DAODiS: Domain-Adapted Object Detection in Surveillance Videos

1. Contest title and abstract

Title: DAODiS: Domain-Adapted Object Detection in Surveillance Videos

Abstract: Automatic analysis of image sequences is a key component of modern visual surveillance. Although very successful object detection methods have been recently developed using deep learning methodologies, their performance dramatically suffers if applied to surveillance videos due to domain shift. Clearly, domain shift is a critical and still open issue. This challenge attempts to advance research on domain adaptation in the context of surveillance videos. Its specific goal is the adaptation of object detection algorithms trained on still-image datasets (e.g., COCO) to surveillance videos by leveraging background subtraction. Algorithms will be judged on three performance metrics and on manual effort expended to annotate additional video frames. We believe this challenge can contribute to the AVSS community by charting a new research direction for visual surveillance. Furthermore, it can contribute to various application fields such as equipment monitoring and e-commerce (new product recognition).

2. Name and contact information of the main organizer and at least 2 other expert committee members

- Atsushi Shimada (Kyushu Univ., atsushi@ait.kyushu-u.ac.jp)
- Janusz Konrad (Boston Univ., jkonrad@bu.edu)
- Vincent Charvillat (ENSEEIHT, vincent.charvillat@toulouse-inp.fr)
- Tsubasa Minematsu (Kyushu Univ., minematsu@ait.kyushu-u.ac.jp)
- Takashi Shibata (NEC Corp., t-shibata@hw.jp.nec.com)
- Yasutomo Kawanishi (Nagoya Univ., kawanishi@is.nagoya-u.ac.jp)

3. General description of the problem

Background: In video surveillance applications, dealing with domain shift is a critical and still open issue. For example, the visual appearance of a scene may depend on date, time, location and camera settings. Recently, many object-detection methods using CNNs, such as YOLO, Faster-RCNN, and Retina-net, have been proposed but their performance is sensitive to domain shift. Similarly, domain shift remains a challenge for many recently-proposed background subtraction methods, as well as methods that combine background subtraction and object detection.

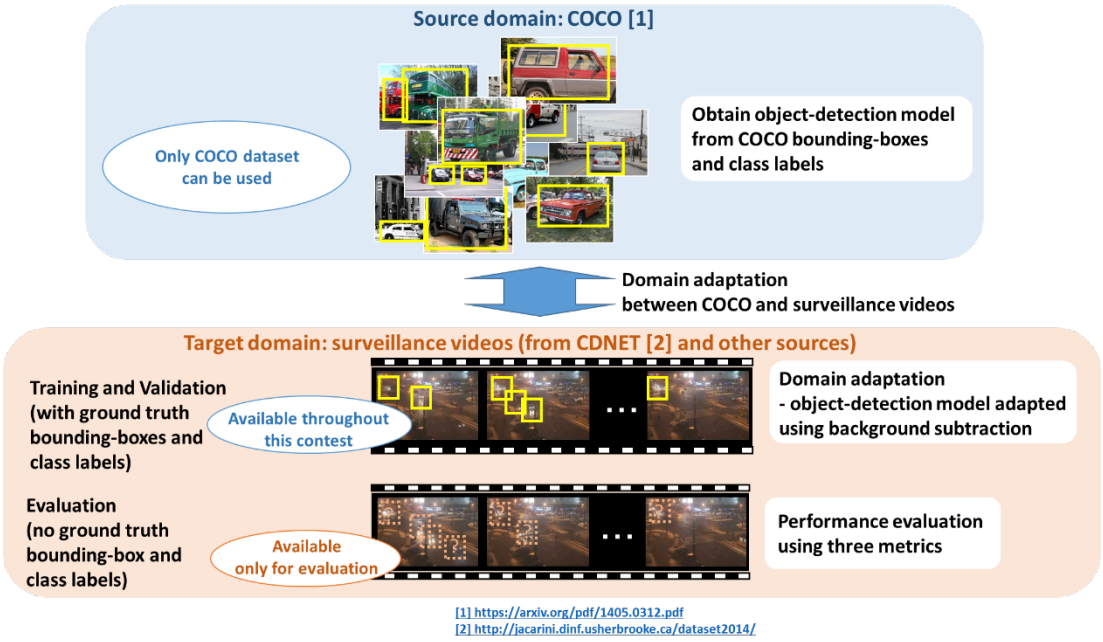
Goal and overview of this contest: This challenge attempts to advance research on domain adaptation in the context of surveillance videos. Its specific goal is the adaptation of object

detection algorithms trained on still-image datasets from COCO to surveillance videos by leveraging background subtraction methods. Algorithms will be judged on three performance metrics and on manual effort expended to annotate additional video frames.

