Underwater Image Processing using Convolutional Neural Networks

A Project Report

submitted by

RIYA JOSEPH

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS

JUNE 2021

CERTIFICATE

This is to certify that the thesis titled **Underwater Image Processing using Convolutional Neural Networks** , submitted by **Riya Joseph**, to the Indian Institute of Technology Madras, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Place: Chennai

Date: 23rd June 2021

Prof. A.N.Rajagopalan
Project Guide
Professor
Dept. of Electrical Engineering
IIT Madras, 600 036

ACKNOWLEDGEMENTS

First of all, I would like to extend my sincere and heartfelt gratitude to my guide Prof. A.N.Rajagopalan who gave me an opportunity to work in the area of my interest. The world is experiencing unprecedented challenges due to the pandemic and this is indeed the most difficult times for everyone alike. He has been very considerate to my concerns and has always motivated me by providing very interesting challenges to work on. Throughout my work he has patiently listened to all my ideas, raised valid criticisms and has constantly given the necessary guidance to carry forward this work in the right direction. For this and more I'm forever grateful to him.

I would like to extend my sincere thanks to all the scholars of Image Processing and Computer Vision (IPCV) Lab, Dept. Of Electrical Engg, IIT Madras who shared very inspiring thoughts and ideas during group meetings on the futuristic research opportunities in the area of computer vision. This has given me a better understanding of the state of the art technologies.

Last but not the least, I would like to thank my family and friends. Words alone cannot express what i owe them, for providing me support, encouragement and hope during this pandemic.

ABSTRACT

KEYWORDS: Imaging Sonar; SfM; 3D reconstruction; Object Tracking; MDNet; VWIE; MSER

Underwater image processing has received considerable attention in the recent times. Detection, Tracking, 3D Reconstruction and Classification of underwater objects are particularly important for maritime security. Convolutional Neural Networks (CNN or ConvNets) have made breakthrough advances in various computer vision tasks and are emerging as a promising technique for underwater images as well. In this work, CNN models are extended to two problems: (1) Enhancement of underwater optical images for 3D reconstruction using Structure from Motion (SfM) algorithm (2) Tracking of underwater objects in Sonar images.

Underwater optical cameras capture high resolution images of underwater scene. The images captured from multiple viewpoints can be used to generate 3D point clouds. It is used in applications such as mine detection, inspection of underwater structures, ocean archaeology & exploration etc. However underwater images are faced with different set of challenges such as poor visibility due to haze, non-uniform illumination, color cast etc. The quality of the images in underwater cameras also deteriorates because of effects of scattering and light absorption. In this work, a CNN based method which is adapted from Water-Net, is studied for enhancing the underwater images. Also an open source 3D reconstruction pipeline COLMAP, is studied for 3D point cloud generation using enhanced images.

Imaging Sonars or acoustic cameras are essential for providing underwater surveillance capabilities in turbid water conditions where the optical cameras mostly fail. When it comes to strategic military applications like harbor security, AUV/ROV navigation, obstacle avoidance etc, the capability to accurately track the objects of threat is of significant interest. Extracting useful information from the sonar images is a chal-

lenging task because of its inherent imperfections like low resolution, lack of color information and foreground objects becoming indistinguishable from background clutter. In this work, a method is proposed to track objects in a sequence of Sonar images produced by BlueView Oculus Sonar which is a multibeam forward looking Sonar. A CNN based tracker adapted from MDNet is implemented. A method based on Maximally Stable Extremal Regions (MSER) and Variance Weighted Information Entropy (VWIE) is proposed to be added in the MDNet pipeline for better handling of track failures due to object going out of view, occlusions and noisy background. Results obtained on real sonar data show that the proposed framework can track the object accurately.

TABLE OF CONTENTS

Table of Contents (continued) **Page**

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

The following are some of the commonly used terms in this thesis:

ROV Remotely Operated Vehicles. These are equipped with optical cameras and Sonars for Underwater surveillance/ exploration.

