MULTI-FISH DETECTION AND TRACKING USING TRACK-MASK REGION CONVOLUTIONAL NEURAL NETWORK

NAWAF FARHAN FANKUR ALSHDAIFAT

UNIVERSITI SAINS MALAYSIA

2023

MULTI-FISH DETECTION AND TRACKING USING TRACK-MASK REGION CONVOLUTIONAL NEURAL NETWORK

by

NAWAF FARHAN FANKUR ALSHDAIFAT

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

September 2023

ACKNOWLEDGEMENT

First of all, profuse thanks to Allah, who helped me and gave me the ability to achieve this work. I would like to thank my supervisors, Professor Dr. Abdullah Zawawi bin Talib and Associate Professor Mohd Azam bin Osman from the School of Computer Sciences at Universiti Sains Malaysia, for their tireless help and support throughout the PhD period, advice, patience and motivation, which allowed me to grow as a research scientist. With the PhD thesis finished, I would like to thank them for being excellent and wonderful supervisors. I must also express my deep gratitude to the School of Computer Sciences. I dedicate this work to the soul of our master Muhammad, may God bless him and grant him peace, and then to the soul of my father. I would also like to extend my heartfelt thanks to my mother, father-in-law, mother-in-law, wife and children, as well as my brothers and sisters for their help and love in my academic endeavours. In particular, the wife's patience and understanding in which without her I would not have achieved any of them.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ADAS	Advanced Driver Assistance Systems
ADS	Automated Driving Systems
AMIR	Motion, and Interaction cues using RNNs
AP	Averaged over union section over IoU thresholds
BB	Bounding Box
CGMMFD	Collection of Gaussian Mixture Model and Frame Differencing algorithm
CNN	Convolutional Neural Network
ConvLASTM	Convolutional Long Short Term Memory
DPM	Deformable Part Model
Fast R-CNN	Fast Region Convolutional Neural Network
Faster R-CNN	Faster Region Convolutional Neural Network
FPS	Frames Per Second
HOG	Histogram of Oriented Gradient
IoU	Intersection over Union
LRN	Local Response Normalization
LSTM	Long Short Term Memory
MDP	Motion Detection with Particle filter algorithm
MDPF-cache	Motion Detection with Particle Filter algorithm cache
MOT	Multiple Object Tracking

MOTSA	Multiple Object Tracking and Segmentation Accuracy
MOTSP	Multiple Object Tracking and Segmentation Precision
PCN	Principal Component analysis
R-CNN	Region Convolutional Neural Network
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
RoI	Regions of Interest
RPN	Region Proposal Network
SOT	Single Object Tracking
SRNs	Simple Recursive Networks
SVM	Support Vector Machine
Yolo	You Only Look

LIST OF APPENDICES

PENGESANAN DAN PENJEJAKAN BERBILANG IKAN MENGGUNAKAN RANGKAIAN NEURAL BERKONVOLUSI RANTAU JEJAK-TOPENG

ABSTRAK

Pembelajaran mendalam telah menjadi suatu kebiasaan beberapa tahun kebelakangan ini kerana kejayaannya yang cemerlang dalam pelbagai bidang. Tesis ini terutamanya memberi tumpuan kepada pengesanan dan penjejakan berbilang ikan dalam video di dalam air. Kaedah pengesanan berbilang ikan yang sedia ada untuk video di dalam air mempunyai kadar pengesanan yang rendah dan mengambil masa dalam proses latihan kerana keadaan di dalam air dan cabaran-cabaran lain; seperti ikan dengan saiz, bentuk, warna dan kelajuan yang berbeza serta pergerakan pelbagai arah dan pertindihan ikan. Banyak kaedah penjejakan berbilang objek untuk video adalah untuk memantau manusia dan bukannya ikan, yang mana bentuk dan pergerakan pelbagai arah ikan perlu dipertimbangkan, dan mengalami ketidakupayaan untuk mengekstrak peta ciri yang penting dari bingkai dan ketidakupayaan untuk mengendalikan setiap objek yang dikesan sepanjang masa. Oleh itu, penyelidikan ini bertujuan untuk menambah baik dan meningkatkan kaedah pengesanan dan penjejakan berbilang ikan dalam video di dalam air berdasarkan algoritma pembelajaran mendalam terkini. Kaedah pengesanan berbilang ikan yang dicadangkan melibatkan tiga langkah utama: 1) menambah baik tulang belakang ResNet-101 untuk pengesanan berbilang ikan, 2) meningkatkan kaedah Rangkaian Cadangan Rantau (RPN), 3) kaedah pengesanan berbilang ikan yang lebih baik dari segi ketepatan dan dengan masa latihan dan pengujian yang lebih rendah dengan menggunakan kaedah-kaedah tersebut. Penjejakan berbilang ikan yang dipertingkatkan (Track-Mask R-CNN) pula dicadangkan berdasarkan pembelajaran mendalam yang menggabungkan LSTM dan kaedah kotak

