FAST AND ROBUST MODEL FOR MULTIPLE OBJECTS TRACKING USING KEY-FRAME DETECTION AND CO-TRAINED CLASSIFIER

Phung Kim Phuong, Nguyen Quang Thi*, Nguyen Huu Hung, Dang Quang Hieu
Military Technical Academy

ABSTRACT

This paper proposes our new approach for multiple objects tracking for real-time video tracking applications. The new tracking method can improve tracking speed and reduce track fragmentation and confusion by using two convolutional neural networks to detect and distinguish the targets. This mechanism ensures real-time capability when you do not have to perform deep learning detector continuously while still ensuring constant and accurate updating of the target's position. This is called a co-training mechanism. The keyframe detection model is a Single Shot Detector that also operates as a data generator; the second neural network is a classifier that will be trained from data collected from the main detector. The tracker is presented as a combination of techniques that we named DCT (Detector-Classifier Tracker). This article will fully explain the working mechanism of DCT and presents the test results for the combined image attachment method according to the frame processing experiments on data of long range thermal imaging cameras.

Keywords: *data association; multi-object tracking; real-time tracking; convolutional neural networks; deep learning*

Received: 06/10/2020; Revised: 30/11/2020; Published: 30/11/2020

MÔ HÌNH BÁM ĐA ĐỐI TƯỢNG ĐẢM BẢO THỜI GIAN THỰC VÀ ỔN ĐỊNH CAO SỬ DỤNG KẾT HỢP BỘ PHÁT HIỆN THEO KHUNG HÌNH KHÓA VÀ BỘ PHÂN LOẠI LUYỆN ĐỒNG BỘ

Phùng Kim Phương, Nguyễn Quang Thi*, Nguyễn Hữu Hùng, Đặng Quang Hiệu
Trường Đại học Kỹ thuật Lê Quý Đôn

TÓM TẮT

Trong bài báo này, chúng tôi đề xuất một cách tiếp cận mới trong bám đa đối tượng cho các ứng dụng trên video thời gian thực. Phương pháp bám mới hướng đến khả năng đảm bảo thời gian thực và chống đứt đoạn quỹ đạo bám bằng cách sử dụng kết hợp hai mạng nơ-ron để phát hiện và phân biệt giữa các mục tiêu. Cơ chế này đảm bảo khả năng thời gian thực khi mô hình không phải thực hiện liên tục các phép tính phát hiện học sâu trong khi vẫn đảm bảo cập nhật liên tục và chính xác vị trí của mục tiêu. Chúng tôi gọi đây là cơ chế luyện đồng bộ. Mô hình thứ nhất là bộ phát hiện học sâu Single Shot Detector đồng thời hoạt động như một bộ tạo dữ liệu, mô hình mạng nơ ron thứ hai là một bộ phân loại sẽ được luyện từ dữ liệu thu thập được từ bộ phát hiện. Bộ bám đa đối tượng được xây dựng dưới dạng sự kết hợp của các kỹ thuật được chúng tôi gọi là DCT (Detector-Classifier Tracker). Bài viết này sẽ giải thích đầy đủ cơ chế hoạt động của cơ chế bám ảnh DCT và trình bày kết quả đánh giá đối với phương pháp theo sơ đồ xử lý bám ảnh kết hợp trên dữ liệu thử nghiệm của camera ảnh nhiệt tầm xa.

Từ khóa: *liên kết dữ liệu; bám đa đối tượng; bám thời gian thực; mạng nơ ron tích chập; học sâu*

Ngày nhận bài: 06/10/2020; Ngày hoàn thiện: 30/11/2020; Ngày đăng: 30/11/2020

* Corresponding author. *Email: thinq.isi@lqdtu.edu.vn*

https://doi.org/10.34238/tnu-jst.3678

1. Introduction

The main task of the video tracking process is associating existing objects with the new image and creating a continuous spatial trajectory of moving objects. Modern trackers use different learning models to represent the appearance, geometric position and movement of the object [1]. Tracking is then addressed as finding the most likely geometrical parameters, usually as a bounding box of the object in the new image. Before the implementation of deep CNN, researchers had many improvements for tracking quality using appearance-based discriminative features [2]. To cope with occlusion, objects and environment changes, multiple similar objects, Kalman's state estimator and data association methods are usually implemented as part of the tracking algorithm [3].

