

# **VTrack: A NOVEL VISUAL OBJECT TRACKING BENCHMARK DATABASE**

THESIS SUBMITTED IN  
PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE DEGREE OF

**MASTER OF ENGINEERING  
IN  
BIOMEDICAL ENGINEERING**

By

**SUPRIYA MONDAL**

**Registration No.: 154631 of 2020-2021**

**Examination Roll No.: M4BMD22006**

Under the guidance of

**Dr. PIYALI BASAK**

School of Bioscience and Engineering

&

**Prof. SHELI SINHA CHAUDHURI**

Department of Electronics and Telecommunication Engineering

**FACULTY OF INTERDISCIPLINARY STUDIES LAW & MANAGEMENT**

**JADAVPUR UNIVERSITY**

**KOLKATA 700032**

## **DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS**

---

I hereby declare that this thesis contains a comprehensive survey on previous works as well as some original technical contribution by the undersign candidate, as part of his Masters of Engineering in Biomedical Engineering in the school of Bio-Science and Engineering. All information in this document has been obtain and presented in accordance with academic rules and ethical conduct. I also declare that I have thoroughly cited and referenced all material and findings which are not original to this research, as provided by these rules and conduct.

**Name: Supriya Mondal**

**Exam Roll No: M4BMD22006**

**Thesis Title: VTrack: A NOVEL VISUAL OBJECT TRACKING  
BENCHMARK DATABASE**

---

Signature of Candidate

**FACULTY OF INTERDISCIPLINARY STUDIES LAW &  
MANAGEMENT  
JADAVPUR UNIVERSITY**

---

**CERTIFICATE**

This is to certify that the thesis title- “VTrack: A NOVEL VISUAL OBJECT TRACKING BENCHMARK DATABASE” has been carried out by Supriya Mondal bearing Class Roll No. **002030201007**, Examination Roll No.: **M4BMD22006** and Registration No. 154631 of 2020-2021, under our guidance and supervision and be accepted in partial fulfillment of the requirement for the degree of Master of Engineering in Biomedical Engineering in the School of Bio-Science and Engineering.

---

Dr. Piyali Basak

Supervisor

School of Bio-Science & Engineering

---

Prof. Sheli Sinha Chaudhuri

Co-Supervisor

Electronics & Tele-communication Engineering

---

Dr. Piyali Basak

Director

School of Bio-Science & Engineering

---

Prof. Subenoy Chakraborty

Dean

Faculty of Interdisciplinary Studies Law & Management

Jadavpur University

**FACULTY OF INTERDISCIPLINARY STUDIES LAW &  
MANAGEMENT  
JADAVPUR UNIVERSITY**

---

**CERTIFICATE OF APPROVAL**

The thesis titled, “**VTrack: A NOVEL VISUAL OBJECT TRACKING BENCHMARK DATABASE**” is hereby approved as a creditworthy study of an engineering subject conducted and presented satisfactorily to warrant its acceptance as a precondition to the degree for which it was submitted. It is understood that the undersigned does not automatically support or accept any argument made, opinion expressed, or inference drawn in it by this approval, but not only approves the thesis for the reason for which it was submitted.

**Committee on Final  
Examination for Evaluation of  
the Thesis**

---

Signature of External Examiner

---

Signature of Supervisor

---

Signature of Co-Supervisor

## ACKNOWLEDGEMENTS

---

This thesis titled, “VTrack: a novel visual object tracking benchmark database” is the result of the work whereby I have been accompanied and supported by many people, my supervisor, my friends and my seniors. It is a pleasant aspect that now I have the opportunity to express my gratitude to all of them.

Firstly, I would like to express my gratitude to my co-supervisor, **Prof. Sheli Sinha Chaudhuri** without whose constant guidance and support this thesis work would not have been possible.

I would like you to express my gratitude to my supervisor **Dr. Piyali Basak**, Director, School of Bio-Science and Engineering and all faculties, lab technicians, my seniors and my friends of the department for being a strong support to me in this journey.

I am thankful to **Prof. Manotosh Biswas**, Head of the Department, Electronics and Telecommunication Engineering Department, all faculties, lab technicians of Electronics and Telecommunication Engineering Department for directly or indirectly helping me to carry out my thesis work.

I am very much obliged to my friend **Asfak Ali** and my senior **Avra Ghosh** for guiding me in this project. This work would not have been possible without their support.

I would like to express my gratitude to my parents and my elder sister without whose sacrifice, unconditional love and support, I could not have achieved anything in my life.

---

Supriya Mondal  
Jadavpur University  
Kolkata 700032

# Thesis

## ORIGINALITY REPORT

2%

SIMILARITY INDEX

### PRIMARY SOURCES

- |   |  |                 |
|---|--|-----------------|
| 1 | <a href="http://www.image-net.org">www.image-net.org</a><br>Internet   | 85 words — 1%   |
| 2 | <a href="http://gcris.iyte.edu.tr">gcris.iyte.edu.tr</a><br>Internet   | 38 words — < 1% |
| 3 | <a href="http://link.springer.com">link.springer.com</a><br>Internet   | 30 words — < 1% |
| 4 | Yangrun Hu, Ruikang Wu, Yanjun Gu. "Action-Improved Actor-Critic Tracking for Accurate Object Tracking", Journal of Physics: Conference Series, 2021<br>Crossref                         | 28 words — < 1% |
| 5 | Seyed Abbas Daneshyar, Nasrollah Moghadam Charkari. "Biogeography based optimization method for robust visual object tracking", Applied Soft Computing, 2022<br>Crossref                 | 18 words — < 1% |
| 6 | Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, Huchuan Lu. "Transformer Tracking", 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021<br>Crossref | 14 words — < 1% |
| 7 | <a href="http://researchbank.rmit.edu.au">researchbank.rmit.edu.au</a><br>Internet   | 14 words — < 1% |

## THESIS ORGANIZATION

<b>Chapter No: Name</b>	<b>Page No.</b>
Abstract	1
Chapter 1: Introduction	2-3
Chapter 2: Background Study & Literature Survey	4-11
Section 2.1. Background study	4-8
Section 2.2.Existing Literature	8-11
Chapter 3: Existing challenges & Thesis motivation	12-14
Section 3.1.Existing challenges in visual object tracking domain	12-13
Sub-Section 3.1.1. Accuracy related challenges	12-13
Sub-Section 3.1.2. Efficiency & Scalability related challenges	13
Section 3.2. Motivation behind conducting this thesis work	13-14
Chapter 4: Comprehensive survey on existing visual tracking benchmark databases	15-34
Chapter 5: VTrack database	35-50
Section 5.1.Examples of frames included in VTrack database	36-50
Chapter 6: Conclusion & Future Scope	51
References	52-54

## LIST OF FIGURES

<b>Figure No. Figure caption</b>	<b>Page No.</b>
Figure.1.1. Generalized representation for object detection and tracking	3
Figure 2.1. An example of a frame containing included in VTrack database with HV attribute	4
Figure 2.2. An example of a frame containing included in VTrack database with NLM attribute	5
Figure 2.3. An example of a frame containing included in VTrack database with LS attribute	5
Figure 2.4. An example of a frame containing included in VTrack database with IPR attribute	6
Figure 2.5. An example of a frame containing included in VTrack database with LR attribute	6
Figure 2.6. An example of a frame containing included in VTrack database with BC attribute	7
Figure 2.7. An example of a frame containing included in VTrack database with LL attribute	7
Figure 2.8. An example of a frame containing included in VTrack database with IWC attribute	8
Figure.2.9.Network architecture of the Siamese Relation Network	9
Figure.2.10.An example of adversarial blur attack against a deployed visual tracker, SiamRPN++ [14]. In the case, two adjacent frames are fed to the designed adversarial blur attack to generate an adversarial blurred frame which misleads the tracker and produces an inaccurate response map.	10
Figure.4.1. Few examples of annotated image frames present in PASCAL VOC datasets	17
Figure.4.2. Examples of some frames belonging to ILSVRC 2013 database	20
Figure.4.3. Examples of output segmented images obtained from COCO 2020 challenge	23
Figure.4.4. Some examples of image frames included in LaSOT database	26
Figure.4.5. Some examples of video frames included in UAV123 dataset	27
Figure.4.6. Some examples of images included in ALOV++ database	28
Figure.4.7. Some examples of images included in VOT 2013 dataset	29
Figure.4.8. Some examples of images included in VOT 2014 dataset	29
Figure.4.9. Some examples of images included in VOT 2015 dataset	30
Figure.4.10. Some examples of images included in VOT -TIR 2015 dataset	31
Figure.4.11. Some examples of images included in VOT 2017 dataset	31
Figure.4.12. Some examples of images included in VOT 2018 dataset	32
Figure.4.13. Some examples of images included in VOT 2019 dataset	32
Figure.4.14. Some examples of images included in VOT-LT 2019 dataset	32
Figure.4.15. Some examples of images included in VOT-RGBD 2019 dataset	33
Figure.4.16. Some examples of images included in VOT-RGBTIR 2019 dataset	33
Figure.4.17. Some examples of images included in VOT-ST 2020 dataset	34
Figure.4.18. Some examples of images included in VOT-ST 2021 dataset	34
Figure. 5.1.1. Some frames extracted from video 1	36
Figure. 5.1.2. Some frames extracted from video 2	36
Figure. 5.1.3. Some frames extracted from video 3	37
Figure. 5.1.4. Some frames extracted from video 4	37
Figure. 5.1.5. Some frames extracted from video 5	38
Figure. 5.1.6. Some frames extracted from video 6	38
Figure. 5.1.7. Some frames extracted from video 7	39
Figure. 5.1.8. Some frames extracted from video 8	40
Figure. 5.1.9. Some frames extracted from video 9	40
Figure. 5.1.10. Some frames extracted from video 10	41



Figure. 5.1.11. Some frames extracted from video 11	41
Figure. 5.1.12. Some frames extracted from video 12	42
Figure. 5.1.13. Some frames extracted from video 13	42
Figure. 5.1.14. Some frames extracted from video 14	43
Figure. 5.1.15. Some frames extracted from video 15	43
Figure. 5.1.16. Some frames extracted from video 16	44
Figure. 5.1.17. Some frames extracted from video 17	44
Figure. 5.1.18. Some frames extracted from video 18	45
Figure. 5.1.19. Some frames extracted from video 19	45
Figure. 5.1.20. Some frames extracted from video 20	46
Figure. 5.1.21. Some frames extracted from video 21	47
Figure. 5.1.22. Some frames extracted from video 22	47
Figure. 5.1.23. Some frames extracted from video 23	48
Figure. 5.1.24. Some frames extracted from video 24	49
Figure. 5.1.25. Some frames extracted from video 25	49

## LIST OF TABLES

<b>Table No. Table caption</b>	<b>Page No.</b>
Table.4.1. Year-wise development achieved in PASCAL VOC datasets	15-16
Table 4.2. Comparative analysis of data present in training/validation/test set of PASCAL VOC 2012 and ILSVRC 2013 challenges	19
Table 4.3. Comparative statistics of validation set	20
Table 4.4. Comparative analysis of data present in training/validation/test set of PASCAL VOC 2012 and ILSVRC 2013 challenges	21
Table 4.5. Attributes considered for representing challenging cases in OTB databases	24-26

## **Abstract**

Visual object tracking is one of the most emerging area in the field of computer vision. In recent years, significant progress has been achieved in this research area. The recent research in this field is mainly focused on data acquisition for creating new benchmark databases suitable for evaluating performances of various tracking applications, designing several tracking methods, etc. The detection ability of any tracking method depends largely on the database on which it is trained. In order to perform efficient training of tracking methods and to perform proper evaluation of the their tracking capability, suitable benchmark databases containing videos or images frames comprising various attributes for e.g. occlusions, blurriness, etc. are required. The work conducted in this thesis is mainly focused on performing effective benchmark of tracking capabilities of multiple methods. The main contribution of this thesis lies in designing a novel database namely, VTrack: a visual object tracking benchmark database comprising of 25 videos containing thousands of frames possessing various attributes.

# Chapter 1

## Introduction

Visual object tracking being a significant research field has found its uses in various computer vision applications like vehicle detection and tracking, pedestrian detection and tracking, surveillance, self-driving cars etc. Although research on object tracking field has been carried out since several decades but in recent years, the practical significance and the challenging nature of this research problem have urged many researchers to conduct their research in this area which has boosted the research development in this field by many folds. Moreover, due to the emerging popularity of deep learning in the present era along with the availability of large amount of data and high computationally efficient systems, significant progress in this research field has been achieved.

The tracking problem is mainly concerned with the initial localization of target objects and then estimating their trajectory throughout the video sequence. This research problem is regarded as the most challenging one as it largely depends on the visual quality of video frames in which the tracking objects are located. The video frames in which the target objects are located can possess various attributes like occlusions, blurriness, varied illumination conditions, low resolution, large scale variation of target objects, etc.

An ideal tracking method should be able to perform effective tracking of objects irrespective of the large scale variations among their properties.

Tracking being a significant research field can be used for various applications based on which it can be divided into several types which are listed as follows:

- Video tracking
- Image tracking

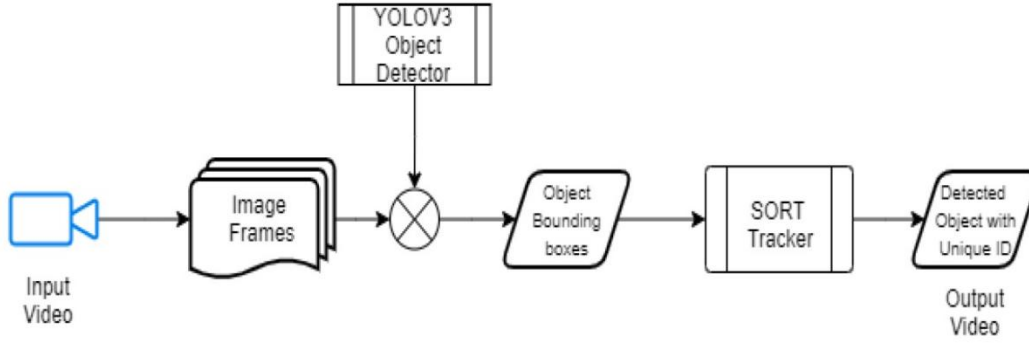
### ***Video tracking:***

This type of tracking is mainly focused on localizing target objects in each video frame and establishing a connection between their positions in each video frame.