pembatas untuk penjejakan yang lebih tepat, dan seterusnya menghasilkan kaedah bersepadu pengesanan dan penjejakan berbilang ikan yang dipertingkatkan berdasarkan pembelajaran mendalam. Akhirnya, kaedah-kaedah yang dicadangkan digunakan pada data umum untuk menyiasat sama ada kaedah-kaedah berkenaan juga boleh menjadi kaedah yang lebih baik untuk pengesanan dan penjejakan berbilang objek. Kaedah pengesanan berbilang ikan yang dicadangkan mempamerkan prestasi yang lebih baik dari segi ketepatan dan memerlukan masa latihan dan pengujian yang kurang berbanding dengan kaedah-kaedah canggih yang ada. Kaedah penjejakan berbilang ikan yang dicadangkan (Track-Mask R-CNN) juga mempamerkan ciri-ciri dipertingkatkan yang sama berbanding kaedah canggih yang ada. Ketepatan 86.7% dan 78.9% telah dicapai masing-masing untuk pengesanan dan penjejakan berbilang ikan. Hasil penggunaan kaedah-kaedah yang dicadangkan pada set data umum juga menunjukkan bahawa kaedah-kaedah ini juga boleh digunakan untuk pengesanan berbilang objek dan penjejakan berbilang objek kerana kaedah-kaedah berkenaan mengatasi kaedah canggih yang ada. Hasil penyelidikan ini boleh digunakan dalam industri akuakultur untuk mengesan dan memantau ikan di dalam air dengan cara yang lebih tepat dan cekap.

MULTI-FISH DETECTION AND TRACKING USING TRACK-MASK REGION CONVOLUTIONAL NEURAL NETWORK

ABSTRACT

Deep learning has become more common in recent years due to its excellent results in many areas. This thesis primarily focuses on multi-fish detection and tracking methods in underwater videos. The existing multi-fish detection methods for underwater videos have a low detection rate and consumes time in the training and testing process due to the underwater conditions and the overfitting during training. Many multi-fish detection and tracking methods for underwater videos (based on deep learning) where low accuracy for multi-fish tracking and occlusion instances during multi-fish tracking leads to inability to distinguish edges, and inability to handle each detected object over time. Therefore, this research aims to improve and enhance methods for multi-fish detection and tracking in underwater videos based on the latest deep learning algorithms. The proposed improved multi-fish detection method involves three main steps: 1) Improving ResNet-101 backbone for better fish detection, 2) Enhancing the Region Proposal Network (RPN) method based on Faster R-CNN for multi-fish detection and 3) An improved multi-fish detection method in terms of accuracy and with a lower training and testing times by utilising the aforementioned methods. The proposed multi-fish tracking method (Track-Mask R-CNN) also exhibits similar enhanced characteristics compared to the state-of-art methods (using fish dataset). An accuracy of 86.7% and 78.9% have been achieved for the proposed multi-fish detection and tracking respectively. The results of applying the proposed methods on general datasets (COCO dataset for multi-object detection and KITTIMOTS dataset for multi-object tracking) also show that the methods can also be used for multi-object detection and

multi-object tracking as they outperform other state-of-the-art-methods. The outcomes of this research could be used in aquaculture industries to track and monitor underwater fish in a more accurate and efficient manner.

CHAPTER 1

INTRODUCTION

1.1 Background

Although the process of tracking objects using mathematical methods such as Kalman filter and background subtraction (Lantsova *et al.*, 2016) has become very common among researchers, and most researches in this area have been performed on humans instead of fish. Meanwhile, the research on fish tracking involves various processes, such as the use of many cameras in different places to determine the shape and movement of fish (Lee *et al.*, 2015). Notably, fish tracking in underwater videos is a major challenge for researchers due to the difficulty in determining the shape, swimming pattern and movement at different speeds and direction of the fish, as well as the underwater speed and changes in the lighting (Mohamed *et al.*, 2020). Most fish tracking algorithms perform tracking on a single fish. Issues regarding overlapping objects are also present. These issues must be resolved to improve tracking of fish in underwater videos. Furthermore, tracking and object detection algorithms depend on the location, shape and movement of the object (Lee *et al.*, 2015; Yang *et al.*, 2005; Labao & Naval, 2017). Most of the tracking algorithms focusing on the shape and movement of an object (Voigtlaender *et al.*, 2019; Lantsova *et al.*, 2016) divide the video into several frames.

1.2 Deep learning approaches for multi-object detection

Deep learning is a branch of machine learning which has been showing significant impacts in recent years due to its effectiveness in many fields (Sarker, 2021). The main challenge in detecting multiple objects in the same image is solved through these architectures. Deep learning comprises a range of object detection methods with AlexNet, VGG, ResNet-101, and GoogleNet being the most common methods. Krizhevsky *et al.* (2012) proposed AlexNet, in which its network comprises five convolution layers; max-pooling layers, dropout layers and three fully-connected layers. Deep learning also contains a range of architectures for multi-object detection with R-CNN, Fast R-CNN and Faster R-CNN being the most common architectures. The main challenge in detecting multiple objects in the same image is solved through these architectures. Notably, Convolutional Neural Network (CNN) is the most effective deep learning technique for detection compared to other algorithms (Krizhevsky *et al.*, 2012).