Recent improvements in applying deep convolutional neural networks (CNN) to target detection and tracking, detect-to-track and track-to-detect with deep learning is a promising solution to improve tracking accuracy but also add a big amount of computational cost. The reason is they focused on particular quality metrics like MOTA (Multiple Object Tracking Accuracy), MOTP (Multiple Object Tracking Precision) [1], that was tested on most popular object classes (human, face, cars). In many practical scenarios, the interested object may not be the popular ones in the open datasets, deep CNN detectors may fail to keep the accuracy inside the acceptable threshold, and tracking quality can be decreased. Deep CNN tracking methods are mostly based on CNN detection modules. In the same way, the tracking accuracy highly depends on the accuracy of the detector.

***Contributions*:**

We present a method based on the combination of traditional tracking techniques and deep CNN approaches. The basic metric is not only based on MOTA but also fitted to a real world implementation. Our tracking method is expected to introduce better tracking quality with the following requirements:

- Scalable and adaptive to hardware capability, the method can be customized easily to adapt with realistic conditions, the processing architecture is based on encapsulated functional modules that can be replaced and customized independently.

- High robustness and less vulnerability to track loss, fragmentation, with ability to distinguish multiple objects of the same class.

- The learning strategy of objects models will be combined from both online training and offline training.

Our approach differs from both traditional tracking techniques and more recent deep CNN methods.

We present a specific experimental implementation of the track method and test the method with realistic data to leverage the efficiency of the method in highly challenging conditions.

2. Challenging MOT problems

A realistic MOT scenarios, a specific tracker can have many limitations, this paper is addressing most popular challenges for highly challenging MOT applications:

- Most offline-trained deep CNN trackers are based on visual based object models, which are vulnerable to long term occlusion or objects leaving and reentering the field of view, this usually leads to track fragmentation, when one track is divided into temporal fragments by visual contact interruption. Confusion between objects is also a weak spot of CNN models, visually similar objects have very weak distinguishing features and they should be tracked with respect to motion and behavior models [3], [4].

- On the other side, state estimation models and other "shallow" features models do not have stable performance with stochastic movement of objects and the camera, visual changes of objects, light and environment.

- Convolution filter based trackers were basically low-level visual models (histogram of gradients, optical flow, kernelized correlation) of objects, which were usually updated, or trained online using new images of the object. In practical scenarios with fast changes of object visual features, due to the lack of general pre-trained model, they are eventually float away by similarity of object and background, or shrink to just a part of the object [5].

Da Zhang et al. [6] proposes a neural-network tracker that combines convolutional and recurrent networks with reinforcement learning algorithms in order to predict objects movement and thus improve state estimation.

The initial motivation for our line of research is combining multiple models in one tracking system to cope with weaknesses of each method and thus improve robustness, while keeping the model optimized for real time tracking applications.

3. Novel tracking frame processing chain

Correlation filters have proved to be competitive with far more complicated approaches when using only a fraction of the computational power, at hundreds of frames-per-second. The proposed processing chain is a combination of several processing algorithms aggregated in a video processing chain which is illustrated in Figure 1. The video processing chain contains following key elements:

- Detector: pre-trained deep learning protection model based on deep CNN architectures. Detector can detect, classify and locate multiple objects in one video frame. This is an expensive processing step and may require hardware acceleration, for real-time requirement, detection only applied to key frames, and the frequency of key frames can vary depending on the processing capability of the hardware. For specific implementation, deep CNN detector and be used in combination or be replaced with other detectors to fit the requirements.

- Track management module: a combination of several functional modules that can operate on video frames with or without detection labels. Operational mechanism of this module will be explained in detail in the next part of this paper.

- Tracking decision: a set of existing tracks that updates after each video frame, each track has the current state vector and history. The state vector contains tracks geometrical parameters (center, bounding box, size, aspect ratio, speed, direction...)