OCULUS OCULUS is a new generation multibeam sonar, designed for a wide variety of underwater applications. It is manufactured by blueprint Subsea. It has got dual frequency of operation at 750kHz and 1.2MHz. More Details can be found at `https://www.blueprintsubsea.com/oculus/`

COLMAP COLMAP is a general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline with a graphical and command-line interface. It offers a wide range of features for reconstruction of ordered and unordered image collections.

Meshlab Meshlab is an open source software which is used for editing, rendering, texturing of meshes. It has been used for viewing the final reconstructed objects.

ABBREVIATIONS

IITM	Indian Institute of Technology Madras
Sonar	Sound Navigation and Ranging
PPI	Plan Position Indicator
CNN	Convolutional Neural Networks
Conv Layers	Convolutional Layers
FC Layers	Fully Connected Layers
GAN	Generative Adversarial Networks
WB	White Balance
HE	Histogram Equalization
GC	Gamma Correction
SfM	Structure from Motion
ROV	Remotely Operated Vehicle
AUV	Autonomous Underwater Vehicle
FLS	Forward Looking Sonar
SSS	Side Scan Sonar
SAS	Synthetic Aperture Sonar
MSE	Mean Squared Error
PSNR	Peak Signal to Noise Ration
SSIM	Structural SIMilarity
UCIQE	Underwater Color Image Quality Evaluation
UIQM	Underwater Image Quality Measurement
MDNet	Multi-Domain Network
MDNet-UW	MDNet Under Water
IoU	Intersection over Union
VWIE	Variance Weighted Information Entropy
MSER	Maximally Stable Extremal Regions

CHAPTER 1

INTRODUCTION

Over the past few decades, underwater image processing has attracted significant amount of research. One of the reasons could be the fast growth of underwater robotics which finds application in maritime security, autonomous underwater vehicles (AUVs), ocean archaeology, inspection of structures etc. The imaging systems have a crucial role to play in enhancing the underwater surveillance capabilities. The underwater environment attenuates the electromagnetic radiation and limits its usable range. Hence there are not many types of sensors available for underwater imaging. One of the sensors that provides high resolution images in underwater environment is the underwater optical cameras. Optical images richly captures great details of the underwater scene and also provides a good visualisation. Sound is known to travel longer ranges than the electromagnetic radiations. Sonar, also known as acoustic camera, is an equipment that uses sound energy for gathering information and is probably the most used sensor for wide range of underwater applications. In order to have a better understanding of the objects and structures underwater, we need to develop algorithms that can process and extract useful information from both optical and acoustic images.

1.1 Motivation

Underwater imaging systems can provide a lot of information about various objects and structures present underwater which is crucial for various civilian as well as military purposes. Imaging systems can be used for a wide range of applications like study and inspection of underwater structures, exploration of the ocean as well as for enhancing surveillance capabilities by detection of unauthorized intrusions either from a diver or an underwater robot, identification of mine like objects, providing assistance in navigation of Autonomous Underwater Vehicle (AUV), obstacle avoidance etc.

The Two modalities of Underwater Imaging Systems include optical as well as acoustic imaging systems. Optical cameras are used in clear waters at shorter ranges. Whereas it mostly fails as the turbidity of water increases. On the other hand, acoustic cameras or Sonars can be used even in turbid water and provide imaging at longer ranges. Both optical as well as acoustic images suffer from certain drawbacks which needs to be mitigated by the application of various image processing techniques. The optical underwater images are inherently degraded due to haze, poor illumination, scattering etc which limits its applicability in underwater vision based tasks.Hence it needs to be enhanced prior to processing. Whereas Sonar images have low resolution and poor visualization. Hence it needs to be processed differently from their optical counterparts in order to extract useful information from it.