Recent researches on deep learning have received significant attention due to its effectiveness in many fields, such as computer vision (Zeiler & Fergus, 2014), image detection (Ren *et al.*, 2015; Girshick *et al.*, 2014) and segmentation (He *et al.*, 2017; Bolya *et al.*, 2019; Cai & Vasconcelos, 2021). Therefore, there is a need to improve detection methods for fish, based on the latest deep learning algorithms. Accordingly, various studies have also developed algorithms to improve localisation and segmentation algorithms (He *et al.*, 2017; Bolya *et al.*, 2019; Cai & Vasconcelos, 2021). These methods used image or video, and the difference between them is that the video is split into frames before inputting into the methods.

1.3 Multi-object tracking

Multi-object tracking methods analyse videos to detect, classify and track objects involving animals or pedestrians. The common methods for detection and tracking in videos include Kalman Filter, Background Subtraction, Deformable Multiple Kernel and Long Short-Term Memory (LSTM). Most methods function on a single object and could also be applied for multi-object detection and tracking. Accuracy of multi-object detection and tracking can be improved by using big data training and developing new methods that offer correct features from the training data. There has been a notable interest in deep learning which has been successfully applied in computer visions, such as in multi-object detection (Girshick *et al.*, 2014; Ren *et al.*, 2015) and tracking of objects (Voigtlaender *et al.*, 2019). The goal of deep learning is to replace hand-crafted features with the high features of raw pixel values. Furthermore, several tracking methods of a single object or multiple objects are present, namely Single Object Tracking (SOT) and Multiple Object Tracking (MOT) (Ciaparrone *et al.*, 2020). The most common object tracking methods based on deep learning involve detecting the objects and grouping each detected object in a bounding box. The bounding boxes are the output from the MOT algorithms, where a bounding box consists of a number representing a detected object. Therefore, in this research, methods based on deep learning with higher performance in solving multi-object tracking for fish and functioning on large videos should be identified.

1.4 Multi-fish detection and tracking

Problems encountered in fish detection and tracking largely originated from the marine environment and they should be solved. Monitoring of fish is important to

obtain information on fish behaviours within the marine ecosystems. Tracking and distribution of the fish within the marine ecosystems allow researchers to gather information regarding the health of the marine ecosystem. This information could then be used in a monitoring system for fish, and for identifying changes in the marine ecosystem (McLaren *et al.*, 2015). Notably, identifying changes in the environmental conditions require monitoring of the movement of different types of fish, especially in areas where some species are susceptible to extinction due to the loss of their habitat, industrial pollution and climate change (Jennings & Kaiser, 1998). Monitoring the effects of preventive measures by estimating and providing biomass requires sampling from the marine environment in the oceans or rivers. Marine biologists hold high interests in using non-destructive sampling techniques (McLaren *et al.*, 2015). Whilst manual processing on underwater videos requires a long time, there is also a risk of high cost and susceptibility to error. Therefore, automatic underwater video manipulation of fish detection and biomass measurement are attractive alternatives. Nevertheless, tracking is challenging due to obstructions that exist within the water system, such as coral reefs, different fish sizes, shapes and colours, and diverse marine behaviours and conditions such as overlapping of fish, lights and noise. The study by Lantsova *et al.* (2016) proposed several approaches using background subtraction, Kalman Filtering and Viola-Jones algorithm to detect movement. However, not all methods produce the ideal result for the detection of the fish. Another study by Aljarrah *et al.* (2012) suggested a fish detection model based on the signature and principal component analysis. However, this model is time-consuming and not suitable for fish tracking when the overlapping of fish occurs. Furthermore, Lee *et al.* (2014) proposed an enhanced movement detection method using the particle filter algorithm (MDPFcache), which incorporates a

cache to store information regarding the fish position. Following that, the next possible move of the fish is predicted. Although an output consisting of a bounding box for each fish is provided in this method, it does not function in an underwater environment which contains high noise and is not stable for rotation movement. Various algorithms were applied in fish detection systems. The Faster R-CNN (Ren *et al.*, 2015) was used for detecting and localisation of fish based on the image. Although previous studies focused on fish detection, the studies on fish movement, tracking and fish overlapping have not been conducted. It is challenging to determine the location and the size of fish including the absence of addressing fish overlapping problem based on deep learning methods. In this research, an enhanced integrated method for fish detection and tracking in underwater video based on deep learning is proposed. Object detection and tracking methods are improved and enhanced for multi-fish detecting and tracking. Notably, integrating these methods would offer more accurate results in multi-fish detection and tracking.

1.5 Research motivation

Notably, deep learning methods have had a significant impact on the field of machine learning in recent years (Hinton *et al.*, 2006). The methods also aim to learn multiple data representation levels from low-level representation to gain an understanding of data, such as text, sound and image to high-level representation to provide more data connotations (Krizhevsky *et al.*, 2012). Moreover, CNN has demonstrated better results, which even surpasses the performance of the human level of object detection (Keuper *et al.*, 2016; Jain *et al.*, 2016). Besides, RNN has demonstrated remarkable success in many learning tasks end-to-end (Donahue *et al.*, 2015; Fragkiadaki *et al.*,

2015). With these advancements, it would be interesting to deal further on enhancing detection and tracking of multiple fish in underwater videos based on deep learning. Labao & Naval (2017) employed the use of CNN and Dense CRFs to localise and segment fish. The process of locating, determining different sizes, establishing the fish status and identifying the behaviour of fish from multiple locations in the underwater environment remains to be a challenging task in marine ecosystems. In this research, an enhanced integrated method for fish detection and tracking in underwater video based on deep learning is proposed. Object detection and tracking methods are improved and enhanced for multi-fish detection and tracking. Notably, integrating these methods would offer more accurate results in multi-fish detection and tracking.