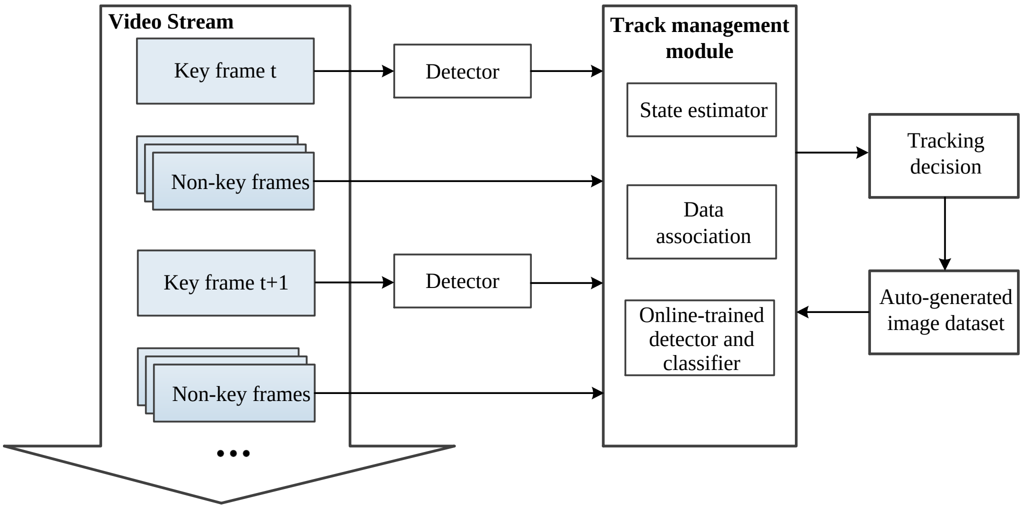


Figure 1. *Temporal sequential model of video frame processing*

- Auto-generated image dataset: a labeled image dataset containing image inside the bounding box of all tracks since the beginning of the processing. This dataset is labeled with target identifications and states before feeding to the correlation filter (CF) detector and online-trained classifier as a training dataset.

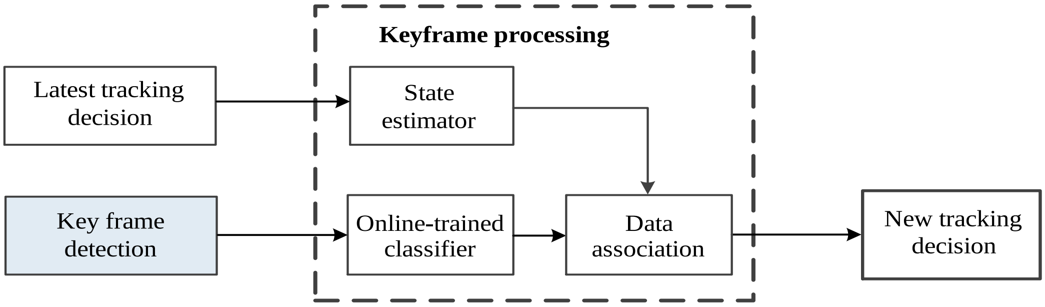
The structure of the tracking manager(TM) module is the main focus of this paper. TM is a data processing block with input data as a sequential stream of images. The output data of TM is the tracking decision. The main purpose of the tracking manager is combining and synchronizing multiple functional modules in one data association module.

Figure 2 explains the work of the tracking manager module in detail.

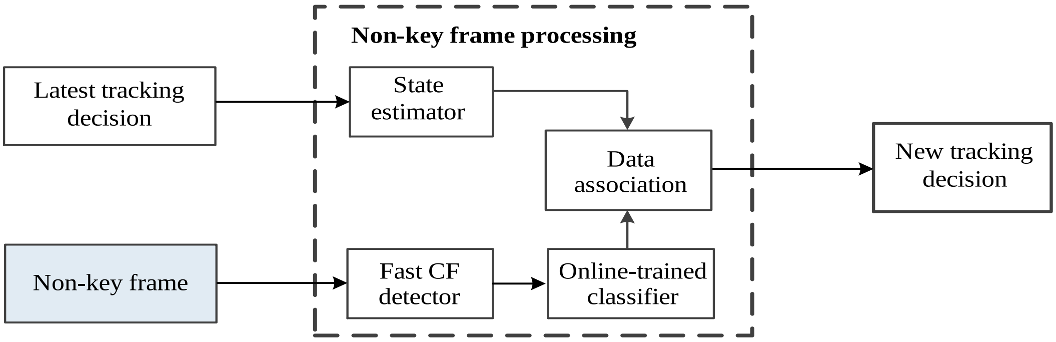
The key elements of the tracking manager can be listed as follows:

- The first element is the state estimator. The state estimator estimates the current state of objects based on its previous measurements with a level of uncertainty and sends it to the data association module. In this case, bounding boxes of objects serve as measurements. In the first frames, the state estimator may have nothing in output, and usually the more instances of the objects are detected, the less uncertainty the output of the state estimator becomes. In this paper, Kalman Filter is chosen for our experimental model. The state for each target is estimated recursively online by applying Kalman filter per each target and feeding to the DA module as current state of the tracked targets.

