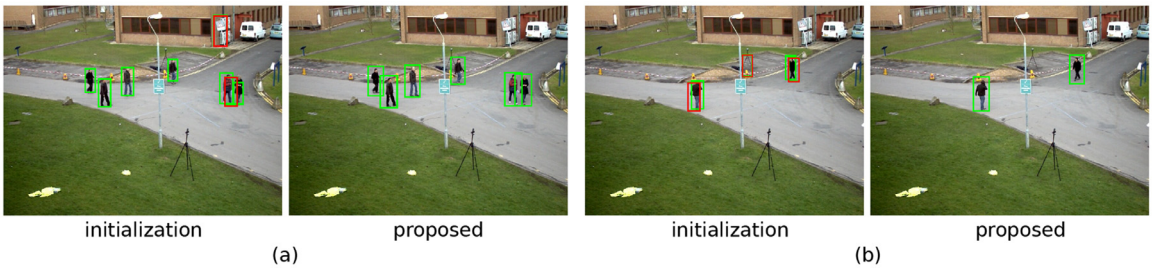


Fig. 9. Illustration of pedestrian detection of the proposed method on the public image sequence PETS09-S2L1 [53]: (a) one example image's detection by initialization [54] (left) and the proposed method (right), (b) another example image's detection by initialization [54] (left) and the proposed method (right). Pedestrians are detected as green bounding boxes. Red boxes highlight the false positive errors.

pedestrian detection demos on two example images are displayed in Fig. 9. It is obvious that the proposed method can significantly reduce the false positive errors.

5. Conclusion and future work

In this paper, we proposed an easy-to-hard learning method to improve the object detections initialized by some weak object detectors. The proposed method alternately run CNN model and object tracking algorithm in several iterations to improve object detection. A proposed unsupervised spatio-temporal analysis is used to involve more complex examples into the training set (a video), so the proposed method could incrementally improve the detection performance. The proposed method takes an image sequence as input without requiring any manual annotations, and after training, a well trained CNN model is obtained which can also be used to detect objects in a single image. The experimental results on three different tasks, i.e., fiber detection, vehicle detection and pedestrian detection, show that the proposed method is accurate and effective in object detection with only some weak supervisions.

The proposed method works well when the interested objects have little inter-occlusions, feature changes and split/merge difficulties in the image sequence, where the spatio-temporal consistency in the image sequence is well preserved and reliable. In the future, we might make efforts to modify the proposed method to solve the challenging problem which has heavy inter-occlusions, large feature changes and split/merge difficulties in the image sequences.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Hongkai Yu: Writing - original draft, Software, Methodology, Conceptualization. **Dazhou Guo:** Data curation, Methodology, Software. **Zhipeng Yan:** Visualization, Software. **Lan Fu:** Software. **Jeff Simmons:** Data curation, Methodology. **Craig P. Przybyla:** Data curation, Methodology. **Song Wang:** Writing - original draft, Writing - review & editing, Supervision.

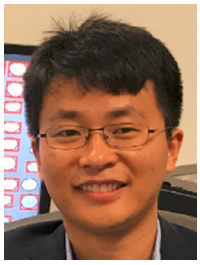
Acknowlgedgments

This work was supported in part by UES Inc./AFRL-S-901-486-002, NSF-1658987, NCPTT-P16AP00373 and NVIDIA GPU Grant.