1.6 Problem statement and research questions

The existing fish detection methods based on deep learning for underwater videos (Agarwal *et al.*, 2020; Ma *et al.*, 2018; Yu *et al.*, 2020) have a low detection rate and are time consuming in the training and testing process due to the adoption of the ResNet-101 as backbone.

ResNet-101 backbone is a deep learning architecture with a series of blocks that were used to overcome the gradient vanishing problem by adding shortcut connections (He *et al.*, 2016). However, they still have other problems, such as using a large filter size for the first convolution layer, identifying ResNet-101 blocks that have not received adequate repetitions and identifying ResNet-101 blocks that have received more than adequate repetitions. As a result, it leads to low accuracy and time-consuming training and testing process. The existing multi-fish detection methods (Zeng *et al.*, 2021; Wu *et al.*, 2018) adopt RPN method based on Faster R-CNN. Faster R-CNN is deep learn-

ing architecture which uses RPN method to generate the region proposals (Ren *et al.*, 2015). However, problems still exist, such as low accuracy for multi-fish detection due to the problem of overfitting during training. As a result of overfitting, it leads to fitting mismatch between intersection over union (IoU) for which the detector is optimal and those of the IoU thresholds during testing phase, and this will decrease the accuracy of multi-fish detection. Therefore, there is a need to overcome these issues especially in multi-fish detection and further exploit deep learning in this domain.

The existing multi-fish detection and tracking methods (Zhiping & Cheng, 2017; Marini *et al.*, 2018; Salman *et al.*, 2020) through videos (based on deep learning) where problems still exist, such as low accuracy and occlusion for multi-fish tracking. Therefore, the proposed method will use general multi-object detection (Ren *et al.*, 2015) tracking (Voigtlaender *et al.*, 2019) due to their proven effectiveness compared with state-of-the-art methods. However, problems still exist, such as low accuracy and when there are occlusion instances.

The existing multi-object tracking method, Track R-CNN (Voigtlaender *et al.*, 2019) has improved the effectiveness of tracking, by incorporating ConvLSTM which provides a feature map with a frame sequence for each frame from ResNet-101 over time to RPN. However, the inability of Track R-CNN to handle the occlusion instances of multi-fish leads to difficulties of segmenting instances. Consequently, it causes low accuracy of tracking multi-fish. Therefore, there is a need to enhance multi-object tracking method to solve the occlusion problem specifically for multi-fish tracking in underwater video and further exploit deep learning in this domain. Even though this research focusses on multi-fish detection and tracking, it would be interesting to inves-

tigate whether the proposed multi-fish detection and tracking methods in this research could also be used for general multi-object detection and tracking.

Accordingly, the research questions developed in this research are as follows:

1. How to improve multi-object detection methods based on deep learning to detect multiple fish in underwater videos for a more accurate detection and a shorter training and testing time?
2. How to enhance existing multi-object tracking methods based on deep learning specifically for multi-fish tracking in underwater videos for more accurate tracking?

1.7 Objectives

This research aims to provide improved and enhanced methods based on deep learning to detect and track the fish movement in underwater videos. Accordingly, the research objectives are as follows:

1. To improve multi-fish detection method based on deep learning for better accuracy and shorter training and testing time.
2. To enhanced Track R-CNN for a better accuracy multi-fish tracking by solving occlusion problem.
3. To investigate whether the proposed methods suitable for general object detection and object tracking.

1.8 Research contributions

The contributions of this research are as follows:

1. An improved Faster R-CNN method for multi-fish detection which includes the following steps.
   (a) An improved ResNet-101 backbone for fish detection by changing the size of the filter in the first convolutional layer, repeating the ResNet-101 blocks and connecting it with a new convolution layer.
   (b) An improved multi-fish detection method by enhancing RPN method based on Faster R-CNN by adding an additional convolutional layer in RPN method. The RPN method is further enhanced by repeating the detection model based on IoU for which the detector is optimal.
   (c) An improved multi-fish detection method in terms of accuracy and with a shorter training and testing time by utilising the aforementioned improved ResNet-101 backbone and enhanced RPN.
2. An enhanced multi-fish tracking method based on deep learning that combines instances segmentation with Convolution Long Short-Term Memory (ConvLSTM). It allows to take a feature that contains a location in instance level with sequences for each fish detection to allow tracking of each detected fish over time. Thus, this method has resulted in an enhanced integrated multi-fish tracking method by solving occlusion problem.
3. A multi-object detection and tracking methods suitable for general multi-object detection and multi-object tracking which are based on the proposed methods

for the improved multi-fish detection method and enhanced multi-fish tracking method.

1.9 Benefit of research

Detection and tracking of multi-fish are important to obtain information on fish behavior within the marine ecosystems without requiring human observers. Detection and tracking of multi-fish within marine ecosystems allow researchers to gather information regarding the health of the marine ecosystem. This information can be used to monitor the changes affecting the marine ecosystem. These changing environmental conditions warrant monitoring the interaction between fish and can be used to protect marine ecosystems. Further, detection and tracking of multi-fish are important because it helps us to understand the distribution and availability of fish and a basis for a long-term fishery management.