- The second element is the data association module. The data association module is a method to solve optimal assignment problems, or global data association [2]. In general, DA takes input as a set of new data measurements and builds a score matrix of assignment variances that includes all possible assignments and then finds a min-cost matching. In this case, DA matches new bounding boxes from detector to existing targets state in the state estimator.



a) Data association with input data labeled by detector (keyframes)



b) Data association with input data from raw frame

Figure 2. *Data association operations in two tracking mode*

- The third element is an online trained classifier that is trained from the auto-generated dataset. The classifier processes bounding box images from the detector and adds a value that indicates the probability of this bounding box to belong to existing tracked objects based on visual features. The online trained classifier uses 3-layers CNN with pyramidal architecture and the last dense layer is the maximum number of tracks that the model can process.

- The fourth element is the correlation filter (CF) detector. CF implements the function of a fast detector for non-key frames, which can detect the proposed location of an object with a template generated by previous detections. Many variants of correlation filters have been used for fast object tracking. In the experiment section of this paper, we will implement KCF as a CF detector. KCF uses the technique described in [5] for fast object tracking.

While processing raw video frames, the measurement update for DA is output data from the CF and classifier. Applying correlation filters on the new image, CF now serves as a detector that can outline bounding boxes of objects in a raw video frame that are visually similar to tracked targets. The output bounding boxes of CF will be classified by the classifier. The classifier adds a feature vector to the bounding boxes that is proportional to likelihood of belonging to each existing object.

$$F_k = [f_{ki}]\,;\ (i = 1..n) \tag{1}$$

where:

n- number of tracked objects

f_{ki}- likelihood of object bounding box k to belong to object class i.

Thus, before being fed into DA as new measurement data, detection data from raw frames are presented as a set of objects with feature vector:

$$V_k = [x, y, w, h, F, C] \tag{2}$$

where:

x- horizontal coordinate of object

y- vertical coordinate of object

w- object horizontal size

h- object vertical size

After each update, the state estimator represents each object with a state vector:

$$S_i = \{x, y, w, h, dx, dy\} \tag{3}$$

where:

dx- horizontal speed of object

dy- vertical speed of object

Based on new feature vectors and existing state vectors, the DA module generates the matrix of possible object assignments and estimates the optimal assignment decision, which is also the tracking decision. The DA can be formulated as a $m \times n$ sized matrix, where *m*: number of new detection, *n*: number of tracked objects. Each matrix element

$$R_{k,i} = \ p(V_k, S_i\,) \tag{4}$$

represents the probability metric of assigning detection V_k to track S_i.

$R_{1,1}$	$R_{1,2}$	...	$R_{1,n}$
$R_{2,1}$	$R_{2,2}$	...	$R_{2,n}$
...	...	...	...
$R_{m,1}$	$R_{m,2}$	...	$R_{m,n}$

***a)** Data association table*

$D_{1,1}$	$D_{1,2}$	...	$D_{1,n}$
$D_{2,1}$	$D_{2,2}$	...	$D_{2,n}$
...	...	...	...
$D_{m,1}$	$D_{m,2}$	...	$D_{m,n}$

***b)** Data association solution table*

Figure 3. *Matrix formulation of data association*

Figure 3 demonstrates the formulation of the data association matrix and the data association solution. The metric of association probability $R_{k,i}$ is a combination of visual classification score and location prediction score.

The association solution has three constraints as following:

$$D_{k,i} \in [0, 1] \tag{5}$$

While $D_{k,i}$ = 1 means the detection k is assigned to track i and $D_{k,i}$ = 0 means the detection $k$is not assigned to track i.

$$\forall k \in [1..m], \sum_{1}^{n} D_{k,i} \leq 1 \tag{6}$$

This constraint indicates that for each detection, there could be no more than 1 assignment to track.

$$\forall i \in [1..n], \sum_{1}^{m} D_{k,i} \leq 1 \tag{7}$$

Similarly, for each track, there could be no more than 1 assignment to a new detection. If a new detection is not assigned to any of existing tracks, it generates a new track, otherwise, if a track is not assigned to any new detection, it keeps waiting in the next updates until a predefined timeout runs out and the missed track is removed from the memory.