### ***Image tracking:***

This type of tracking is mostly concerned with localization and tracking of target objects within an image which has mostly found its use in augmented reality.

Generalized representation for object detection and tracking is given in Figure.1.1.



**Figure.1.1.** Generalized representation for object detection and tracking [1]

Tracking being a significant research field, many researchers have designed various tracking methods to perform effective tracking of target objects. The efficiencies of those methods are largely dependent on the type of data which are used for training those algorithms. To perform effective benchmarking of the state-of-the-art tracking methods videos possessing diverse properties are required, so that the pros and cons of the existing methods can be highlighted.

The main contribution of this thesis lies in the creation of the novel database namely, VTrack: a visual object tracking benchmark database containing 25 videos comprising of thousands of video frames possessing diverse characteristics like varied degree of occlusions, illumination conditions, blurriness, etc. The videos included in this database also contains target objects moving with wide range of velocities, having varied dimensional properties, etc.

The primary objective behind designing this database is to effectively benchmark the performances of tracking methods in several challenging cases. For this purpose, 25 videos comprising of thousands of frames possessing various attributes are included in this database. The videos included in this databases are either captured from real world by the creator of this database or downloaded from YouTube. Detailed description of the created database is given in Chapter 5.

The rest of the thesis is organized as follows: Background Study & Literature Survey is given in Chapter 2, Existing challenges in visual tracking research area which are yet to be explored are summarized in Chapter 3 and how these unexplored challenges have served as the main motivation behind conducting this research work is discussed in the same Chapter. A comprehensive survey on existing visual tracking benchmark databases is conducted in Chapter 4 of this thesis and inspite of the presence of these databases what is the need to design more such databases are also highlighted in the same Chapter. VTrack database: a novel visual tracking benchmark database which is the main contribution of this thesis is introduced in Chapter 5. Examples of frames included in VTrack database, attributes considered while designing this database are also given in the same Chapter. Finally, the thesis work is concluded in Chapter 6 where the future scope of the conducted work is also highlighted.

## Chapter 2

### Background Study & Literature Survey

#### 2.1. Background Study

Researchers working in the visual object tracking area have addressed different challenges by developing new tracking methods. An ideal object tracking method should track target objects having diverse characteristics. The properties of target objects which an ideal tracking method should handle are as follows:

- **High velocity (HV):** Velocities of objects often play crucial role in determining the performance efficiencies of tracking methods. The challenges which can arise due to this attribute are listed as follows:

1. The high velocities of target objects often leads to the production of visually blurred video frames which often leads to erroneous tracking of objects.
2. Sometimes, the size of the target objects and the feature maps generated corresponding to those objects are so small that those objects cannot be accurately located in video frames.
3. High velocity objects are mostly present in outdoor scenes which have varied boundary structures and diverse lighting conditions. These additional factors often increase the complexity of the tracking problem.

An example of a frame containing a target object possessing high velocity included in VTrack database is given in Figure 2.1.



**Figure 2.1.** An example of a frame containing included in VTrack database with HV attribute

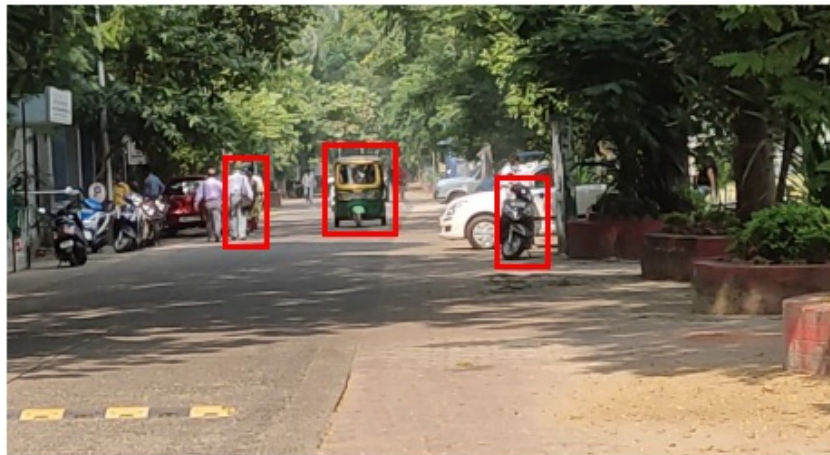
- **Non-linear movement (NLM):** This type of movement of tracking objects often increases the challenging nature of the tracking problem as the target objects possessing non-linear velocity may often go out of the video frame or their view can be blocked by some other objects, etc.

An example of a frame containing a target object possessing non-linear movement included in VTrack database is given in Figure 2.2.



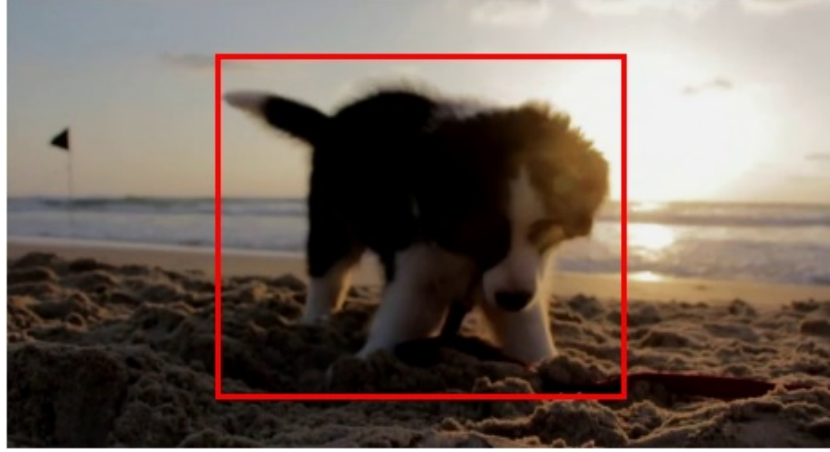
**Figure 2.2.** An example of a frame containing included in VTrack database with NLM attribute

- **Large scale (LS) variation of dimensions of target objects:** This attribute often has significant impact on the performance efficiency of tracking objects due to large scale variations among the size of feature maps used for detecting objects. An example of a frame with LS attribute included in VTrack database is given in Figure 2.3.



**Figure 2.3.** An example of a frame containing included in VTrack database with LS attribute

- **In-plane rotation (IPR):** When the target object rotates around an axis, it often gives erroneous tracking results due to the sudden change in bounding co-ordinates. An example of a frame with IPR attribute included in VTrack database is given in Figure 2.4.



**Figure 2.4.** An example of a frame containing included in VTrack database with IPR attribute

- **Low-resolution (LR):** Resolution of any image or video frame plays a significant role in determining its visual quality. Low resolution of any image or video frame often results in degradation of their visual quality. The feature maps generated for target objects located in low resolution frames may often be erroneous due to their poor visual quality. Hence, performing tracking of target objects using those erroneous feature maps sometimes produce ambiguous results. An example of a frame with IPR attribute included in VTrack database is given in Figure 2.5.



**Figure 2.5.** An example of a frame containing included in VTrack database with LR attribute

- **Background Clutter (BC):** Locating a target object from a cluttered environment often increases the complexity of the tracking problem as it becomes very challenging to segment out the target object from its' surrounding objects due to their overlapping nature. An example of a frame with BR attribute included in VTrack database is given in Figure 2.6.





**Figure 2.6.** An example of a frame containing included in VTrack database with BC attribute

- **Low light condition (LL):** The poor illumination condition often significantly impacts the tracking performance of any method as it makes the localization of target objects in an image frame or video frames and then plotting their trajectories considering their positions in sequential video frames quite difficult. An example of a frame with LL attribute included in VTrack database is given in Figure 2.7.



**Figure 2.7.** An example of a frame containing included in VTrack database with LL attribute

- **Inclement weather conditions (IWC):** Adverse weather conditions for e.g. rainy weather condition or hazy weather condition often degrades the visual quality of image frame or video frames due to scattering or attenuation of scene light by the rain droplets or fog, mist, dust particles present in the atmosphere respectively. These types of attenuation or scattering often degrades the contrast, edge information as well color information of the frames which in turn makes the target object localization and tracking tasks very challenging. An example of a frame with IWC attribute included in VTrack database is given in Figure 2.8.



**Figure 2.8.** An example of a frame containing included in VTrack database with IWC attribute

## ***2.2. Existing Literature***

Inspired by the practical significance of the object tracking problem and its' challenging nature, many authors have designed various methods in recent years for performing effective tracking of target objects. In this Section, brief description of methodologies of the existing tracking methods are given as follows:

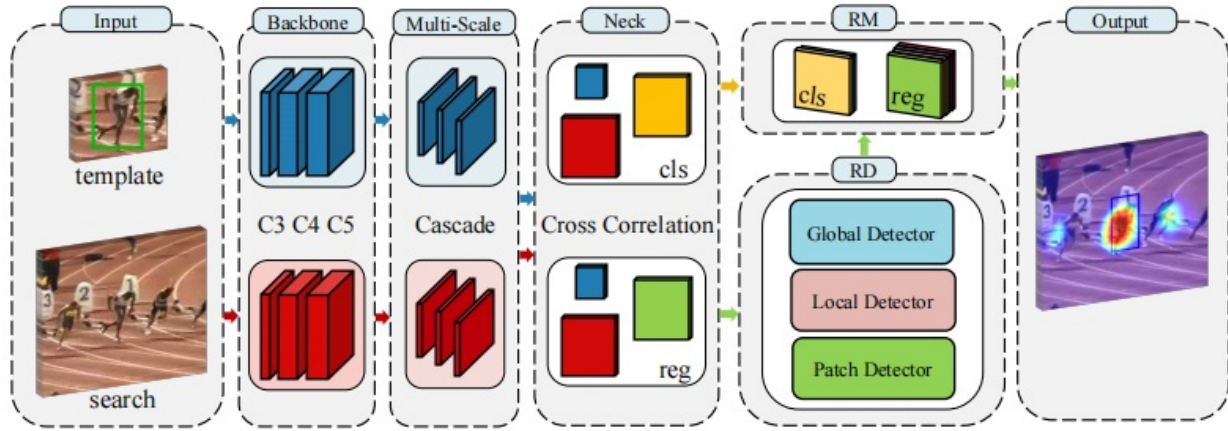
In the paper published by A. Bewley et. al. [2], the authors have designed a tracker by merging the Kalman filter and the Hungarian filter which has helped them to attain a high accuracy. The less complex nature of their designed tracker enables it to update at a rate of 260 Hz which makes the tracking process much faster.

A comprehensive survey [3] based on various aspects of object detection like generic object detection, object feature representation, detection frameworks (for e.g. Region Based (Two Stage) Frameworks, Unified (One Stage) Frameworks), context modeling, object proposal generation, training strategies, and evaluation metrics is done in this paper. The authors have mentioned the significance of the different datasets used for object detection like PASCAL VOC data sets [4], ImageNet [5], Common Objects in Context (COCO) [6], etc. in this survey. The comparison of performances of different object representation methods like local descriptions (SIFT), (HOG), (Bag of words), (shift vector) and deep learning architectures which are most commonly used for performing automated object detection and tracking are also included in this paper.

D. Hoiem et. al. [7] focuses on analyzing the impact of object characteristics in performing object detection in their work. The authors analyzed the performances of two types of detectors for e.g. Vedaldi et al. multiple kernel learning detector and different versions of the Felzenszwalb et al. detector in performing object detection and also highlighted effects of occlusion, aspect ratio, size, localization error, viewpoint, visibility of parts, etc. in performing object detection. Their study has found that localization error, sensitivity to size and confusion with similar objects are the most relevant forms of error in the detection task.

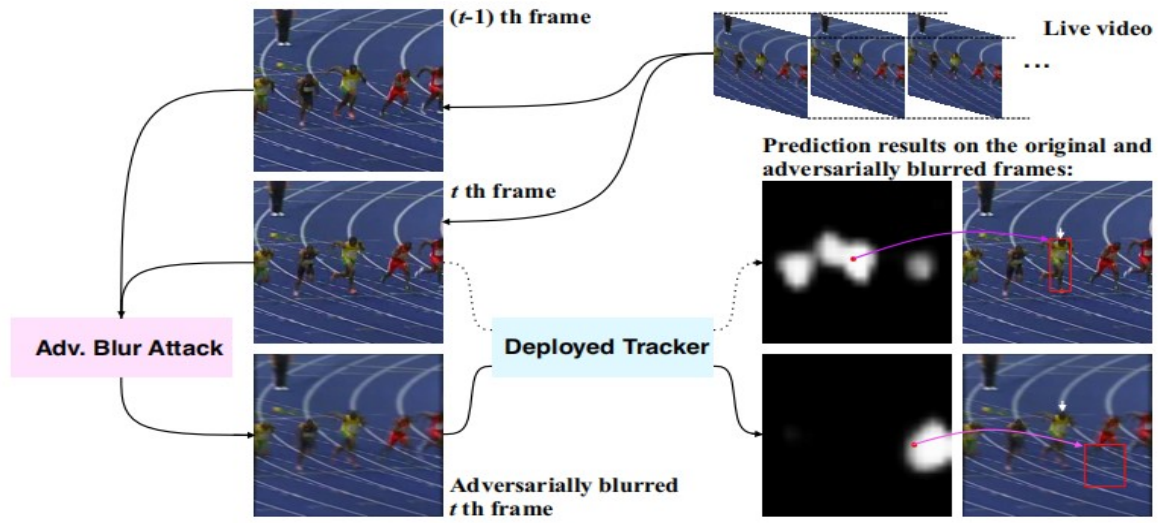
Siyuan cheng et. al. [8] have proposed a novel Siamese relation network based on the Siamese Network framework [9]. The designed network comprises of two efficient models, firstly Relation Detector (RD) which filters the distractions from the background in a meta-learning way. The second model is

Refinement Module (RM) which has the ability to effectively integrate the proposed RD and the Siamese framework to get the non-erroneous tracking result. This work also employs contrastive tracking strategies to perform feature matching between similar and dissimilar objects. The authors in this work have performed benchmarking of their proposed algorithm with respect to tracking benchmark databases for e.g. OTB100 [10], LaSOT [11], UAV123[12]. The block diagram of the designed Siamese relation network is given in Figure 2.9.