This research also provides a better method for detecting and tracking objects that can be used to monitor other objects besides fish, such as cars and pedestrians, that can be used in Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS). One of the essential components of an ADAS and ADS are the perception and understanding of the environment through detection and tracking pedestrians, lane lines, and other cars on the road to make correct decisions. ADAS and ADS have been developed to reduce traffic and increase safety on roads, translating to considerable economic benefits.

1.10 Scope and limitations

The scope of this research is restricted in several ways. This research relates to multi-fish detection and tracking in underwater videos i.e., they are performed on underwater fish in videos. The video clips were taken from many places of the seas and oceans, and various environments. There are many shapes and sizes of fish in the video clips, and they move in different directions. There are also varying numbers of fish in a video. To obtain a broader range of datasets containing different densities of fish, two datasets were used. The first is from the Fish4Knowledge (Fisher *et al.*, 2016) project which uses underwater videos to study the marine ecosystem. 15 videos from the Fish4Knowledge videos were used in the training part and 2 videos were used in the testing part. In order to enhance the fish dataset, a second dataset was collected from various other sources (YouTube, 2019). The second dataset consists of 2000 frames, and out of which 1,800 frames were used for the training part and 200 frames were used for the testing part. The videos depict seas, oceans, or marine complexes that contain complex and more realistic imagery especially underwater videos, In addition padding technique was used for training and testing videos to make the frames in the videos have the same size which is 1024 x 1024 . However, an existing preprocessing method is required to split and clean the underwater video. The videos were split into frames using the Blender software (Gschwandtner *et al.*, 2011) and subsequently preprocessing method with the constant luminance method (for brightness optimisation) with sharpness correction (Turkowski, 1990) was applied to control the light and remove blurring, and thus obtaining high-quality images. The constant luminance method aims to reproduce clearer images and focus the lighting on places where the lighting is low. Then, in the implementation, sharpness correction was applied to

increase or decrease the sharpness of the frame and reduce the blurring within each frame. The underwater fish have many features such as different colours, sizes, shapes and patterns of movement. The outcomes of this research were also applied on general datasets in order to see whether the proposed methods could also be applied on general objects. Therefore, benchmark databases for multi-object detection (COCO dataset) and tracking (KITTIMOTS dataset) were also used in this research.

1.11 Overview of research approach

The research consists of four stages. The first stage improves object detection methods by improving RestNet-101, resulting in an improved the ResNet-101 backbone to solve low detection rates, as shown in Section 3.3.1. Then, RPN method based on the Faster R-CNN is enhanced to solve low accuracy for multi-fish detection as well as training. This stage facilitates the improved multi-fish detection in underwater videos to deal with a more accurate detection and lower computational time, as shown in Section 3.3.2. This research also investigates whether the proposed methods are suitable for general multi-object detection.

The second stage is to enhance multi-fish tracking method based on deep learning to solve the occlusion problem. This stage has resulted in an enhanced and integrated multi-fish tracking method that increase the performance of an existing multi-fish tracking method by combining instance segmentation with Convolution Long Short Term Memory (ConvLSTM), as shown in Section 3.4. In addition, this research also investigates whether the proposed methods are suitable for general multi-object tracking. This is by the third stage in which an integrated multi-fish detection and tracking

system incorporating the proposed methods is implemented, as shown in Section 3.5. Training and evaluation of the proposed enhanced methods from the previous stage will be performed, as shown in Section 3.6.

1.12 Thesis organization

The remaining sections of this thesis are structured as follows:

- **Chapter 2:** In this chapter, a literature review is presented on various object detection and tracking methods based on non-deep learning as well as on deep learning. This chapter also presents a review of the existing methods for multi-fish detection and tracking.

- **Chapter 3:** This chapter describes the research methodology of the proposed work for this research. This chapter presents the methods for the improved backbone ResNet-101 and an enhanced RPN for multi-fish detection. This chapter also presents an enhanced tracking method by combining instances segmentation and ConvLSTM to track fish in underwater. The methodology to investigate whether the proposed methods in this research can also be used for general multi-object detection and tracking is also presented. The construction of the datasets and the training process (learning phase) and the evaluation methodology for multi-fish detection and tracking are also presented. Finally, this chapter presents the implementation of a complete system for multi-fish detection and tracking.

- **Chapter 4:** The results and discussion are presented in Chapter 4. This chapter includes the results of the evaluation of the proposed methods in this research.

- **Chapter 5:** This chapter presents the conclusion and the future work of this research

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, a literature review of many existing detection and tracking methods is presented. Section 2.2 presents the design and overview of the literature review. Section 2.3 presents a literature review on various multi-object detection methods. In this section, methods based on non-deep learning are reviewed. Also, in this section, a literature review on object detection and multi-object detection methods based on deep learning is presented. Section 2.4 presents a literature review on various methods of multi-objects tracking. Methods reviewed are based on non-deep learning and deep learning. Lastly, Section 2.5 presents literature review on multi-fish detection and tracking methods, which is also based on non-deep learning and deep learning.

The Overview of the literature review is as shown in Figure 2.1.