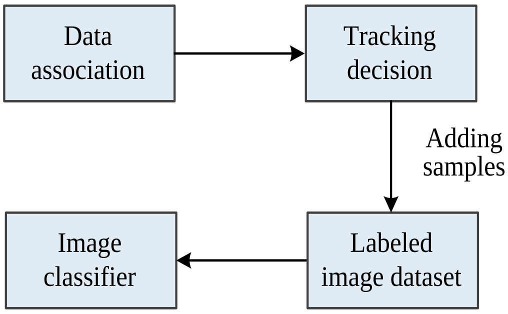


Figure 4. *Training scheme of the classifier*

Figure 4 shows the mechanism of the training process for image classifier models. Based on each tracking decision, an image dataset is updated and serves as a training dataset of the tracking module that keeps training data for the classifiers.

4. Experiments

4.1. Metrics

The main purpose of DCT is to focus on robustness, for this reason, the main metric used for comparison between tracking models in our experiments is the rate of track fragmentation and confusion.

The evaluation processing was applied to DCT, SORT and KCF with the same main detector - MobileSSD for comparison. Data used for experiment was collected from a long-range pan-tilt thermal image of sea ships. The similar shape of ships and camera movement are the most challenging factors that failed conventional trackers.

4.2. Experimental multiple ship tracking (MST) solution

We evaluates the performance of our tracking implementation using a multiple ship tracker (MST) specially customized for real time video tracking of long range objects:

- Mobilenet SSD as keyframe object detector,
- Hungarian Algorithm as data association module,
- KCF as a fast detector.

The keyframe object detector is a Mobilenet SSD (28 layers) that uses the architecture of Single Shot Detector, pre-trained with 80 classes COCO labeled dataset and trained with a database of sea ship images using transfer learning. The keyframe detection model was trained to detect sea ships in long-range thermal imaging with cooled sensors.

The tracking implementation was tested on long range tracking video with moving camera and moving objects, and short-term occlusion and showed stable tracking results compared to standard tracking techniques. The first testing video, as shown in Figure 5, despite the simple model architecture, Mobilenet SSD performance is almost the same as deeper networks at detecting objects with low resolution image.

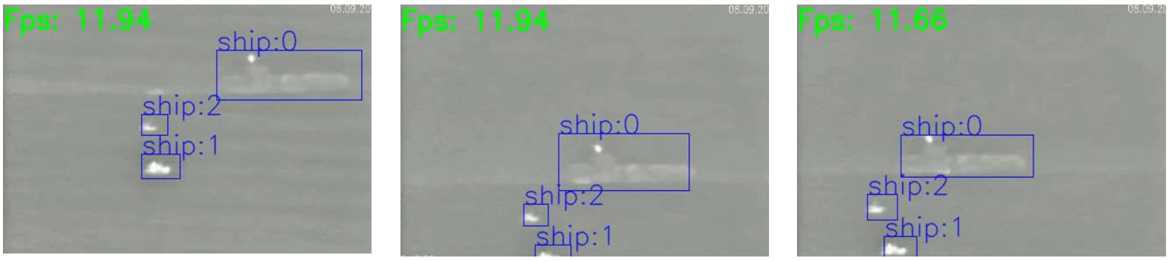


Figure 5. *Series of keyframes and bounding boxes of detected objects*

Table 1. *Fragmentation and confusion comparison of DCT with other tracking system*