**Figure.2.9.**Network architecture of the Siamese Relation Network

In the paper published by Qing Guo et. al. [13], the authors have explored the effectiveness of visual object trackers against motion blur from a completely different aspect of adversarial blur attack (ABA). Authors have designed motion blur synthesis model based on motion information and light shifting process. The main aim of the work conducted in this paper is to transform input image frames to their motion-blurred counterparts. The authors have initially designed a generation principle of motion blur based synthesizing method taking into account the motion information and light accumulation process parameters. Using their designed motion blur synthesizing method, the authors have proposed optimization-based ABA (OP-ABA) method which performs iterative optimization of the adversarial objective function against tracking in accordance to the above-mentioned parameters. This iterative optimization process has quite a high time-complexity which makes the proposed OP-ABA unsuitable for use in real-life trackers. To mitigate this process, the authors have designed an one-step ABA (OS-ABA) by jointly training accumulation predictive network and adversarial network which facilitates the optimization of adversarial objective function in one step. The authors have also proved the efficiency of their designed method by testing the performance of their designed tracker using well-known datasets for e.g. OTB100 [10], UAV123[12], and LaSOT [11]. The block diagram of the designed OS-ABA based tracker is given in Figure.2.10.



**Figure.2.10.**An example of adversarial blur attack against a deployed visual tracker, SiamRPN++ [14]. In the case, two adjacent frames are fed to the designed adversarial blur attack to generate an adversarial blurred frame which misleads the tracker and produces an inaccurate response map.

Xin Chen et. al. [15] presented a novel attention based feature fusion network to effectively combine the template and search region features which mitigates the semantic information loss problem occurring in local linear template matching processes. The method proposed in this work includes a cross-feature augment module based on cross-attention and an ego-context augment module based on self-attention. The authors have designed a Siamese feature extraction framework based Transformer tracking method (namely TransT) in this work.

In [16], the authors have proposed a detection and tracking approach for performing vehicle and pedestrian recognition from sequential video frames. In their work, the authors have performed detection of objects using You Only Look Once (YOLO) detector and performed tracking of objects using Simple Online and Real-time Tracking (SORT) algorithm. Similar type of work is proposed in [1].

In the paper published by C. Bao et. al. [17], the authors developed a new  $l_1$  norm related minimization model based real time  $L_1$  tracker. The designed tracker also includes a  $l_2$  norm regularization term which is applied on coefficients associated with the templates. In this work, the authors have used quadratic convergence method to solve the  $l_1$  norm related minimization model to obtain a faster solution to the tracking problem which reduces the time-complexity of the problem.

A local fragment-based object tracking technique has been proposed in this research [18]. The method designed in this paper performs tracking of tracking objects using unique, distinctive and valid fragments which are selected based on spatial constraint, similarity and displacement criteria. Structural consistency and feature similarity phenomenon are used to update the object template. The authors have tested the performance of the designed tracking method using OTB 2013 [19] benchmarking database. In [19], the authors have created series of test image sequences annotated with various attributes to perform effective benchmarking of well-known tracking algorithms.

The authors in [20] have designed a unique target localization and representation method using a feature histogram based target representation technique. The feature maps generated using this technique is regularized by an isotropic kernel based spatial masking procedure. The target localization problem in

this work is portrayed as a local maxima estimation problem. The local maxima is estimated in this work using a gradient based optimization technique, Bhattacharyya Coefficient, a similarity measure parameter and mean shift procedure. The authors have proved the efficiency of their proposed method considering various challenging cases like partial occlusions, clutter, camera motion and target scale variations.

In [21], the authors have proposed a novel scale adaptive tracking approach to reduce the computational complexity of the exhaustive scale search method which is used to perform target size estimation in most of the methods. The proposed method performs target size estimation by learning the discriminative target features which are extracted using correlation filters. Instead of performing exhaustive scale search, the authors performed target scale estimation in their proposed method by detecting the changes in the appearance of target objects at various scales. The authors have proved the excellence of their method using popular visual tracking benchmark datasets like OTB [19] and the VOT2014 [22] datasets.

## Chapter 3

### Existing challenges & Thesis motivation

Despite significant research progress in the object tracking domain during the past decade, there are many challenges in this field which are yet to be explored. Those challenges have served as the major inspiration behind conducting this thesis work.

This Chapter comprises of two sections where in the first Section the unexplored challenges in this research area are highlighted followed by another section where, how these existing challenges have served as the motivation behind conducting the research work in this particular domain in this thesis is stated.

#### *3.1. Existing challenges in visual object tracking domain*

An ideal object tracking method should detect target objects possessing various properties with high accuracy. High quality detection of target objects means effectively localizing the target objects in different video frames and stitching their positions in each video frame so that their movement can be tracked across consecutive video frames. In real-world, tracking objects can have varied properties like:

- Large scale dimension variations among target objects
- Occlusion
- Vast velocity range of target objects
- Varied types of movements of target objects
- Visual degradation of video frames either due to poor visibility or low resolution

The challenges in this research domain can be broadly divided into two divisions namely,

- Accuracy Related Challenges
- Efficiency and Scalability Related Challenges

##### *3.1.1. Accuracy Related Challenges*

This type of challenge mostly arises for the reasons which are listed as follows:

- Intra-class variations among the objects' properties
- Large number of object categories
- Variations in the illumination properties of video frames
- Visual degradation of video frames due to scattering and attenuation of scene light by water droplets, dust, haze, fog, etc. present in the atmosphere during inclement weather conditions

Intra-class variations among target objects belonging to one object category can be defined as the variations in the color, morphology, dimension, texture, etc. of different instances of those objects.

Variations in the imaging conditions due to various environmental factors like poor visibility during adverse weather conditions (for e.g. during rainy weather condition, hazy weather condition, etc.), spatially variant illumination condition or poor illumination condition mostly during nighttime, occlusion significantly affects the visual quality of video frames which often decrease the accuracies of the tracking algorithms.

Apart from these issues, presence of motion blur, shading, clutter, etc. in video frames often have adverse effect on the performance efficiencies of the tracking algorithms.

In addition to intra-class variations among different instances of objects belonging to same category, inter-class variations among objects belonging to different categories also severely degrades the performances of tracking algorithms.

The researchers working in this domain in recent years mostly used well-known benchmarking datasets like ImageNet: A Large-Scale Hierarchical Image Database (ILSVRC) [23], PASCAL VOC [4], and MS COCO [6] to benchmark the performances of tracking methods. These databases comprise of target objects belonging to 200, 20 and 91 object categories respectively which can be considered almost negligible compared to the number of object categories present in real-world. Hence, the object trackers which are trained using video frames present in these databases comprising of objects belonging to limited number of object categories and are often unsuitable for performing effective tracking of innumerable target objects belonging to diverse object categories in real world.

### ***3.1.2. Efficiency and Scalability Related Challenges***

The prevalence of mobile devices and social media networks have increased the generation of data by many folds in the past few decades but the availability of limited storage space and computational capabilities of mobile devices make the object detection and tracking process much more challenging.

The efficiency challenges mostly arise while locating and recognizing objects belonging to large number of object categories. These challenges mostly occur due to large-scale variations among properties of target objects belonging to different categories (inter-class object properties variations) and target objects belonging to same class of objects (intra-class object properties variations).

Most of the tracking methods which are proposed till date still rely on manual annotation of target objects present in video frames for preparing the training dataset. As the number of object categories are increasing every day and as the properties of objects belonging to the same object class and objects belonging to different classes have large-scale variations in their properties, so there is a need to design an automated target object annotation method instead of performing manual annotation to conduct proper training of tracking methods using accurate bounding boxes of Ground Truths.

The large-scale variations in object properties and inefficiency of manual annotation method in performing accurate hand annotation of objects belonging to diverse categories also significantly affects the efficiencies of tracking algorithms.

## ***3.2. Motivation behind conducting this thesis work***

A huge number of research papers are published in this research domain due to immense significance of this research problem in real world. Apart from designing accurate object detection and tracking methods, another most important aspect of this research area is to create databases comprising of vast categories of objects possessing diverse intra-class variations and inter-class variations among their properties. The databases should comprise of several visually degraded or low-resolution video frames containing target objects possessing varied dimensional properties, velocities, etc., so that the trackers which will be trained using frames from these databases can efficiently handle such challenges .

Benchmarking databases play significant role in training, validating and testing of the object tracking methods as well as performing accurate evaluation of the efficiencies of the tracking methods.

Well-known databases which are mostly used in the existing works are PASCAL VOC datasets [4], , Common Objects in Context (COCO) [6], OTB100 [10] , LaSOT [11] , UAV123[12], OTB 2013 [19], VOT2014 [22] , ILSVRC[23], etc.

Brief descriptions of the existing databases are given in Chapter 4. The object categories included in these databases along with few examples of video frames included in the databases are also given in the same Chapter.

The novel database, VTrack which is created within the scope of thesis work is also introduced in Chapter 4 and how the frames included in the novel database differs from those belonging to the existing databases are highlighted in Chapter 5.



## Chapter 4

### Comprehensive survey on existing visual tracking benchmark databases

Brief descriptions of well-known databases which are used to perform effective benchmarking of object tracking methods proposed till date are given in this Chapter. The novel database, VTrack which is designed in this thesis to perform benchmarking of tracking methods satisfactorily even in challenging conditions is also introduced in this Chapter.

#### Existing databases:

1. **PASCAL VOC datasets [4]:** The main aim of PASCAL VOC project (2005-2012) is to design datasets comprising of frames containing objects belonging to various object categories. The datasets also include annotations which increases their popularity as a benchmarking database. The databases created under this project are used in performing several object detection tasks.

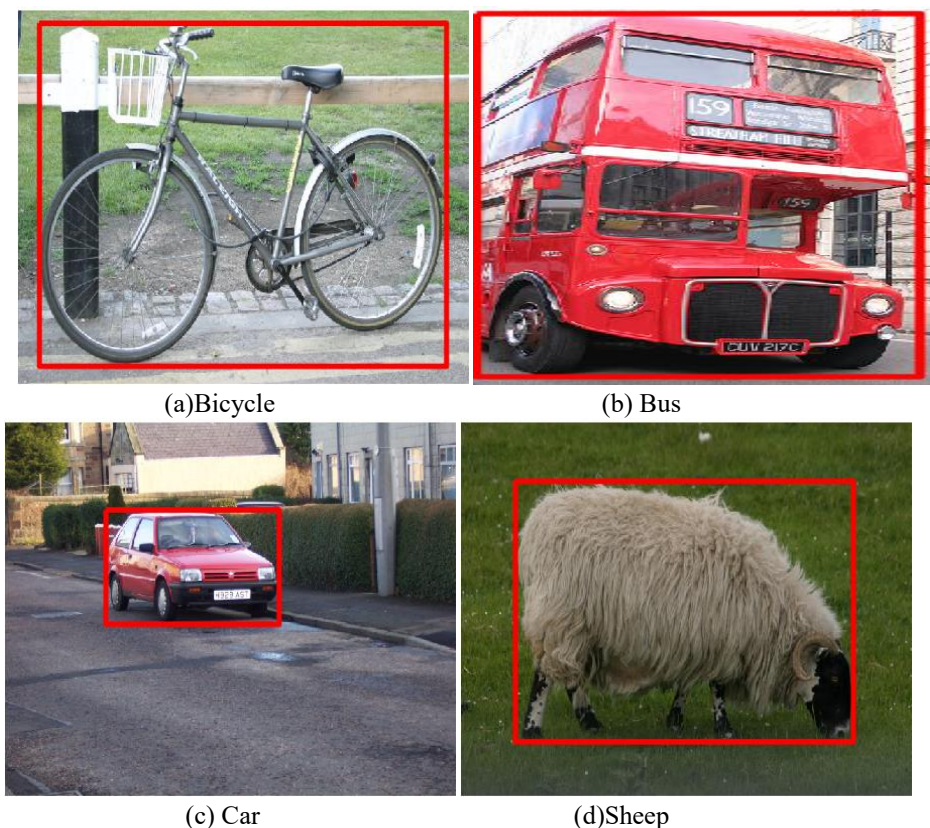
Brief details of VOC databases which are designed under PASCAL VOC project and how the development achieved in this project over each year during 2005-2012 are highlighted in Table 4.1.

**Table.4.1.** Year-wise development achieved in PASCAL VOC datasets

<i>Year</i>	<i>Object Categories</i>	<i>Developments achieved</i>	<i>Remarks</i>
2005	Number of classes: 4 Object categories: 1. Bicycles 2. Cars 3. Motorbikes 4. People Number of images present in training/validation/test dataset: 1578 Number of annotated objects: 2209	This dataset is used for performing several classification and detection tasks.	Frames included in this database are taken from several publicly available sources. These contain no challenging cases and hence, VOC 2005 is no longer used for performing benchmarking tracking and detection methods.
2006	Number of classes:10 Object categories: 1. Bicycles 2. Bus 3. Car 4. Cat 5. Cow 6. Dog 7. Horse 8. Motorbike 9. Person 10. Sheep Number of images present in training/validation/test dataset:2618 Number of annotated objects:4754	Images included in this database are taken from Flickr [24] and Microsoft Research Cambridge (MSRC) dataset [25].	Performing localization of objects present in MSRC dataset is often easier to perform compared to images taken from Flickr as those images are most concentrated on target objects.
2007	Number of classes: 20 Object categories: 1. Person 2. Bird	Segmentation taster and person layout taster are introduced as additional features in PASCAL VOC 2007 challenge. Truncation flags are added to	The number of object categories of VOC datasets is fixed to 20 from this year onwards. The final annotations for testing data is released in this year.