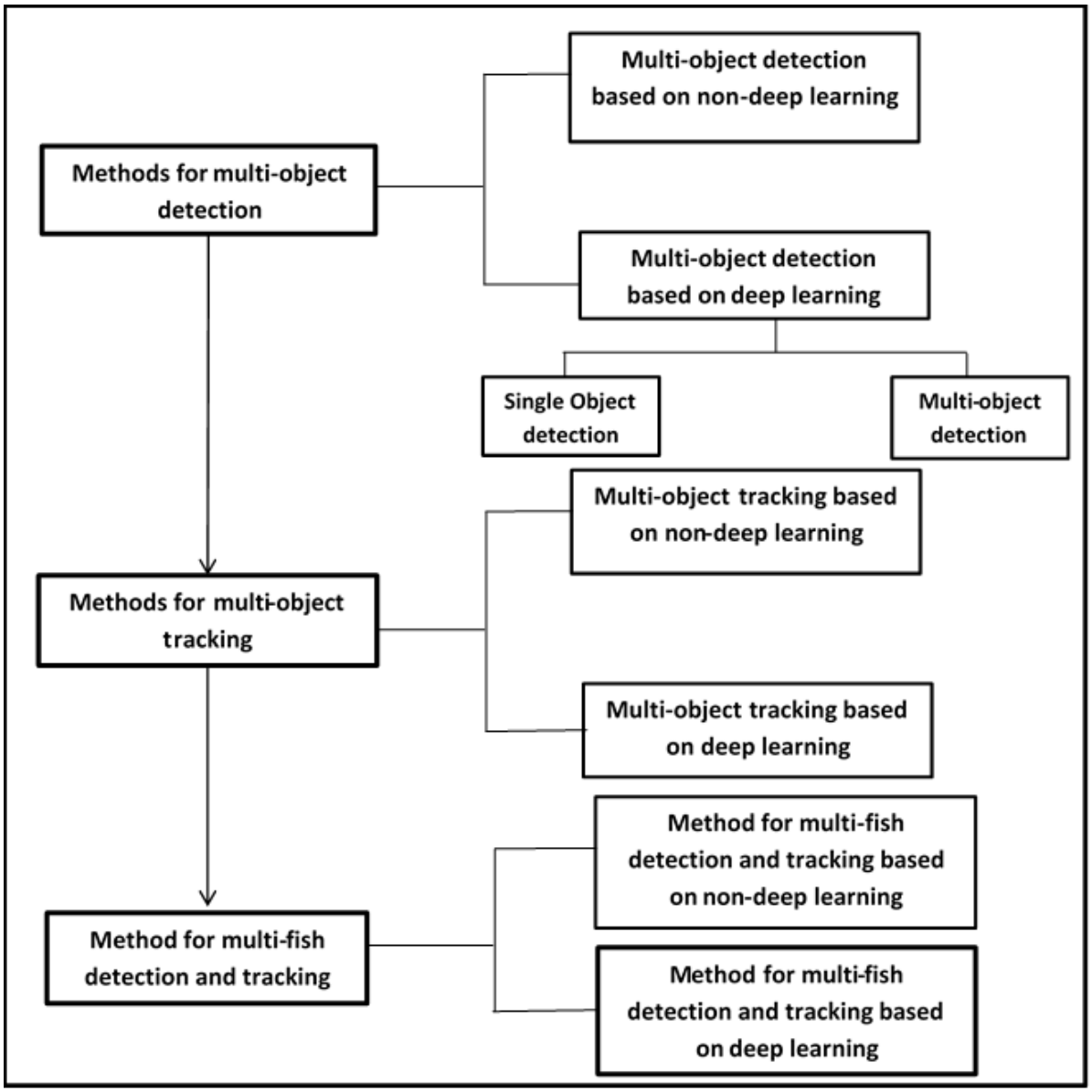

Figure 2.1: Overview of literature review

2.2 Multi-object detection methods

2.2.1 Multi-object detection methods based on non-deep learning

Object detection methods deal with objects based on input image. Object detection is an essential part of computer vision. Chang & Krumm (1999) used the colour Co-occurrence Histogram (CH) for detection. It allows adding of geometric details to the colour histogram to save the track of pixel pairs. In the test stage, the model is

matched in the sub-region to find all objects through using a false alarm rate to choose the better parameters suitable for the model. Ramesh & Mohan (2007) presented an algorithm that applies many steps through a distributed system. The algorithm is split into two levels, the upper level, and the lower level. The upper level is interested in the cognitive process, and the other level is interested in the biological process of the human brain. Olaode & Naghdy (2019) detailed a shape context, colour histogram and completed a local binary pattern (CLBP) approach to classify various classes of objects. The database for their research is ETH-80, in which the accuracy attained is higher.