Contributions: We believe this challenge will contribute to the AVSS community by charting a new research direction for visual surveillance. While domain adaptation is an active research area in machine learning, there is less effort in this area in video surveillance community. Additionally, this challenge can contribute to various application fields such as equipment monitoring and e-commerce (new product recognition).

4. Description of the dataset to be used

The challenge will leverage the well-known *2014 Change detection dataset* (CDNET, http://jacarini.dinf.usherbrooke.ca/dataset2014/) with permission from its authors. The selection of videos from CDNET is guided by difficulties encountered in the captured scenes (low light, rain, snow, acquisition modality, etc.) We are in the process of annotating selected CDNET videos by drawing a bounding box around each visible object and assigning a label (e.g., car. person). The challenge dataset will be divided into three subsets: 1) training, 2) validation, and 3) evaluation. We will provide training and validation sequences with bounding-box coordinates and object labels for algorithm development and evaluation sequences without such coordinates/labels for algorithm's performance assessment.



5. Description of the actual competition task

The goal of this challenge is to make baseline object detectors, such as YOLO, adapt to a new

domain by leveraging background subtraction. The idea is that background/foreground predictions can be informative for the task of object detection (e.g., an object detected in an area of predicted foreground may positively reinforce the detection). In order to fairly compare different methods, participants will be asked to develop their baseline object detectors using the COCO dataset only. During evaluation, participants will be asked to submit predictions of bounding-box coordinates and class probabilities for evaluation sequences. Details of the evaluation metrics are described in Sec. 6.

We believe that approaches to domain adaptation can be roughly categorized into two groups: 1) detection without training, and 2) detection with training. A detection method in the former category would simply integrate baseline object detection with a background subtraction method to deal with domain shift. On the other hand, a method in the latter category would re-train the baseline object detection model using target domain videos with some ground-truth bounding boxes. Participants will be asked to report the type of method submitted, i.e., with or without re-training.

6. Evaluation metrics

The performance of each method will be evaluated in terms of precision (Pre), recall (Rec), and mean average precision (mAP) scores. In particular, precision and recall are given by

$$\mathrm{Pre} = \frac{TP}{TP + FP}, \qquad \mathrm{Rec} = \frac{TP}{TP + FN},$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively, defined as follows:

- TP: IoU between prediction and ground-truth bounding boxes ≥ 0.5 and matching class labels,
- FP: FP: IoU between prediction and ground-truth bounding boxes < 0.5 or mismatched labels,
- FN: ground-truth bounding box does not overlap with any predicted bounding box.

Table 1 shows this summary.

Table 1. Summary of TP, FP, and FN.

Class label	0.5≤IoU	IoU < 0.5
Prediction = Ground truth	TP	FP ($\mathrm{FP}_{\mathrm{iou}}$)
Prediction ≠ Ground truth	FP ($\mathrm{FP}_{\mathrm{class}}$)	FP ($\mathrm{FP}_{\mathrm{class,iou}}$)

The mean average precision (mAP) is the mean of the average precision score for each class. These three metrics will be evaluated based on submitted bounding box coordinates for object detection with class label.

7. Additional information

We will ask the participants to report additional information. This information will be used for categorizing proposed methods and will NOT be used for ranking them.

1) Method type: We will ask the participants whether their object detection method uses manually-extracted training data from CDNET videos in addition to the training and evaluation data provided within the challenge, i.e., whether they extracted additional bounding boxes from CDNET videos themselves.

2) Number of annotations: In order to improve the performance of an object detection method under domain shift, it is beneficial to add new annotations to target data. However, manual labeling is a time-consuming, and therefore expensive, task. Thus, to account for additional manual labels, we will measure *Annotation Rate* as follows:

$$Anotation\ Rate = \frac{N}{N+K}$$

, where K is number of frames with (ground truth) bounding boxes provided within the challenge for training and validation while N is the number of additional frames that participants have manually annotated themselves.

8. Plan of how to organize the contest

This schedule is **tentative**.

- Dec. 1, 2019: Ad-hoc website open.
- Jan. 1, 2020: Dataset available.
- Jul. 1, 2020: Deadline for the submission of results.
- Aug. 1, 2020: Challenge results announced and of a 2-3 page report ready for distribution to participants.
- Sep. 22-25, 2020: Results of the competition are announced at the conference

9. Estimated number of participants

Five to ten groups

10. Preference of a quarter day session or a half-day session

Quarter-day session