1.2 Background

1.2.1 Underwater Imaging systems: Optical Camera Vs Imaging Sonar

Even though underwater optical cameras provide higher resolution and better visualization, it mostly fails to provide good vision in turbid waters due to scattering of the light. Hence it cannot be relied on for long range applications. The Imaging Sonar is a category of Sonars that uses sound energy for generating useful two-dimensional images of underwater objects and have been used as a replacement for underwater optical cameras. Imaging Sonars are active ranging devices that produces images by using the echoes of sound energy reflected from objects. Imaging sonars can operate at higher ranges than their optical counterparts and are particularly useful in enhancing the underwater surveillance capabilities by early detection of objects of threat like divers, underwater robots, mines etc.

When it comes to image formation, optical cameras generate the elevation view of the underwater environment whereas Sonars generate the cross-section. Unlike optical cameras that uses x-y coordinate system for representation of 2D images, Sonars typ-

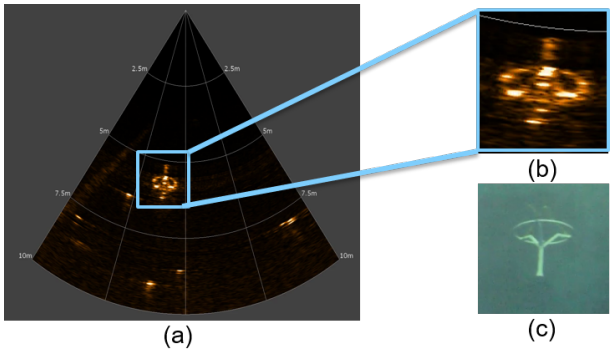

Fig. 1.1: : (a) Sonar image of a steering wheel (b) Zoomed in view of steering wheel (c) Corresponding optical image of steering wheel

ically use PPI (Plan Position Indicator) type of displays which is like a portion of the polar plots. This makes the visualization difficult in Sonar imagery.

1.2.2 Imaging Sonar: Working Principles

Imaging Sonars are active Sonars that transmits pulses of sound into water and receives the echoes returning from the object or scene to produce a two dimensional image of that object. The presence of objects in the scan area is marked by strong reflections (highlights) and the absence is marked by weak reflections or no reflections (shadows).

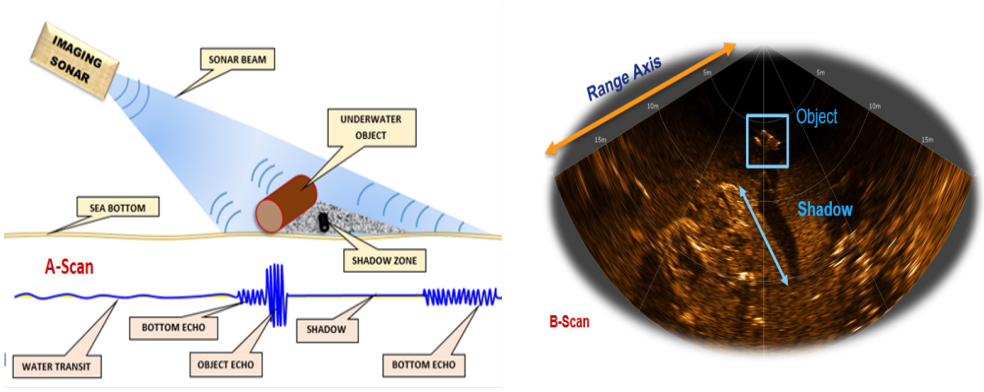

Fig. 1.2: (a) Imaging Sonar Concept (b) Typical Sonar Image

Imaging sonars are mainly of two types: (1) Side Scan Sonars (SSS) which produce high resolution images at long ranges (2) Forward Looking Sonars (FLS) which can

produce more details at shorter ranges.

Imaging Sonars forms fan like acoustic beams which are narrow in the horizontal plane and broader in the vertical plane. Multiple such beams are formed simultaneously in horizontal plane to get the cross-sectional view of the environment which is displayed as an image in the PPI format.