Kim & Kweon (2007) presented an algorithm used as a codebook to minimise the intra classes. The algorithm depends on a cookbook to minimise the effect of surface marking. In this algorithm, there are three stages. The first stage removes the surface marking part in the training stage, the second releases the codebook, and the last stage uses Nearest Neighbour Classifier (NNC) and support vector machine (SVM) to classify different matrices. The algorithm was applied to the Caltech-101 database which consist of pictures of objects belonging to 101 categories. About 40 to 800 images per category. Otoom *et al.* (2008) rated the performance of various feature groups for choosing the better feature group that is more suitable for the classification of objects. However, the experiment result shows that the algorithm depends on the classification of the object's statistics of the geometric primitives feature group, and is better than that of the Scale Invariant Image Transform (SIFT) key points histogram using different classifications and evaluation schemes. The result is higher for detection accuracy compared to the second-best approach based on the SIFT keypoint programs. Wang *et al.* (2013) indicated that a model depends on comparative object similarity for

learning object models having fewer training. The model modifies the detection and classification algorithms to combine similarity constraints. Although, the model has drawbacks since it had a small number of items in the dataset.

Mokji & Bakar (2007) presented a new algorithm for the Grey Level Co-occurrence Matrix (GLCM) computation that depends on the Haar wavelet transform technique to minimise the computational problem through pixel entries, and thereby increasing the accuracy. Rockinger (1997) proposed a new technique based on a shift-invariant wavelet transform for the fusion of spatially registered images and image sequences. This technique has better results in image sequence problems. Moreover, the combined techniques have an advantage in temporal stability and consistency. Dao & Vemuri (2002) proposed that the Neural Network (NN) model could apply controlled input data files for intrusion detection in the computer network. The model was compared with many different NN models, such as the gradient descent backpropagation (BP), the gradient descent with momentum, the variable learning rate gradient descent BP, the conjugate gradient BP, and the quasi-Newton method where the best model depends on the users when logging into the computer network (Cover & Hart, 1967) suggested K-NN as an algorithm to classify the object. The algorithm's notion is centred on the nearest feature space in the training process, which is considered as the simplest algorithm in non-deep learning algorithms. The algorithm consists of two stages, the training stage and test training. In the training stage, the algorithm keeps feature vectors for the label object, while in the test stage, the un-labelled object is transferred to the nearest label, as shown in Figure 2.2. The advantage of the k-nearest neighbour algorithm is that it can be used with various models. However, the disadvantage of K-NN algorithm is when the dataset is small and does not have many features, leading

to an error in classifying the object.

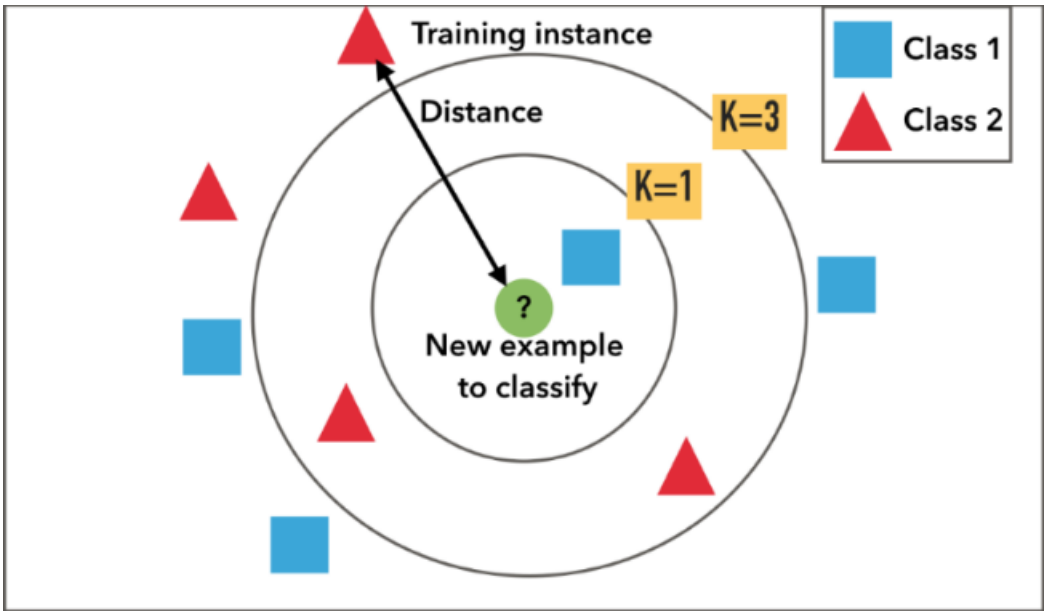

Figure 2.2: K-NN classification (Cover & Hart, 1967)

Burges (1998) applied the SVM method using various levels of space to classify an object. The SVM method uses a different perspective that maximises the edge when having other classes by dividing it. The SVM method consists of two parts: the training and test parts. In the training part, the process splits the points to the nearest point of classes, while the test part predicts the end in space to which the classes belong to, which depends on the point's location. The SVM method for classification depends on the training data that predicts the class labels in the test stage. The advantage of SVM classification is that the SVM method provides a good result in different datasets, even for a dataset having a small number of classes in the training stage. However, the disadvantages of the SVM method does not perform very well when the data set has noise such as overlapping, and is not suitable for large data sets.

Nevertheless, Shah & Gandhi (2004) suggested that developing an Artificial Neu-

ral Network (ANN) algorithm to design and improve a hierarchical network depends on incorporating textural features. They noted the importance of textural features to enhance image classification using the ANN algorithm. Haykin & Network (2004) proposed an (ANN) algorithm to solve linearity and loss associated with mathematical problems by using neurons to deal with available data following feature extraction from the image, and using a backpropagation algorithm in the training stage to train, choose and update the better weight for neurons towards a suitable the dataset. ANN has many advantages since it can be used for the classification or regression of images but, it suffers from overfitting and vanishing gradient problems.