	3. Cat 4. Cow 5. Dog 6. Horse 7. Sheep 8. Aeroplane 9. Bicycle 10. Boat 11. Bus 12. Car 13. Motorbike 14. Train 15. Bottle 16. Chair 17. Dining Table 18. Potted plant 19. Sofa 20. tv/monitor Number of images present in training/validation/test dataset: 9,963 Number of annotated objects:24,640	annotations in this year's dataset. Average Precision is considered as the evaluation measure for estimating the performances of tracking methods in PASCAL VOC 2007 challenge whereas in PASCAL VOC 2005 and PASCAL VOC 2006, ROC-AUC is considered as the evaluation measure.	
2008	Number of classes: 20 Object categories: Same as VOC 2007 Number of images present in training/validation dataset:4340 Number of annotated objects:10,363	Occlusion flags are added in this year's dataset.	Test data annotation are made publicly available.
2009	Number of classes:20 Object categories:Same as VOC 2007 Number of images present in training/validation dataset:7054 Number of ROI annotated objects:17,218 Number of Segmentations: 3211	From this year onwards, previous year's images are augmented with new images which facilitates the increase in the number of images included in the database each year. This also allows comparison of the current year's test results with previous year's results.	Test data annotation are made publicly available. No difficult flags were given for additional images.
2010	Number of classes:20 Object categories:Same as VOC 2007 Number of images present in training/validation dataset:10,103 Number of annotated objects:23,374 Number of segmentations: 3211	Introduction of Action Classification taster. Early stage annotation is performed using Amazon Mechanical Turk.	The method of Average Precision computation is changed in PASCAL VOC 2010 compared to earlier years. Test data annotation are made publicly available.
2011	Number of classes: 20 Object categories:Same as VOC 2007 Number of images present in training/validation dataset:11,503 Number of ROI annotated objects:27,405 Number of segmentations: 5034	Action Classification taster is performed for objects belonging to all object categories.	Layout annotation are introduced for some images belonging to object category, 'people'.
2012	Number of classes:20 Object categories:Same as VOC 2007 Number of images present in training/validation dataset:27450 Number of ROI annotated objects: 6929	More segmentations are added to the database compared to the previous year's images.	No new layout annotation is performed.

Few examples of frames belonging to PASCAL VOC dataset are given in Figure 4.1.



**Figure.4.1.** Few examples of annotated image frames present in PASCAL VOC datasets

**2. ImageNet: A Large-Scale Hierarchical Image Database (ILSVRC)** [23]: The main aim of designing this challenge is to evaluate the performance accuracies of image classification and object recognition methods irrespective of the large scale intra-variations and inter-variations among the characteristics of objects belonging to the same category as well as different categories respectively.

The main motivation behind designing this challenge is to allow the researchers to perform comparative analyses of the performances of algorithms designed to perform various computer vision tasks for e.g. object classification and object detection by exploiting the huge labelling effort.

This challenge is organized each year during the time period of 2010-2017 and a workshop corresponding to the challenge is hosted along with some popular computer vision conference. The main aim behind organizing these workshops is to provide a global platform to the researchers working in the visual object tracking domain to present their designed methods and the results obtained from those methods.

ILSVRC 2012-2017 have been declared as the most popular challenge for performing image localization and classification tasks. The databases designed under this challenge are given in Kaggle [24].

Brief descriptions of the ILSVRC challenge organized each year during 2010-2017 are given as follows:

- **ILSVRC 2010** [26] - This challenge holds “taster competition” along with PASCAL VOC 2010 challenge.

The main aim of this challenge is to perform automatic annotation and retrieval of objects present in image frames. Hand-annotated images belonging to a subset of 1000 object categories included in the ImageNet database [5] are used for training the algorithms which competed in this challenge. Test images which are used in this version of the challenge do not contain any hand annotation, segmentation or labeling as the main objective of this challenge is to evaluate how accurately the algorithms participating in this challenge are able to provide correct labels corresponding to the target objects located in the test images. The primary aim of this challenge is to perform automatic labeling of target objects, localization of target objects is not within the scope of this challenge.

200,000 images collected from Flickr [24] and other publicly available databases are used as validation set and test set in this challenge. These objects fall within 1000 object categories of ImageNet database which are used in the training phase. Training set includes 1.2 million images from the selected 1000 object categories belonging to the ImageNet database.

The algorithms which participated in this challenge are evaluated based on two types of evaluation criteria:

- a. Hierarchical (Labelling is done based on the hierarchical structures of object categories)
- b. Non-hierarchical (Labelling is done considering all object categories to be equally significant)

The algorithms are ranked based on how accurately the automatic object labels provided by them for top most 5 object categories match with their original labels provided in their corresponding Ground Truth images. This procedure of object category detection also facilitates the algorithms to identify multiple objects located in video frames.

The algorithms are not penalized if they detect objects which are not labeled in the Ground Truth images.

1000 object categories which are considered for training the algorithms participating in ILSVRC 2010 challenge are given in [26].

- **ILSVRC 2011[27]**- This challenge is held jointly with PASCAL VOC 2011 challenge. In contrary to ILSVRC 2010 challenge, ILSVRC 2011 also focuses on object localization task in addition to image labeling task. Test set used in ILSVRC 2010 are modified in this challenge as some test images containing hand annotations are included in the test set used in ILSVRC 2011.

150,000 images collected from Flickr [24] and other publicly available sources are used as validation set and test set in this challenge. These objects fall within 1000 object categories of ImageNet database which are used in the training phase. Training set used in ILSVRC 2011 includes 1.2 million images from the selected 1000 object categories belonging to the ImageNet database similarly as done in ILSVRC 2010.

The algorithms which participated in this challenge are also ranked similarly as done in ILSVRC 2010 but in addition to the classification tasks, as this challenge also focuses on object localization so while evaluating the performance efficiencies of the participating algorithms, it is also considered that besides object classification how closely the bounding boxes of objects belonging to the top 5 object categories detected by these algorithms coincide with the bounding boxes of those objects present in their corresponding Ground Truths by performing error computation.

- **ILSVRC 2012 [28]**- This challenge also focuses on performing object detection in addition to object localization and classification. Test set used in ILSVRC 2011 are modified in this challenge as some new test images containing labeled objects are included in the test set.

Test, training and validation sets are created similarly as done in the previous year's challenges.

The algorithms which participated in this challenge are ranked similarly as done in the earlier challenges but in addition to the classification and object localization tasks, as this challenge also focuses on object detection so while evaluating the performance efficiencies of the participating algorithms it is also considered that how these algorithms perform fine-grained classification. Fine-grained classification is performed using dog images belonging to more than 100 categories. This type of classification is concerned with correctly labeling the category of the dog detected by the bounding box.

Mean Average precision is used as an evaluation parameter to rank the algorithms based on their performances for all object categories and Average Precision is used as an evaluation parameter to rank the algorithms based on their performances for one object category.

● **ILSVRC 2013 [29]**- This challenge includes three sub-challenges as stated below:-

- A PASCAL-style detection challenge which is performed using fully labeled data belonging to 200 object categories.
- An image classification challenge which is performed using frames containing objects belonging to 1000 different object categories.
- An image classification and object localization challenge which is performed using images containing objects belonging to 1000 different object categories.

Test data used in this year's challenge comprises of fully annotated and fully labeled images containing objects belonging to 200 object categories. The object categories are chosen taking into account some factors like level of image clutter, average number of object instances, object scale, etc. Comparative analyses of the number of images and objects present in the training, validation and test sets of databases used in ILSVRC 2013 challenge and PASCAL VOC 2012 challenge and statistics of data belonging to validation set are given in Table 4.2 and Table 4.3 respectively.

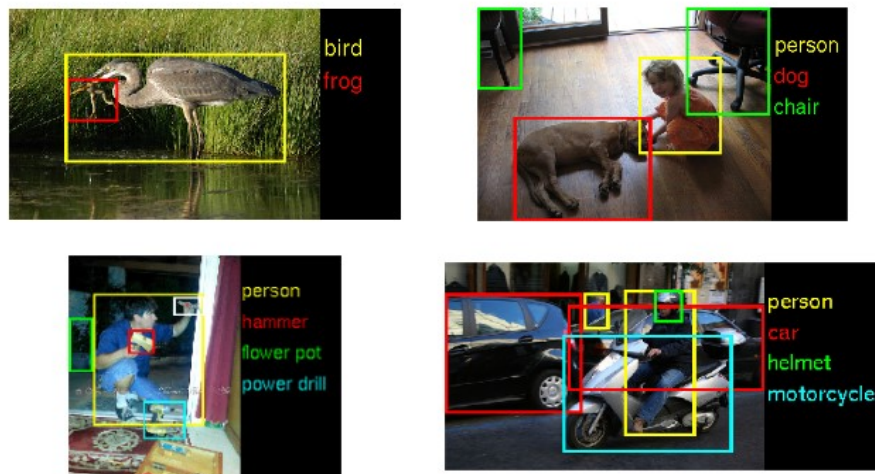
**Table 4.2.** Comparative analysis of data present in training/validation/test set of PASCAL VOC 2012 and ILSVRC 2013 challenges [29]

		<i>PASCAL VOC 12</i>	<i>ILSVRC 2013</i>
Number of object categories		20	200
Training set	Number of images	5717	395909
	Number of objects	13609	345854
Validation set	Number of images	5823	20121
	Number of objects	13841	55502
Test set	Number of images	10991	40152
	Number of objects	-	-

**Table 4.3.** Comparative statistics of validation set [29]

	<i>PASCAL VOC 12</i>	<i>ILSVRC 2013</i>
Average Image resolution	469x387 pixels	482x415 pixels
Average object classes (per image)	1.521	1.534
Average object instances (per image)	2.711	2.758
Average object scale (Ratio between bounding box and image area)	0.207	0.170

Some examples of frames used in ILSVRC 2013 challenge are given in the following Figure.4.2.



**Figure.4.2.** Examples of some frames belonging to ILSVRC 2013 database [29]

The algorithms which participated in this challenge are ranked similarly based on their object detection, classification and localization capabilities. The algorithms are penalized if they are unable to provide annotations for any object belonging to the 200 categories or they perform duplicate annotation i.e. they provide two bounding boxes for a single object.

For classification task and classification-localization tasks, the algorithms are ranked similarly as done in previous year.

- **ILSVRC 2014 [30]** - Like ILSVRC 2013[29], in ILSVRC 2014 challenge too, the object detection task is performed similarly as done in PASCAL VOC challenge using fully annotated and fully labeled objects belonging to 200 object categories. The training set used in ILSVRC 2014 challenge has been expanded significantly compared to the data used in ILSVRC 2013 challenge by including images collected from Flickr [24] database.

Comparative analysis of the number of images and objects present in the training, validation and test sets of databases used in ILSVRC 2014 challenge and PASCAL VOC 2012 challenge is given in Table 4.4.

**Table 4.4.** Comparative analysis of data present in training/validation/test set of PASCAL VOC 2012 and ILSVRC 2013 challenges [30]

		<i>PASCAL VOC 12</i>	<i>ILSVRC 2014</i>
Number of object categories		20	200
Training set	Number of images	5717	456567
	Number of objects	13609	478807
Validation set	Number of images	5823	20121
	Number of objects	13841	55502
Test set	Number of images	10991	40152
	Number of objects	-	-

The ranking of algorithms in ILSVRC 2014 challenge are done similarly as done in ILSVRC 2013.

- **ILSVRC 2015 [31]:** This challenge is focused on ranking methods based on their image/scene classification, object detection and localization properties. This challenge includes four competitions (two main competitions and two taster competitions) which are listed as follows:

a. Main competitions:

I. Detecting objects belonging to 200 fully connected object categories.

II. Performing localization of objects belonging to 1000 object categories.

b. Taster competitions:

I. Detecting objects from video frames belonging to 30 fully labeled categories.

II. Classification of scenes belonging to 401 categories.

The training set and validation set of database used in ILSVRC 2015 are similar to that of database used in ILSVRC 2014. The test set is partially modified. It contains fully labeled and fully annotated video frames containing objects belonging to 200 categories. The test images used in ILSVRC database also contains frames which do not belong to 200 categories.

The data used in this challenge for performing object localization task is similar to the data used in ILSVRC 2012 challenge.

Apart from main competitions, this challenge also includes taster completions.

The first taster competition focuses on detecting objects belonging to 30 object categories from videos. These 30 object categories are chosen from 200 object categories used in the main competition. These object categories are selected considering the same factors as mentioned in ILSVRC 2013 challenge.

The second taster competition is focused on identifying the scene categories of image frames included in this database. The data for this challenge is taken from Places2 dataset [32] which comprises of more than 10 million images belonging to more than 400 unique object categories. This data is divided into training set, validation set and test set comprising of 8.1 million images, 20 thousand images and 381 thousand images respectively. The distribution of training data set is non-uniform.

The algorithms participating in this challenge are evaluated similarly as done in previous year's challenges for object detection, classification and localization tasks. For scene classification task, the algorithms are ranked based on how closely the automatic scene labeling done by the algorithms match with the Ground Truth labels for top most 5 scene categories.

- **ILSVRC 2016 [33]:** This challenge is focused on performing object detection and localization as well as scene classification and scene parsing. The tasks performed in this challenge are summarized as follows:
  - a. Localizing objects belonging to 1000 categories.
  - b. Detecting objects belonging to 200 fully labeled categories.
  - c. Detecting objects belonging to 30 fully labeled categories from videos.
  - d. Classifying scenes belonging to 365 scene categories.
  - e. Parsing scenes belonging to 150 stuff categories and discrete object categories.

The data used for object localization and object detection tasks are similar to the data used in ILSVRC 2012 and ILSVRC 2014 challenges. The validation set and test set used for performing object detection from video frames in ILSVRC 2016 challenge are refreshed slightly. For scene classification task, the data is taken from Places2 database [32] as done in ILSVRC 2015. The training set, validation set and test set comprise of 8 million images, 36 thousand images and 328 thousand images respectively.