Method	Fragmentation	Confusion
DCT (proposed)	12	3
SORT	17	8
KCF	33	9

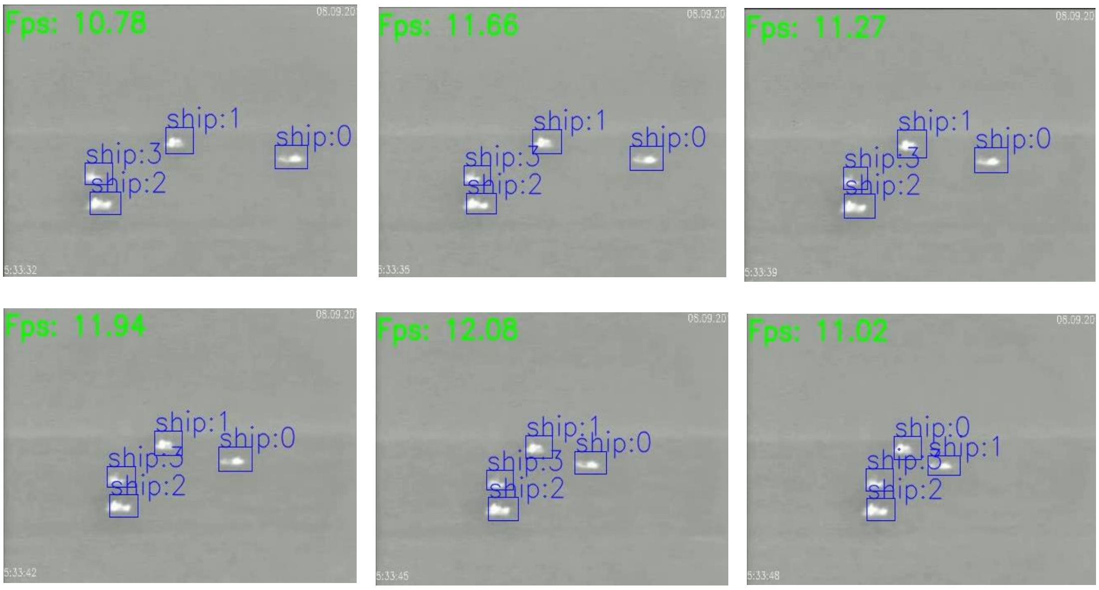


Figure 6. *Associated objects tracks between key-frames*

The test case demonstrated in Figure 5 demonstrates a typical challenge for long range pan tilt cameras, beside the movement of objects, the camera itself can move with significant speed. Trackers that mostly rely on detection frequently suffer from fragmentation in this scenario. Table 1 demonstrates the DCT model performance to keep the right tracking object with significantly reduced count of track fragmentation.

Table 1 compares the robustness of DCT for other tracking systems on 22 minutes of video data and 24 ground-truth tracks.

Another challenging situation is confusion between multiple similar moving objects. Figure 6 demonstrates a test case for long range sea ship tracking when four tracked moving objects have similar shape and size. As shown in the last image, DCT was able to track objects as they move in different directions and there is one object confusion between object 1 and object 0 as shown in the last image.

5. Conclusions

This paper introduces a tracking model based on multiple AI models that significantly improves robustness and speed of multiple target tracking in videos. The key difference of the method is the combination between three elements in tracking: (1) deep CNN keyframe object detector; (2) construction of correlation filters for detecting objects based on keyframe detections; (3) classification between multiple objects using a simple 3-layer CNN classifier.

Our tracking model differs from previous CNN and CF based trackers in two important ways. First, every keyframe creates a trusted status of existing objects that generates labeled image dataset of objects and kernel for correlation filter. For every non-key frame, the fast detector and data association module will keep following the exact position of the objects until the next keyframe. For all frames, including key and non-key, the data association algorithm will combine multiple status vectors and generate a tracking decision.

For the moment, the testing dataset for implementation in long range thermal ship images is not big enough for MOTA and MOTP metrics. Experiments on our testing data only use the count of fragmentation and confusion as a comparison metric. The DCT model is shown to be more robust to track fragmentation compared to conventional Simple Online Real Time Tracking (SORT) and Kernelized Convolution Filter (KCF) algorithms.

We believe that the DCT model using key-frame detection and co-training classifiers will be more accurate than conventional tracking approaches on many real-time tracking tasks with extreme conditions. We have already extended this work to popular tracking datasets to test the performance. In the future, we hope to test MMT in a variety of scenarios related to real time multiple object tracking.

REFERENCES

[1]. A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, “A benchmark for multi-object tracking,” May 2016, *arXiv:1603.00831v2 [cs.CV]* 3.

[2]. Q. Yu, T. B. Dinh, and G. Medioni, “Online Tracking and Reacquisition using Co-trained generative and discriminative trackers,” *Proceedings of 10th European Conference on Computer Vision*, Marseille, France, October 12-18, 2008, Part II, pp. 678-691.

[3]. A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. “Simple online and realtime tracking,” Jul. 2017, *arXiv:1602.00763v2 [cs.CV]*, vol.7.

[4]. C. Feichtenhofer, A. Pinz, and A. Zisserman. “Detect to track and track to detect,” Mar 2018, *arXiv:1710.03958v2 [cs.CV] 7*.

[5]. J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. “High-speed tracking with Kernelized correlation filters,” Nov. 2014, *arXiv:1404.7584v3 [cs.CV] 5*.

[6]. D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. “Visual object tracking using adaptive correlation filters,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010.

[7]. D. Zhang, H. Maei, X. Wang, and Y.-F. Wang, “Deep reinforcement learning for visual object tracking in video,” Apr. 2017, *arXiv:1701.08936v2 [cs.CV] 10*.