A comparative study of various techniques used for object classification and detection based on non-deep learning algorithms such as SVM, KNN, and ANN has been undertaken. Based on the analysis, each method has both advantages and disadvantages based on the dataset. However, it would be more interesting to deal further with deep learning for object detection. Table 2.1 summarises the advantages and disadvantages of the object detection methods based on the non-deep learning.

Table 2.1: Summary of advantages and disadvantages of object detection methods based on non-deep learning

Author (year)	Technique	Dataset	Advantages	Limitations
Cover & Hart (1967)	KNN	Category classification problem for nearest neighbour (NN) based on the infinite data set	Can work in any model	Utilises all features each time, and consumes more time
Rockinger (1997)	Shift-invariant Wavelet transform	Fusion of spatially registered images and image sequences	Ability to work with temporal stability and consistency	Slow to detect and classify the object
Chang & Krumm (1999)	Colour cooccurrence ogram Ragged Right	Wood's model images have 12 curves match images of Woody, each of the curves represents model images	Ability to work in unclear background	Slow to detect and classify objects and poorly adapted on a large dataset
Shah & Gandhi (2004)	ANN and textural Features	Textural dataset	Ability to work with different classes	Could not deal well with different overfitting problems and could not avoid the vanishing gradient problem

Continued on next page

Table 2.1 – *Continued from previous page*

Author (year)	Technique	Dataset	Advantages	Limitations
Kim & Kweon (2007)	Codebook, NNC, and SVM	Caltech-101 dataset	Deals with the intra classes	Poorly adapted on large datasets
Mokji & Bakar (2007)	Grey Level Co-occurrence Matrix and Haar wavelet	Brodatz dataset	Minimises computation time	Poorly adapted on large datasets
Wang *et al.* (2013)	Comparative object similarity to learning in training	PASCAL VOC 2007	Can work with few training stages	The system solution is not specific for solving classification problems

2.2.2 Single object detection based on deep learning

Deep learning comprises a range of object detection methods, with AlexNet, VGG, ResNet-101, and GoogleNet being the most common methods. Krizhevsky *et al.* (2012) proposed AlexNet, in which the network consists of five convolution layers, maxpooling layers, dropout layers and three fully-connected layers. The designed structure was used for classification with 1000 possible categories using SoftMax function and Rectified Linear Unit (ReLU) for the nonlinearity functions, including the Local Response Normalization (LRN). The database was trained on image net data, which stored over 15 million annotated images from over 22,000 categories. The model uses batch stochastic gradient descent, with specific values for momentum and weight decay. Although the algorithm achieved the highest accuracy in the experiments, the classification of objects and training requires long duration and many parameter.

Simonyan & Zisserman (2014) proposed to build a network of up to 19-layer CNN called VGG, which accurately utilises 3x3 filters and a padded step of more than 2x2 max layers combining with Step 2 using SoftMax. The localised normalisation response, which was only used once throughout the entire network, was functional on image classification and localisation tasks. The algorithms achieved accuracy rates of 91.2% and 75.2% based on ILSVRC-2014.

In another study by (Szegedy *et al.*, 2015), they presented the inception model, which used 22 layers of Convolution Neural Network (CNN) to improve performance and computational load. In this algorithm, multiple layers could function in parallel. They also used fully connected layers with soft-max probabilities to achieve the final recognition,

He *et al.* (2016) developed a system to detect image using residual learning, which involves a network consisting of 101 CNN layers, to perform detection and localisation. Following the use of residual learning on the recurrent unit after every two layers are the compression of the spatial volume from 224 x 224 to 56 x 56. In this method, the average pool was used instead of the fully connected layers, while the SoftMax probabilities were used to achieve final detection.

A review has been performed on a comparative study of a range of methods for object detection based on deep learning. The review focuses more on architectures for object detection based on deep learning. Furthermore, common methods for the above-mentioned purposes were based on deep learning methods, such as AlexNet, VGG, ResNet-101, and GoogleNet, to achieve a more accurate classification. The

ResNet-101 is deemed to be the best method to obtain high accuracy in detection object. However, they still have other problems such as identification of ResNet-101 blocks that have not received adequate training, identification ResNet-101 blocks that have received more than adequate training, and usage of a large filter size for the first convolution layer. For better accuracy and lower training time, researchers have suggested predicting the object to increase accuracy by training the methods to achieve higher accuracy. The methods to improve the accuracy are essential to obtain better accuracy in detecting the objects in the image. Table 2.2 summarises different techniques based on deep learning for object detection. Overall, from the review, it was found that the ResNet-101 architecture offers the ideal method to achieve higher accuracy in detecting objects.

Table 2.2: Summary of different CNN architectures on object detection based on deep learning

Author (year)	CNN Implementation	Dataset	CNN Architecture	Advantages	Limitations
Krizhevsky *et al.* (2012)	CNN method consisting of five convolution layers, max-pooling layers, dropout layers, and three fully-connected layers	Image-net	AlexNet	Reduces the complexity of the network and works with a big dataset and high accuracy with traditional classification	Takes a long time to classify an object

Continued on next page