Scene data which are used for performing the scene parsing task are taken from ADE20K dataset [34] and [35]. The main objective of this challenge is to segment an image into different image regions and parse through those segmented regions. Each segmented image region considered in this task is associated with an image category. Image categories considered in this task are road, sky, etc. The training set, validation set used in this challenge comprise of 20 thousand images, 2 thousand images respectively. 150 distinct semantic categories including stuffs like grass, road, etc. and discrete objects like people, car, etc. are considered in this task.

The ranking of algorithms participating in this challenge based on object detection, object classification, scene classification are done similarly as in ILSVRC 2015 challenge. The ranking of algorithms based on scene parsing are done based on class-wise Intersection over Union (IoU) and pixel-wise accuracy as the final score. Pixel-wise accuracy is measured as the ratio of correctly predicted pixels to the total number of pixels present in a frame while class-wise IoU is measured as the IoU of pixels averaged for all the 150 semantic categories.

- **ILSVRC 2017 [36]:** This challenge is focused on performing object detection and localization from images and videos. The tasks which are performed under this challenge are listed as follows:
  - a. Localizing objects belonging to 1000 object categories.
  - b. Detecting objects belonging to 200 fully labeled object categories.
  - c. Detecting objects belonging to 30 fully labeled categories from video.



The data used for performing object localization task are similar to the data used in ILSVRC 2012 database. The training data and validation data used for performing object detection task are similar to that of ILSVRC 2014 while the test data used in this challenge is an updated version of the test data used in ILSVRC 2016 challenge.

The validation data and test data used to perform object detection from video is also an updated version of similar data used in ILSVRC 2016.

The ranking of the algorithms participating in this challenge is done similarly as done in previous years' challenges.

**3. Common Objects in Context (COCO) [6]:** COCO database comprises of data which are used for benchmarking various computer vision tasks for e.g. object detection, image segmentation and image captioning. COCO challenge is organized every year starting from 2014 onwards. The main objective of challenge organized each year is different from other years. Some examples of output segmented images obtained from COCO 2020 challenge are given in Figure.4.3.



**Figure.4.3.** Examples of output segmented images obtained from COCO 2020 challenge



**4. Object Tracking Benchmark (OTB) (OTB100 [10], OTB 2013 [19] and OTB 2015[37]):** These popular databases are widely used to evaluate the performances of visual tracking algorithms. OTB benchmark has two versions namely, OTB 2013 and OTB 2015 where OTB 2013 comprises of 51 sequences and OTB 2015 comprises of 100 sequences. Each frame of every sequence included in this database is annotated with bounding boxes and 11 challenging attributes.

The test sequences included in these databases are manually tagged with several attributes representing several challenging cases. A list of those attributes are given in the following Table.4.5.

**Table 4.5.** Attributes considered for representing challenging cases in OTB databases

<i>Attribute name</i>	<i>Attribute description</i>	<i>Example</i>
Illumination Variation (IV)	Illumination of target objects vary significantly	 <p>basketball</p>
Scale Variation (SV)	Ratio of bounding box of first frame and bounding box of current frame is out of range	 <p>biker</p>
Occlusion (OCC)	Targets are either fully or partially occluded	 <p>bird2</p>
Deformation (DEF)	Deformation of non-rigid objects	 <p>human3</p>

Motion blur (MB)	Target is blurred either due to motion of camera or motion of target	
Fast Motion (FM)	The motion of Ground Truth exceeds 20 pixels	
In-plan Rotation (IPR)	Target objects rotate in the image plane	
Out-of-plane Rotation (OPR)	Target objects rotate out of the image plane	
Out-of-view (OV)	Some regions of target objects go out of the frame	

Background Clutters (BC)	Close resemblance between the color and texture of target objects and their adjoining background	
Low Resolution (LR)	Total number of pixels located within the Ground Truth bounding box is less than 50 pixels.	

**5. Large-scale Single Object Tracking(LaSOT) [11]:** This benchmarking database is created to train the tracking methods which are designed using deep learning architectures as their backbone. As such trackers require huge amount of data to train properly, so to fulfil their requirements, this database is designed comprising of 1400 sequences having more than 3.52 million high-quality frames. Each frame is annotated manually by experts. 70 object categories are considered while designing this database where twenty sequences are included for each object category. The average length of videos included in this database is 83 seconds and the average number of frames present in the videos is 2512. Few examples of images included in LaSOT database are given in Figure. 4.4.



**Figure.4.4.** Some examples of image frames included in LaSOT database



6. **UAV123[12]**: Unlike all the databases whose details are discussed so far, UAV123 dataset is designed totally from aerial tracking viewpoint. It comprises of 123 video sequences captured from low-altitude Unmanned Aerial Vehicles (UAV). The inclusion of 110 thousand frames and 123 video sequences in this database makes it one of the largest visual tracking database ranking only after Amsterdam Library of Ordinary Videos for tracking (ALOV++).ALOV300++[38]. All the frames and video sequences included in this database are fully annotated. The entire dataset can be downloaded from [39]. Some examples of images included in this database are given in Figure.4.5.



**Figure.4.5.** Some examples of video frames included in UAV123 dataset [39]

7. **Amsterdam Library of Ordinary Videos for tracking (ALOV++) database [38]**: This database is probably the largest benchmarking database which is designed for evaluating the performances of tracking algorithms. The videos included in this database are mostly short videos whose duration ranges between 9.2 seconds to 35 seconds. In addition to these short videos, this database also contains 10 long videos whose duration ranges between 1 minute to 2 minutes.

In this database, the focus is kept mostly on short videos to maximize diversity. The videos included in this database cover various challenging cases which are listed as follows:

1. Illumination variation
2. Specularity
3. Transparency
4. Close resemblance between objects belonging to different object categories
5. Clutter
6. Zoom
7. Occlusion
8. Large-scale variation of shapes of tracking objects
9. Variations in motion patterns
10. Low contrast

To maintain compatibility with other databases, videos from existing databases are also included in this database. For example, 11 video sequences from existing databases which deal with occlusion and smoothness are included in this database. Moreover, another 11 popular video sequences which are used for performing tracking methods in most of the recently published papers are incorporated in this database. These videos deal with attributes like occlusion, shaking camera, transparency, etc.

In addition to these video sequences, this database also contains 65 videos which were earlier published in PETA workshop and 250 new video sequences resulting in the total count of videos included in this database to 315.

The videos included in this database are mostly collected from YouTube. The target objects included in this database are categorized to 64 different categories. Some examples of object categories are microscopic cells, plastic bag, octopus, etc.

The video sequences are also divided into thirteen different classes depending upon the level of difficulty of tracking the target objects present in those videos. For example different levels of difficulty are associated with tracking different objects like flock of birds, an octopus, a dancer, soldier in camouflage, etc.

The total 89364 number of frames included in ALOV++ database are fully annotated using rectangular bounding boxes. Some examples of video frames included in this database are given in Figure. 4.6.



**Figure.4.6.** Some examples of images included in ALOV++ database [38]

**8. Visual Object Tracking (VOT):** VOT challenges are primarily organized to provide a global platform to researchers working in the visual tracking research area to discuss about the recent advancement occurring in this field and also compare the performances of tracking algorithms.

VOT challenge is first organized in the year 2013 and from then onwards this challenge is organized each year.

a. VOT 2013 dataset [40]: This challenge is organized in the year 2013 at the conference venue of International Conference on Computer Vision (ICCV) in Sydney, Australia. This dataset comprises of 16 video sequences of short duration containing target objects in challenging backgrounds. The sequences which are included in this dataset are selected from a larger set of sequences using a visual features based clustering method. All the 16 sequences are annotated using axis-aligned bounding boxes provided by

VOT committee. This dataset has been archived now. Examples of data included in VOT 2013 dataset are given in Figure.4.7.



**Figure.4.7.** Some examples of images included in VOT 2013 dataset [40]

b. **VOT 2014 dataset [41]:** This dataset comprises of 25 video sequences of short duration containing target objects in challenging backgrounds out of which 8 sequences are included from VOT 2013 challenge. The sequences which are included in this dataset are selected from a larger set of sequences including ALOV++ dataset [38] similarly as done in VOT 2013 challenge. All the sequences are annotated using rotating bounding boxes instead of axis-aligned bounding boxes as done in VOT 2013 dataset. This dataset has been archived now. Examples of data included in VOT 2014 dataset are given in Figure.4.8.



**Figure.4.8.** Some examples of images included in VOT 2014 dataset [41]

c. **VOT 2015 dataset [42]:** This dataset comprises of 60 video sequences of short duration containing target objects in challenging backgrounds. The sequences which are included in this dataset are selected from a larger set of sequences which includes data taken from the publicly available sources listed below:

i. Computer Vision Online [43]

ii. Center for Research in Computer Vision, University of Central Florida, USA [44]

iii. Professor Bob Fisher's Image Database [45]

iv. Videezy [46]

v. Learning and Recognition in Vision Group, INRIA, France [47]

vi. Open Access Directory [48]

vii. Data Wrangling [49]

viii. NYU Center for Genomics and Systems Biology [50]

The sequence selection procedure used in this challenge is designed to select the sequences representing various challenges cases in the final dataset from the larger set of sequences which is composed of sequences taken from the publicly available sources listed above.

All the sequences included in this dataset are annotated by the VOT committee using rotating bounding boxes to provide accurate Ground Truths for performing effective comparative analysis of the tracking algorithms participating in this challenge.

The bounding boxes used in this dataset are placed over the target objects in such a way that almost 30% of the background pixels are included within the bounding boxes. This type of bounding box annotation can be explained using the following example:

If the target object is a person with extended arms, then in such case the bounding box is defined in such a way that the arms of the person is not included within the bounding box.

If the target objects were partially occluded or partially out of frame, then in such case the bounding boxes of those target objects were set in such a way that entire target objects are included within their corresponding bounding boxes. This type of bounding box annotation can be explained using the following example:

If the target object is a person with occluded legs, then in such case the bounding box is defined in such a way that the entire person with invisible legs is included within the box.

Some examples of images included in VOT 2015 challenge are given in Figure.4.9.



**Figure.4.9.** Some examples of images included in VOT 2015 dataset [42]

In addition to main challenge, VOT 2015 also holds a tracking sub-challenge VOT-TIR2015 which is focused on performing annotations of thermal images. The VOT-TIR2015 dataset comprises of a total 20 sequences, out of which 8 sequences are captured for exclusively for this dataset while the other 12 sequences are taken from the sources listed below:

- i. Termisk Systemteknik AB
- ii. EU FP7 project P5
- iii. Department of Electrical Engineering at Linköping University
- iv. Aalborg University
- v. School of Mechanical Engineering at University of Birmingham, ETH Zürich
- vi. Fraunhofer IOSB



The images included in this database are captured using thermal infrared sensor and are stored using a 16 bit format but as most of the trackers cannot handle 16 bit thermal images, so the sequences included in VOT-TIR 2015 dataset are truncated to 8 bit range.

The sequences included in this dataset are annotated using axis-aligned bounding boxes. Few examples of sequences included in VOT-TIR 2015 challenge are given in Figure.4.10.



**Figure.4.10.** Some examples of images included in VOT -TIR 2015 dataset [42]

d. VOT 2016 dataset [51]: Like VOT 2015 challenge, VOT 2016 challenge too co-hosts a tracking sub-challenge namely, VOT-TIR2016. The sequences included in VOT 2016 dataset are similar to the sequences included in VOT 2015 dataset but the Ground Truths of the sequences are more accurately defined in VOT 2016 compared to VOT 2015 challenge. VOT-TIR2016 dataset although contain new sequences.

e. VOT 2017 dataset [52]: VOT 2017 dataset is an updated version of VOT 2016 dataset while the thermal imaging sequences included in VOT-TIR 2017 dataset are similar to the sequences included in VOT-TIR 2016 dataset. Few examples of frames included in VOT 2017 dataset are given in Figure 4.11.



**Figure.4.11.** Some examples of images included in VOT 2017 dataset [52]

f. **VOT 2018 dataset [53]:** In addition to VOT 2018 challenge, VOT 2018 also co-hosts two sub-challenges which are listed as follows:

- i. Long-term sub-challenge.
- ii. Real-time sub-challenge.

Some examples of annotated frames included in VOT 2018 dataset are given in Figure 4.12.



**Figure.4.12.** Some examples of images included in VOT 2018 dataset [53]

g. **VOT 2019 datasets [54]:** Apart from the main challenge, VOT 2019 also holds two new challenges which are mentioned as follows:

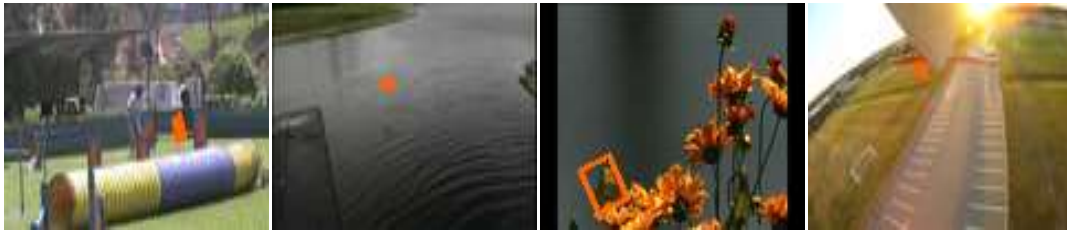
i. VOT-RGBT 2019 challenge

ii. VOT-RGBD 2019 challenge

VOT-RGBT 2019 challenge is organized to evaluate the performances of tracking methods which uses four channels (RGB+IR (Infrared)) channels as inputs.

VOT-RGBD 2019 challenge is organized to evaluate the performances of tracking methods which uses four channels (RGB+Depth) channels as inputs.

Some examples of annotated frames included in VOT 2019 dataset are given in Figure 4.13.



**Figure.4.13.** Some examples of images included in VOT 2019 dataset [54]

Some examples of annotated frames included in VOT-LT 2019 dataset are given in Figure 4.14.



**Figure.4.14.** Some examples of images included in VOT-LT 2019 dataset [54]

Some examples of annotated frames included in VOT-RGBD 2019 dataset are given in Figure 4.15.