References

[1] H. Yu, H. Yu, H. Guo, J. Simmons, Q. Zou, W. Feng, S. Wang, Multiple human tracking in wearable camera videos with informationless intervals, Pattern Recognit. Lett. 112 (2018) 104–110.
[2] Q. Zou, H. Ling, S. Luo, Y. Huang, M. Tian, Robust nighttime vehicle detection by tracking and grouping headlights, IEEE Trans. Intell. Transp. Syst. 16 (5) (2015) 2838–2849.
[3] Y. Lin, K. Abdelfatah, Y. Zhou, X. Fan, H. Yu, H. Qian, S. Wang, Co-interest person detection from multiple wearable camera videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4426–4434.
[4] H. Guo, K. Zheng, X. Fan, H. Yu, S. Wang, Visual attention consistency under image transforms for multi-label image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[5] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[6] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[7] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[8] H. Yu, Y. Zhou, J. Simmons, C. Przybyla, Y. Lin, X. Fan, Y. Mi, S. Wang, Group-wise tracking of crowded similar-appearance targets from low-continuity image sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[9] Y. Zhou, H. Yu, J. Simmons, C.P. Przybyla, S. Wang, Large-scale fiber tracking through sparsely sampled image sequences of composite materials, IEEE Trans. Image Process. 25 (10) (2016) 4931–4942.
[10] S. Li, H. Yu, J. Zhang, K. Yang, R. Bin, Video-based traffic data collection system for multiple vehicle types, IET Intel. Transp. Syst. 8 (2) (2013) 164–174.
[11] Y. Cui, J. Zhang, Z. He, J. Hu, Multiple pedestrian tracking by combining particle filter and network flow model, Neurocomputing 351 (2019) 217–227.
[12] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 41–48.
[13] V. Cirik, E. Hovy, L.-P. Morency, Visualizing and understanding curriculum learning for long short-term memory networks, 2016. arXiv preprint arXiv: 1611.06204.
[14] J. Hong Yoon, C.-R. Lee, M.-H. Yang, K.-J. Yoon, Online multi-object tracking via structural constraint event aggregation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1392–1400.
[15] V. Pătrăucean, P. Gurdjos, R.G. Von Gioi, A parameterless line segment and elliptical arc detector with enhanced ellipse fitting, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 572–585.
[16] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1, 2005, pp. 886–893.
[17] P. Hu, D. Ramanan, Finding tiny faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[18] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1.
[19] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, S. Wang, Deepcrack: learning hierarchical convolutional features for crack detection, IEEE Trans. Image Process. 28 (3) (2019) 1498–1512.
[20] Q. Zou, L. Ni, Q. Wang, Q. Li, S. Wang, Robust gait recognition by integrating inertial and rgbd sensors, IEEE Trans. Cybern. 48 (4) (2017) 1136–1150.
[21] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2018) 1531–1544.
[22] Z. Zhang, Q. Zou, Y. Lin, L. Chen, S. Wang, Improved deep hashing with soft pairwise similarity for multi-label image retrieval, IEEE Trans. Multimedia (2019) 1–13.
[23] Y. Li, M. Paluri, J.M. Rehg, P. Dollár, Unsupervised learning of edges, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1619–1627.

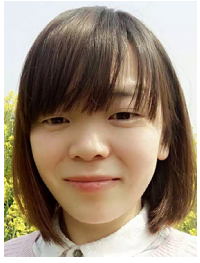
[24] R. Goroshin, J. Bruna, J. Tompson, D. Eigen, Y. LeCun, Unsupervised learning of spatiotemporally coherent metrics, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4086–4093.
[25] D. Zhang, J. Han, Y. Zhang, Supervision by fusion: towards unsupervised learning of deep salient object detector, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.
[26] H.-Y. Lee, J.-B. Huang, M. Singh, M.-H. Yang, Unsupervised representation learning by sorting sequences, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.
[27] I. Croitoru, S.-V. Bogolin, M. Leordeanu, Unsupervised learning from video to detect foreground objects in single images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.
[28] X. Wang, A. Gupta, Unsupervised learning of visual representations using videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2794–2802.
[29] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, F. Porikli, See more, know more: unsupervised video object segmentation with co-attention siamese networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, p. 1.
[30] P. Tang, X. Wang, Z. Huang, X. Bai, W. Liu, Deep patch learning for weakly supervised object classification and discovery, Pattern Recognit. 71 (2017) 446–459.
[31] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, A.L. Yuille, Pcl: proposal cluster learning for weakly supervised object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (1) (2018) 176–191.
[32] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, M.-H. Yang, Deep regression tracking with shrinkage loss, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 353–369.
[33] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for multitarget tracking, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 58–72.
[34] D. Oneata, J. Revaud, J. Verbeek, C. Schmid, Spatio-temporal object detection proposals, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 737–752.
[35] J. Meng, J. Yuan, J. Yang, G. Wang, Y.-P. Tan, Object instance search in videos via spatio-temporal trajectory discovery, IEEE Trans. Multimedia 18 (1) (2016) 116–127.
[36] X. Liang, Y. Wei, L. Lin, Y. Chen, X. Shen, J. Yang, S. Yan, Learning to segment human by watching youtube, IEEE Trans. Pattern Anal. Mach. Intell. 39 (7) (2017) 1462–1468.
[37] W. Wang, J. Shen, R. Yang, F. Porikli, Saliency-aware video object segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2017) 20–33.
[38] W. Wang, J. Shen, F. Porikli, R. Yang, Semi-supervised video object segmentation with super-trajectories, IEEE Trans. Pattern Anal. Mach. Intell. 41 (4) (2019) 985–998.
[39] W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, IEEE Trans. Image Process. 24 (11) (2015) 4185–4196.
[40] C. Feichtenhofer, A. Pinz, A. Zisserman, Detect to track and track to detect, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.
[41] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng, J. Wang, Object detection in videos by high quality object linking, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1.
[42] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, S. Yan, Stc: a simple to complex framework for weakly-supervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2016) 2314–2320.
[43] S. Song, H. Yu, Z. Miao, D. Guo, W. Ke, C. Ma, S. Wang, An easy-to-hard learning strategy for within-image co-saliency detection, Neurocomputing 358 (2019) 166–176.
[44] C.L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 391–405.
[45] M.L. Comer, E.J. Delp, The em/mpm algorithm for segmentation of textured images: analysis and further experimental results, IEEE Trans. Image Process. 9 (10) (2000) 1731–1744.
[46] Y. Xie, Q. Ji, A new efficient ellipse detection method, in: Proceedings of the International Conference on Pattern Recognition, 2002, pp. 957–960.
[47] H. Qin, J. Yan, X. Li, X. Hu, Joint training of cascaded cnn for face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3456–3465.
[48] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast r-cnn for pedestrian detection, IEEE Trans. Multimedia 20 (4) (2017) 985–996.
[49] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014) 1–14.
[50] H.W. Kuhn, The hungarian method for the assignment problem, Nav. Res. Logist. Q. 2 (1–2) (1955) 83–97.
[51] UES, https://www.ues.com/robo-met-3d/, 2018.
[52] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[53] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, K. Schindler, MOTCHallenge 2015: towards a benchmark for multi-target tracking, 2015, pp. 1–15. arXiv: 1504.01942")1504.01942[cs]
[54] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532–1545.