**Figure.4.15.** Some examples of images included in VOT-RGBD 2019 dataset [54]

Some examples of annotated frames included in VOT-RGBTIR 2019 dataset are given in Figure 4.16.



**Figure.4.16.** Some examples of images included in VOT-RGBTIR 2019 dataset [54]

h. **VOT 2020 datasets [55]:** VOT 2020 includes five challenges whose details are mentioned as follows:

i. VOT-ST2020: This is a short term tracking challenge which focuses on estimating the variations in performance efficiencies of trackers when the frames given as inputs to the trackers contain various attributes like clutter, occlusion and appearance variation, etc. Targets used in this challenge are annotated using segmentation masks.

ii. VOT-RT2020: This is a short term real-time challenge which focuses on evaluating the performance efficiencies of trackers in real-time cases. Targets used in this challenge are annotated using segmentation masks.

iii. VOT-LT2020: This is a long term challenge which focuses on evaluating the performance efficiencies of trackers when target objects disappear.

iv. VOT-RGBT2020

v. VOT-RGBD2020

The objectives of VOT-RGBT2020 and VOT-RGBD2020 are similar to the objectives of VOT-RGBT2019 and VOT-RGBD2019.

Some examples of annotated frames included in VOT-ST 2020 dataset are given in Figure 4.17.



**Figure.4.17.** Some examples of images included in VOT-ST 2020 dataset [55]

VOT-LT 2020, VOT-RGBT 2020 and VOT-RGBD 2020 datasets are similar to the corresponding datasets used in the year 2019 but the frames included in VOT-RGBT 2020 are annotated with extra points to support short-term experimental methodology.

i.VOT 2021 [56]: VOT 2021 co-hosts challenges namely, VOT-ST2021,VOT-RT2021, VOT-LT2021 and VOT-RGBD2020. The objectives of these challenges are similar to the corresponding challenges organized in earlier years.

Some examples of annotated frames included in VOT-ST 2021 dataset are given in Figure 4.18.



**Figure.4.18.** Some examples of images included in VOT-ST 2021 dataset [56]

VOT-LT2021 dataset and VOT-RGBD2021 dataset are same as VOT-LT2020 dataset and VOT-RGBD2020 dataset.

Although the frames included in these databases cover wide range of challenging cases but still there is a need to design databases which shall include frames from different scenarios for e.g. in order to make a self-driving car work satisfactorily in all situations, it should be trained using databases containing road images from urban areas as well as rural areas, images of cars, trackers, trollers, rickshaws etc. as it is not known beforehand in which country and in types of areas, these cars will be used.

So keeping these facts in mind, VTrack database is designed in this thesis. VTrack database comprises of videos from YouTube as well as videos which are recorded from real field. Detailed description of VTrack database, examples of frames included in this database, attributes considered, etc. are given in Chapter 5.

## Chapter 5

### VTrack database

VTrack database is designed in this thesis containing frames of videos collected from publicly available sources such as YouTube as well as videos collected from real field.

The videos included in this database cover a wide range of challenging cases. A list of attributes considered while designing this database are listed as follows:

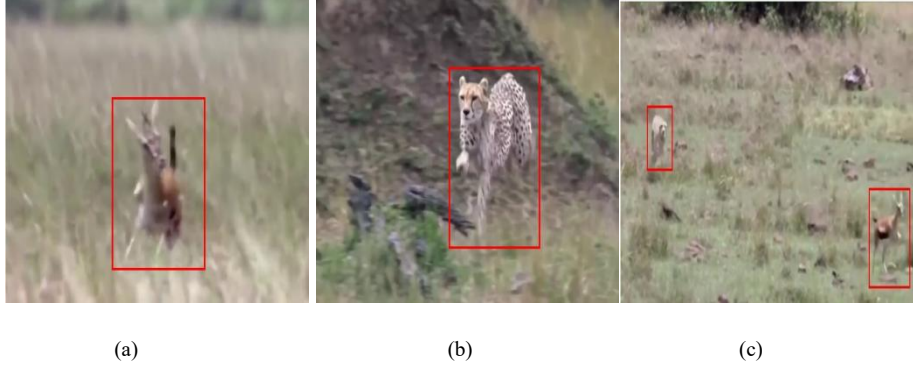
- Non-linear velocity (NLV)
- Scale-variation (SV)
- In-plane rotation (IPR)
- Low Resolution (LR)
- Background Cluster (BC)
- Illumination variation (IV)
- Occluded object (OO)
- Out-of-view (OFV)
- Linear velocity (LV)
- Videos shot far away from the target object (FT)
- Outer-plane rotation (OPR)
- Object with rotational movement (ORM)
- Frames captured during inclement weather condition for e.g. hazy weather condition (IWC)
- Blurred target objects (BrT)
- High velocity (HV)
- Absence of target objects in the initial video frames (InA)
- Non-Linear motion (NLM)

Some examples of frames included in the VTrack database are given in the following Section and the list of attributes associated with those frames are also mentioned.



### 5.1. Examples of frames included in VTrack database

- Examples of frames belonging to video 1



**Figure. 5.1.1.** Some frames extracted from video 1

*Attributes associated with video 1 are listed as follows:*

- a. NLV
- b. SV
- c. IPR
- d. LR
- e. BC
- f. IV
- g. OO
- h. OFV

- Examples of frames belonging to video 2

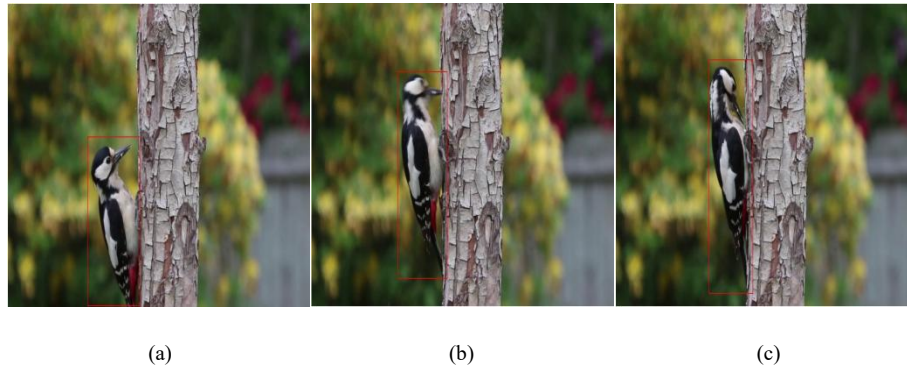


**Figure. 5.1.2.** Some frames extracted from video 2

*Attributes associated with video 2 are listed as follows:*

- a. LV
- b. SV
- c. OO

- Examples of frames belonging to video 3

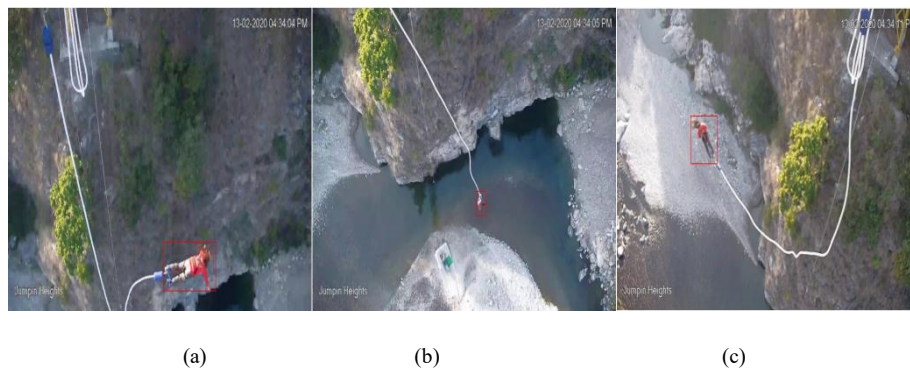


**Figure. 5.1.3.** Some frames extracted from video 3

*Attributes associated with video 3 are listed as follows:*

- a. NLM

- Examples of frames belonging to video 4



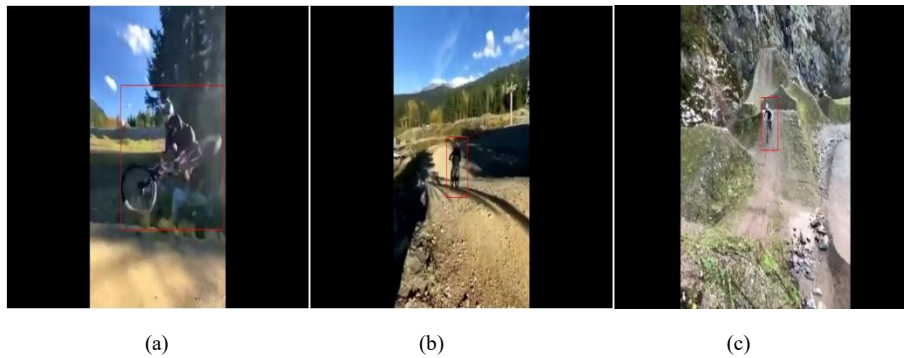
**Figure. 5.1.4.** Some frames extracted from video 4

*Attributes associated with video 4 are listed as follows:*

- a. NLV
- b. SV
- c. OPR

- d. LR
- e. ORM
- f. BrT

- Examples of frames belonging to video 5

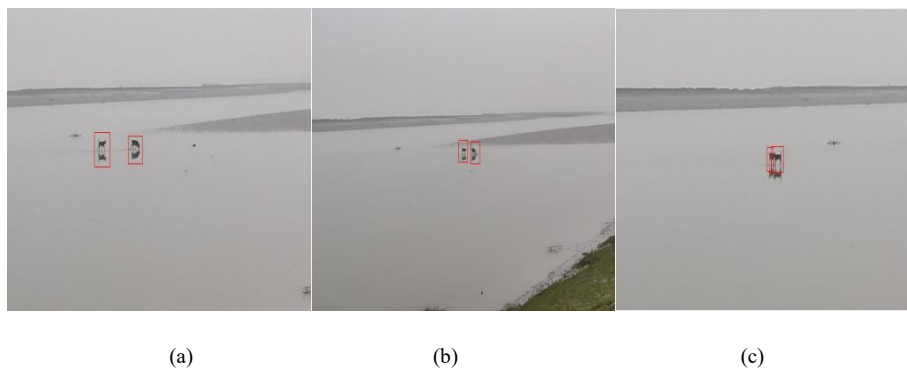


**Figure. 5.1.5.** Some frames extracted from video 5

*Attributes associated with video 5 are listed as follows:*

- a. HV
- b. NLV
- c. SV
- d. LR
- e. LL
- f. IV

- Examples of frames belonging to video 6



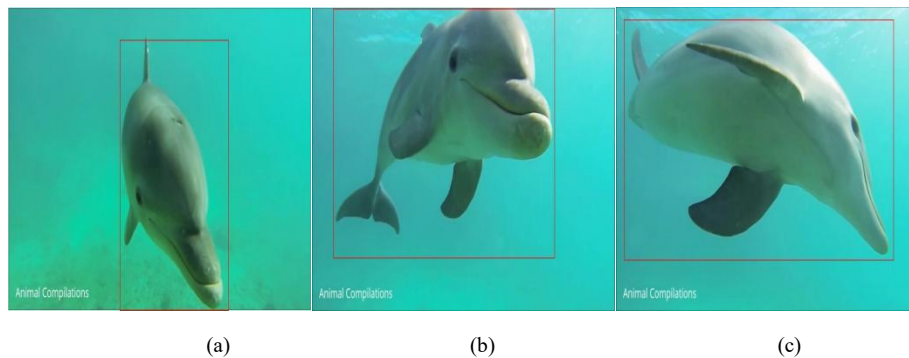
**Figure. 5.1.6.** Some frames extracted from video 6



*Attributes associated with video 6 are listed as follows:*

- a. NLV
- b. SV
- c. IPR
- d. LR
- e. LL
- f. ORM
- g. OO
- h. BrT

- Examples of frames belonging to video 7

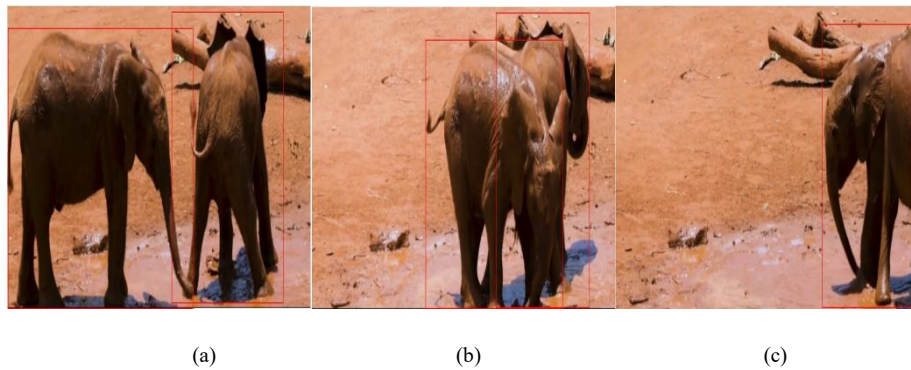


**Figure. 5.1.7.** Some frames extracted from video 7

*Attributes associated with video 7 are listed as follows:*

- a. NLV
- b. ORM
- c. SV
- d. IPR

- Examples of frames belonging to video 8

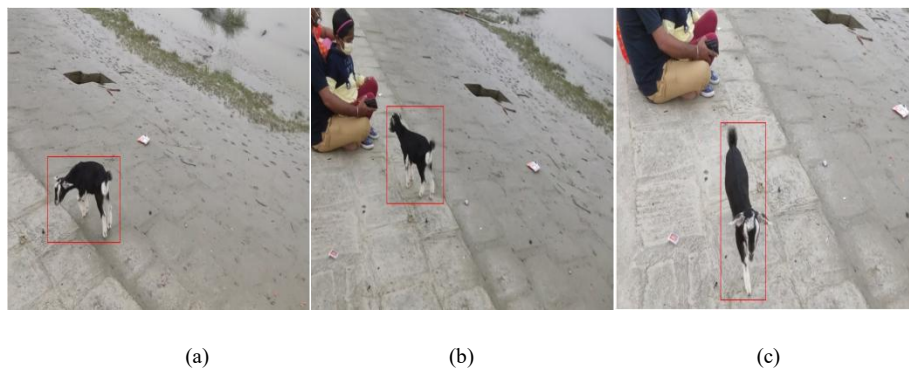


**Figure. 5.1.8.** Some frames extracted from video 8

*Attributes associated with video 8 are listed as follows:*

- NLV
- IPR
- OO
- BC
- ORM

- Examples of frames belonging to video 9

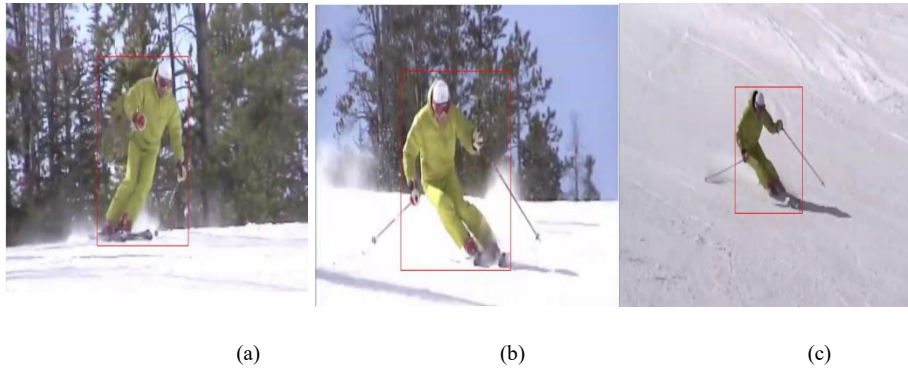


**Figure. 5.1.9.** Some frames extracted from video 9

*Attributes associated with video 9 are listed as follows:*

- NLV
- SV
- ORM
- IPR
- OO

- Examples of frames belonging to video 10

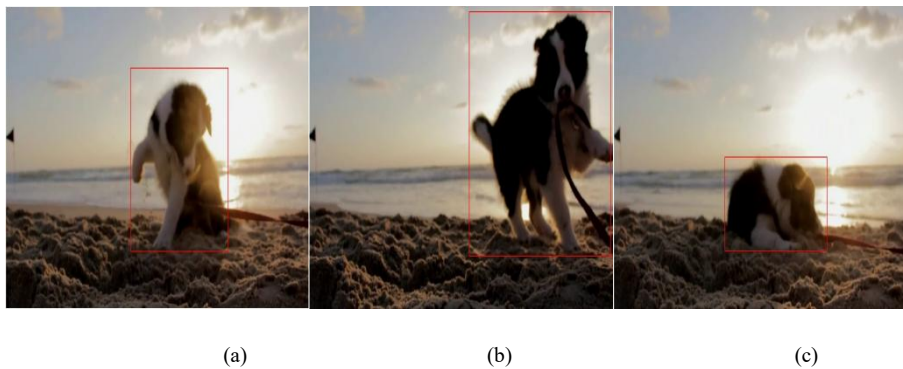


**Figure. 5.1.10.** Some frames extracted from video 10

*Attributes associated with video 10 are listed as follows:*

- NLV
- SV
- OO
- OFV

- Examples of frames belonging to video 11

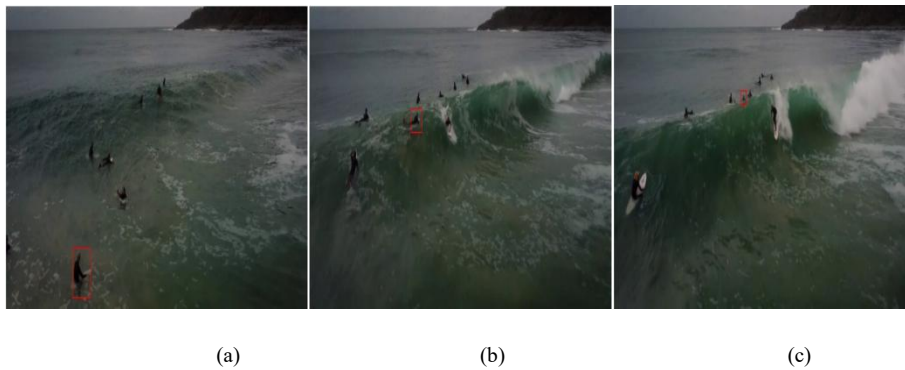


**Figure. 5.1.11.** Some frames extracted from video 11

*Attributes associated with video 11 are listed as follows:*

- NLV
- SV
- ORM
- IPR
- IV

- Examples of frames belonging to video 12

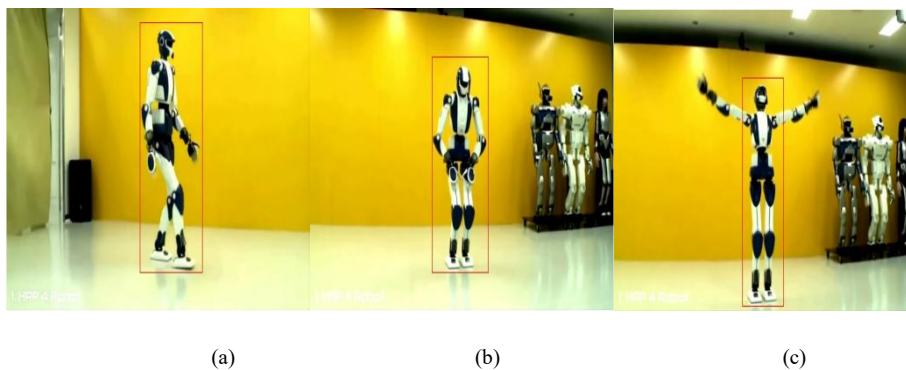


**Figure. 5.1.12.** Some frames extracted from video 12

*Attributes associated with video 12 are listed as follows:*

- HV
- NLV
- SV
- IPR
- BC
- LR
- IV
- OO
- ORM
- BrT

- Examples of frames belonging to video 13



**Figure. 5.1.13.** Some frames extracted from video 13

*Attributes associated with video 13 are listed as follows:*

- a. LV
- b. SV
- c. IPR

- Examples of frames belonging to video 14



(a) (b) (c)

**Figure. 5.1.14.** Some frames extracted from video 14

*Attributes associated with video 14 are listed as follows:*

- a. HV
- b. OO
- c. SV

- Examples of frames belonging to video 15



(a) (b) (c)

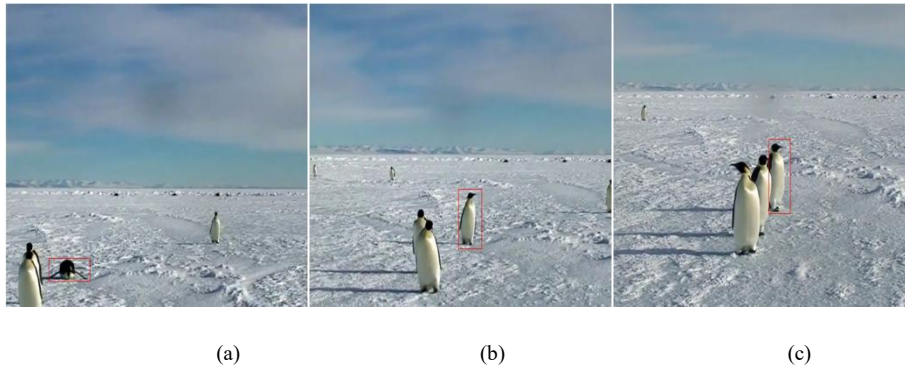
**Figure. 5.1.15.** Some frames extracted from video 15

*Attributes associated with video 15 are listed as follows:*

- a. NLV
- b. OPR
- c. BC



- Examples of frames belonging to video 16

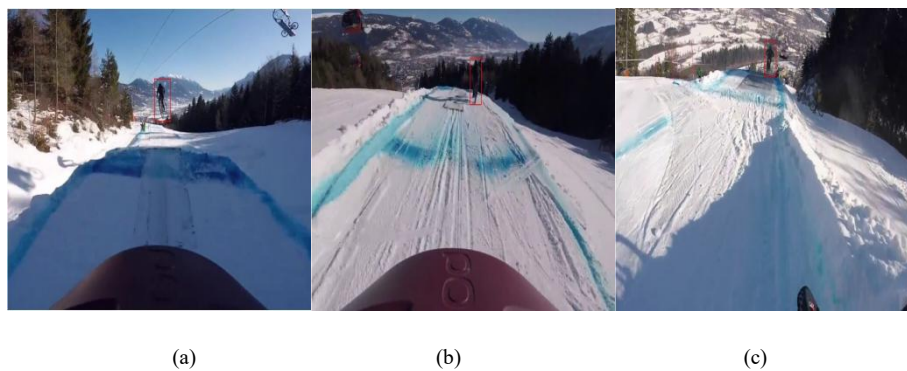


**Figure. 5.1.16.** Some frames extracted from video 16

*Attributes associated with video 16 are listed as follows:*

- HV
- NLV
- SV
- BC
- IV
- OO

- Examples of frames belonging to video 17



**Figure. 5.1.17.** Some frames extracted from video 17

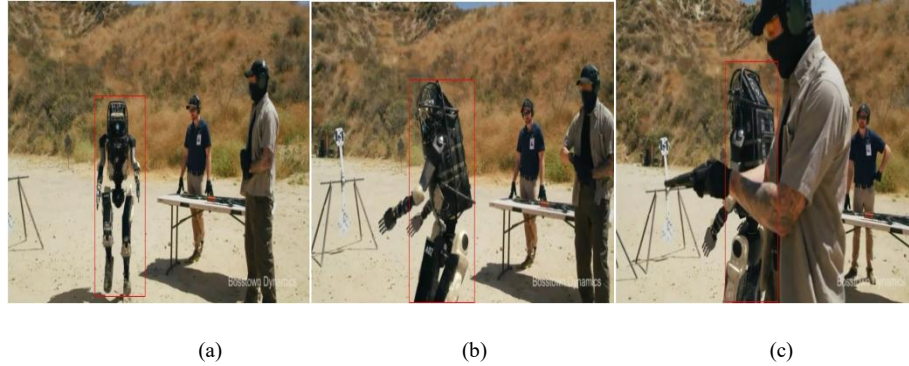
*Attributes associated with video 17 are listed as follows:*