Hongkai Yu received the Ph.D. degree in computer science and engineering from University of South Carolina, Columbia, SC, USA, in 2018. He is a tenure-track Assistant Professor in the Department of Electrical Engineering and Computer Science at Cleveland State University, Cleveland, OH, USA, since 2020. His research interests include computer vision, machine learning, deep learning and intelligent transportation system. He is a member of the IEEE.

Dazhou Guo received the Ph.D. degree in computer science and engineering from University of South Carolina, Columbia, SC, USA in 2019. He is currently a Senior Research Scientist at PAII Inc, Bethesda, MD, USA. He received the B.S degree in Electronic Engineering from Dalian University of Technology, Dalian, China, in 2008, the M.S. degree in Information and Informatics Engineering from Tianjin University, Tianjin China, 2010. His research interests include computer vision, medical image processing, and machine learning.

Zhipeng Yan received his M.S. degree from the Department of Computer Science and Engineering, University of California, San Diego, CA in 2018. He is currently a research engineer in TuSimple. His research interests include computer vision, machine learning, deep learning, and optimization.



Lan Fu received her B.S. and M.S. from the Department of Biomedical Engineering, Yanshan University, Qinhuangdao, China, in 2011 and Tianjin University, Tianjin, China, in 2014, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of South Carolina, SC, USA. Her current research interests include computer vision and machine learning.

Jeff Simmons is a materials and imaging scientist in the Metals Branch, Materials and Manufacturing Directorate of the Air Force Research Laboratory (AFRL). He received the B.S. degree in metallurgical engineering from the New Mexico Institute of Mining and Technology, Socorro, NM, USA, and M.E. and Ph.D. degrees in Metallurgical Engineering and Materials Science and Materials Science and Engineering, respectively, from Carnegie Mellon University, Pittsburgh, PA, USA. After receiving the Ph.D. degree, he began work at AFRL as a post-doctoral research contractor. In 1998, he joined AFRL as a Research Scientist. His research interests are in computational imaging for microscopy and has developed advanced algorithms for analysis of large image datasets. Other research interests have included phase field (physics-based) modeling of microstructure formation, atomistic modeling of defect properties, and computational thermodynamics. He has lead teams developing tools for digital data analysis and computer resource integration and security. He has overseen execution of research contracts on computational materials science, particularly in prediction of machining distortion, materials behavior, and thermodynamic modeling. He has published in the Materials Science, Computer Vision, Signal Processing, and Imaging Science fields. He is a member of ACM and a senior member of IEEE.

Craig P. Przybyla is a Research Team Leader at the Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio. Specifically, he is the technical lead of the Composites Performance Research Team in the Ceramics Branch whose mission is the performance characterization and prediction of advanced polymer and ceramic matrix composites for aerospace applications. Craig and his team support both composite material technology issues related to fielded systems and the transition of next generation composite technologies for future systems. Prior to his current position, he was a Senior Materials Engineer in the Composites Branch where he worked from 2010 to 2015 as a research engineer and subject matter expert for performance characterization and modeling of ceramic matrix composites for hot section components in gas turbine engines and/or thermal protection systems for hypersonic vehicles. He earned his B.S. and M.S. degrees in the Department of Mechanical Engineering at Brigham Young University in 2004 and 2005, respectively. He was awarded his Ph.D. from the Georgie Institute of Technology in the Department of Materials Science and Engineering in 2010.

Song Wang received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2002. He was a Research Assistant with the Image Formation and Processing Group, Beckman Institute, UIUC, from 1998 to 2002. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA, where he is currently a Professor. His current research interests include computer vision, image processing, and machine learning. Dr. Wang is a senior member of IEEE and a member of the IEEE Computer Society. He is currently serving as the Publicity/Web Portal Chair for the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society and as an Associate Editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition Letters, and Electronics Letters.