- HV
- NLV
- SV
- BC

e. IV

f. OO

- Examples of frames belonging to video 18



**Figure. 5.1.18.** Some frames extracted from video 18

*Attributes associated with video 18 are listed as follows:*

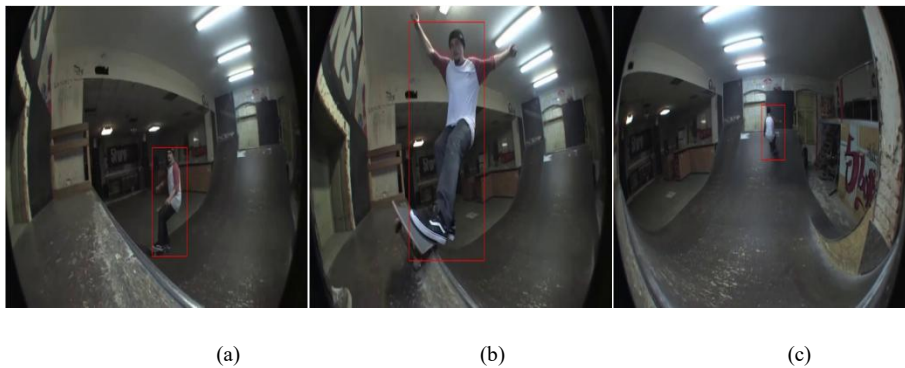
a. NLV

b. SV

c. IPR

d. OO

- Examples of frames belonging to video 19



**Figure. 5.1.19.** Some frames extracted from video 19

*Attributes associated with video 19 are listed as follows:*

a. HV

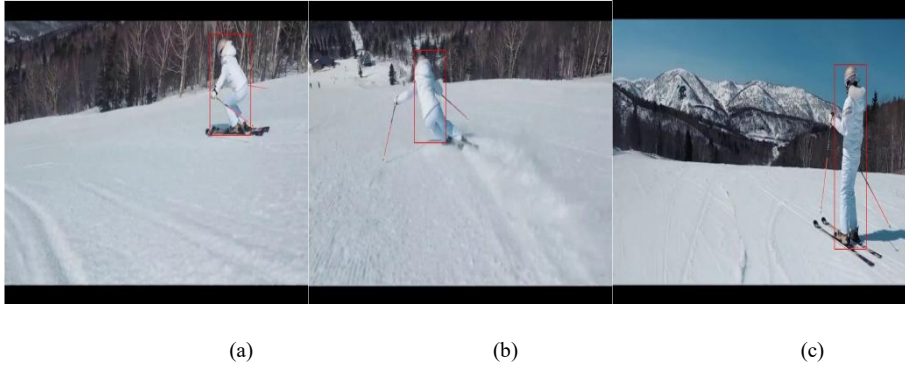
b. NLV

c. SV

d. OPR

e. ORM

- Examples of frames belonging to video 20



**Figure. 5.1.20.** Some frames extracted from video 20

*Attributes associated with video 20 are listed as follows:*

a. HV

b. NLV

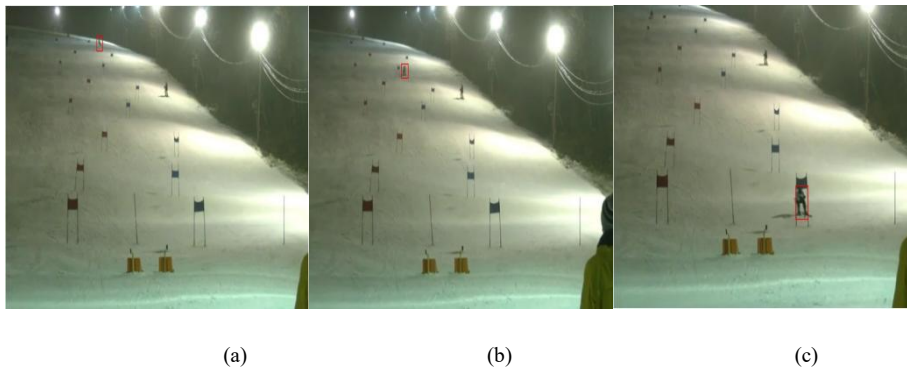
c. SV

d. BC

e. OFV



- Examples of frames belonging to video 21

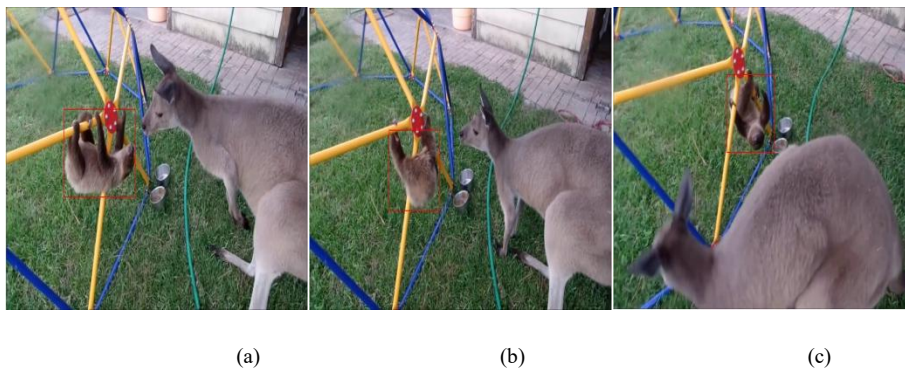


**Figure. 5.1.21.** Some frames extracted from video 21

*Attributes associated with video 21 are listed as follows:*

- a. HV
- b. NLV
- c. SV
- d. LR
- e. BC
- f. LL
- g. IV
- h. BrT

- Examples of frames belonging to video 22

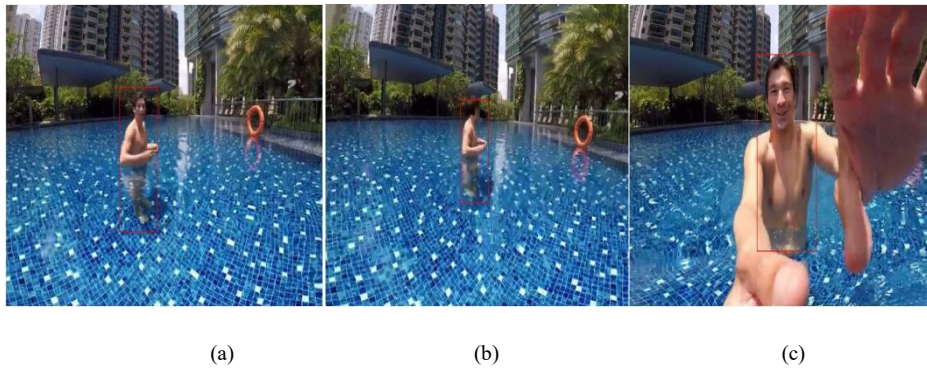


**Figure. 5.1.22.** Some frames extracted from video 22

*Attributes associated with video 22 are listed as follows:*

- a. NLV
- b. SV
- c. IPR
- d. OPR
- e. ORM

- Examples of frames belonging to video 23



**Figure. 5.1.23.** Some frames extracted from video 23

*Attributes associated with video 23 are listed as follows:*

- a. NLV
- b. SV
- c. IPR
- d. OPR
- e. ORM

- Examples of frames belonging to video 24



(a)

(b)

(c)

**Figure. 5.1.24.** Some frames extracted from video 24

*Attributes associated with video 24 are listed as follows:*

- HV
- NLV
- SV
- LR
- LL
- IV
- OO
- BrT

- Examples of frames belonging to video 25



(a)

(b)

(c)

**Figure. 5.1.25.** Some frames extracted from video 25

*Attributes associated with video 25 are listed as follows:*

- a. SV
- b. IPR

The videos included in the VTrack database are collected from diverse backgrounds like wildlife videos, videos from different sports activities, etc.

Videos of different sports activities like skiing, bungee jumping, cycling, skateboarding, etc. exhibiting varied types of movements, velocities, occlusions, etc. are selected in this thesis work for inclusion in the VTrack database because of the vast diversity in their characteristics. Videos of wildlife are also collected for similar reasons. In wildlife videos activities of several animals exhibiting wide range of behaviors from movement point of view and also depicting different degree of camouflage are recorded. These factors often make the tracking task even more challenging.

## Chapter 6

### Conclusion & Future Scope

The work conducted in this thesis focuses on designing a novel visual object tracking database namely, VTrack for benchmarking the performances of tracking methods. Inspired by the real-life significance and emerging popularity of the visual object tracking research area, this topic is chosen here to carry out the thesis work .

Databases containing frames taken from videos associated with various attributes and containing target objects belonging to many object categories are required to properly train object tracking methods in order to enable them to perform efficiently in real-world applications.

Although a large number of visual object tracking databases namely, PASCAL VOC datasets [4], ILSVRC [23], OTB100 [10], OTB 2013 [19], OTB 2015[37],COCO [6], etc. (whose details are given in Chapter 4) exist till date, still there is a need to create more databases as the number of challenges associated with this research area are increasing with each passing day.

Visual object tracking is a popular research avenue of the computer vision field. It has found its use in wide range of applications due to availability of large amount of data in current era, progress in the architectural development of deep neural networks which is one of the main tools used in the computer vision field, accessibility of high computationally efficient graphics card,etc.

In order to enable tracking methods work satisfactorily in real-world, they should be trained using fully labeled and fully annotated images of target objects. As the number of target objects encountered in real-world is huge compared to the total number of object categories included in the existing databases, there is a need to create more databases to improve the performances of the designed trackers in real world.

For this reason, VTrack, a novel visual object tracking benchmark database comprising of 25 videos associated with various attributes are designed in this thesis. The videos included in the designed database are either taken from YouTube or captured from real-world. The videos selected for preparing the database comprise of diverse scenes for e.g. the videos are mostly wildlife videos, videos of different sports activities, videos of road scenes, etc. VTrack database comprises of videos containing target objects exhibiting varied types of movements, velocities, occlusions, illuminations, rotational movements, etc. to perform efficient training of tracking methods and thus enhance their performances.

The future scope of this work can be focused on increasing the number of object categories, creation of more accurate Ground Truth bounding boxes, benchmarking the performances of tracking methods designed in recent years using this database, etc.

## References

- [1] Akansha Bathija and Grishma Sharma, "Visual Object Detection and Tracking using YOLO and SORT", International Journal of Engineering Research & Technology (IJERT), pp. 2278-0181, vol.8 (11), 2019.
- [2] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking." IEEE Int. Conf. on Image Processing, Phoenix, AZ, 2016, pp. 3464-3468.
- [3] L. Liu, W. Ouyang, X. Wang, P. Fieguth, "Deep Learning for Generic Object Detection: A Survey," International Journal of Computer Vision, vol. 128, 2020.
- [4] <http://host.robots.ox.ac.uk/pascal/VOC/>, Retrieved 16.7.2022.
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database." IEEE Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 2009, pp. 248-255.
- [6] <https://cocodataset.org/#home>, Retrieved 16.7.2022.
- [7] D. Hoiem, Y. Chodpathumwan and Q. Dai, "Diagnosing Error in Object Detectors", European Conference on Computer Vision (ECCV), Italy, 2012, pp. 340-353.
- [8] S. Cheng, B. Zhong, G.Li, X.Liu, et.al. , "Learning to Filter: Siamese Relation Network for Robust Tracking", arXiv:2104.00829v1, 2021.
- [9] G. Koch, R. Zemel and R. Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition", ICML deep learning workshop, vol. 2. 2015.
- [10] Y. Wu, J. Lim, M-H. Yang, "Object Tracking Benchmark", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1834 – 1848, vol. 37 (9), 2015.
- [11] H. Fan, L. Lin, F. Yang, P. Chu, et.al."LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking", IEEE Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, 2019.
- [12] M. Mueller, N. Smith and B. Ghanem, "A Benchmark and Simulator for UAV Tracking", European Conference on Computer Vision (ECCV), Amsterdam, 2016, pp. 340-353.
- [13] Q. Guo, Z. Cheng, F. Juefei-Xu, L.Ma, et.al., "Learning to Adversarially Blur Visual ObjectTracking", arXiv:2107.12085v4, 2021.
- [14] H. Fan, L. Lin, F. Yang, P. Chu, et.al., "Lasot: A high-quality benchmark for large-scale single object tracking", IEEE Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, pp. 5369–5378, 2019.
- [15] X.Chen, B.Yan, J.Zhu, D.Wang, et.al., "Transformer Tracking", IEEE Computer Vision and Pattern Recognition (CVPR), Nashville, 2021.
- [16] H.Thakkar, N.Tambe, S.Thamke and V.K. Gaidhane, "Object Tracking by Detection using YOLO and SORT", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol.6 (2), 2020.
- [17] C.Bao, Y.Wu, H.Ling and H.Jui, "Real Time Robust L1 Tracker Using Accelerated Proximal Gradient Approach", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 2012.
- [18] J.Zheng, B.Li, M.Xin and G.Luo, "Structured fragment-based object tracking using discrimination, uniqueness, and validity selection", Multimedia Systems, pp. 487-511, vol. 25, 2019.

- [19] Y. Wu, J. Lim and M-H. Yang, “Online Object Tracking: A Benchmark”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 2013.
- [20] D. Comaniciu, V. Ramesh and P.Meer, “Kernel-Based Object Tracking”, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.564-577, 2003.
- [21] M. Danelljan,G. Hager, F. S. Khan and M. Felsberg, “Discriminative Scale Space Tracking”,IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-14, vol. 39, 2017.
- [22] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, et.al., “VOT2014 Benchmark”, European Conference on Computer Vision (ECCV), Zurich, 2014.
- [23] O. Russakovsky, J. Deng, H. Su, J. Krause, et.al., “ImageNet Large Scale Visual Recognition Challenge”, International Journal of Computer Vision, 2015.
- [24] <https://www.kaggle.com/c/imagenet-object-localization-challenge/overview/description>, Retrieved 16.07.2022.
- [25]<https://www.microsoft.com/en-us/download/details.aspx?id=52644>,Retrieved 16.07.2022.
- [26] <https://www.image-net.org/challenges/LSVRC/2010/browse-synsets.php>, Retrieved 16.07.2022.
- [27] <https://www.image-net.org/challenges/LSVRC/2011/index.php> ,Retrieved 16.07.2022.
- [28] <https://www.image-net.org/challenges/LSVRC/2012/index.php> ,Retrieved 16.07.2022.
- [29] <https://www.image-net.org/challenges/LSVRC/2013/index.php> ,Retrieved 16.07.2022.
- [30] <https://www.image-net.org/challenges/LSVRC/2014/index.php#> , Retrieved 16.07.2022.
- [31] <https://www.image-net.org/challenges/LSVRC/2015/index.php#> , Retrieved 16.07.2022.
- [32] <http://places2.csail.mit.edu/> ,Retrieved 16.07.2022.
- [33] <https://www.image-net.org/challenges/LSVRC/2016/index.php#> , Retrieved 16.07.2022.
- [34] B. Zhou, H. Zhao, X. Puig, et.al., “Scene Parsing through ADE20K Dataset”, Computer Vision and Pattern Recognition (CVPR), Hawaii,USA,pp. 5122-5130, 2017.
- [35] B. Zhou, H. Zhao, X. Puig, et.al., “Semantic Understanding of Scenes through ADE20K Dataset”, International Journal on Computer Vision (IJCV), vol.127, pp. 302-321,2019.
- [36] <https://www.image-net.org/challenges/LSVRC/2017/index.php#> , Retrieved 16.07.2022.
- [37] <https://paperswithcode.com/dataset/otb-2015> , Retrieved 17.07.2022.
- [38] A. W. M. Smeulders, D. M. Chu, R.Cucchiara,et.al., “ Visual Tracking: an Experimental Survey”, IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI), vol. 36(7), pp.1442-1468, 2014.
- [39] <https://cemse.kaust.edu.sa/ivul/uav123> ,Retrieved 18.07.2022.
- [40] <https://www.votchallenge.net/vot2013/dataset.html> , Retrieved 18.07.2022.
- [41] <https://www.votchallenge.net/vot2014/dataset.html> ,Retrieved 18.07.2022.
- [42] <https://www.votchallenge.net/vot2015/dataset.html> ,Retrieved 20.07.2022.
- [43] <https://computervisiononline.com/> ,Retrieved 20.07.2022.



- [44] <https://www.crcv.ucf.edu/data/> , Retrieved 20.07.2022.
- [45] <https://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm> , Retrieved 20.07.2022.
- [46] <https://www.videezy.com/> ,Retrieved 20.07.2022.
- [47] <http://lear.inrialpes.fr/data> , Retrieved 20.07.2022.
- [48] [http://oad.simmons.edu/oadwiki/Data\\_repositories](http://oad.simmons.edu/oadwiki/Data_repositories) , Retrieved 20.07.2022.
- [49] <http://www.datawrangling.com/some-datasets-available-on-the-web/> , Retrieved 20.07.2022.
- [50] <http://celltracking.bio.nyu.edu/> , Retrieved 20.07.2022.
- [51] <https://www.votchallenge.net/vot2016/dataset.html> , Retrieved 21.07.2022.
- [52] <https://www.votchallenge.net/vot2017/dataset.html> , Retrieved 21.07.2022.
- [53] <https://www.votchallenge.net/vot2018/index.html> , Retrieved 21.07.2022.
- [54] <https://www.votchallenge.net/vot2019/index.html> , Retrieved 21.07.2022.
- [55] <https://www.votchallenge.net/vot2020/index.html> , Retrieved 21.07.2022.
- [56] <https://www.votchallenge.net/vot2021/index.html> , Retrieved 21.07.2022.