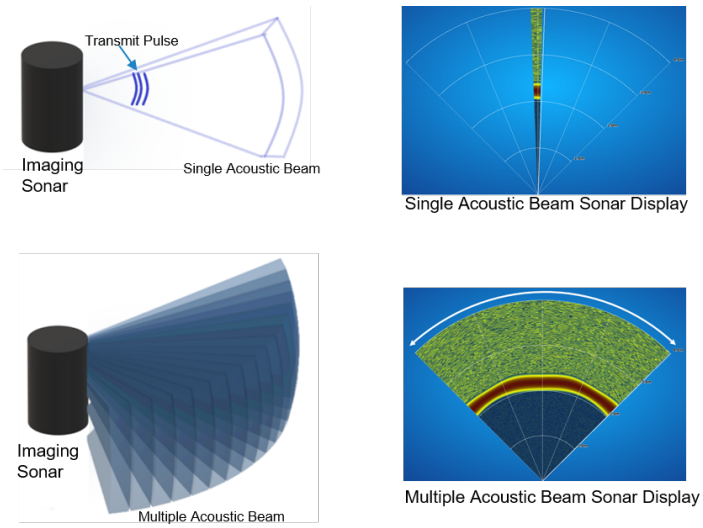

Fig. 1.3: Sonar Scanning Beams and Corresponding Sonar Images (https://bluerobotics.com/)

1.2.3 Underwater Imaging Experiments

Experiments were conducted in a water testbed which has a depth of about 18m for capturing both sonar and optical datasets. Different solid bodies such as cylinder, cube, steering wheel etc were used as targets for imaging. The objects were suspended in water testbed at a depth of 5-10m. A ROV equipped with an optical camera as well an imaging sonar was made to move around and capture videos of the submerged target objects.

The imaging sonar used in the experiments is Blueprint Subsea Oculus M750d[1] which is a general purpose dual-frequency sonar offering 120m imaging range capability at 750 kHz and 40m imaging range at 1.2MHz. It has field of view of 130 deg

[1] https://www.blueprintsubsea.com/pages/product.php?PN=BP01032

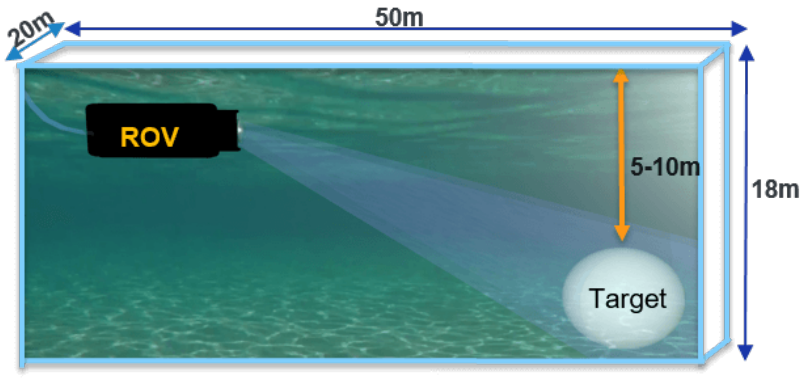

Fig. 1.4: Conceptual diagram : Data collection using ROV

and vertical beamwidth of 12 deg for the operating frequency of 750 kHz. It is deal for AUV navigation and imagery for near field target identification.

1.2.4 Structure from Motion (SfM) : Open Source packages

Structure from Motion (SfM) is a method for 3D reconstruction of objects from 2D images. In Multi-View SfM, a large number of images with overlapping contents, are used to estimate the 3D point cloud of the scene.The open source photogrammetric pipelines available for SfM are reliable and have the capability to process large number of unordered images. The input to these packages is an image sequence and camera intrinsic parameters whereas the output is sparse/dense point cloud or a dense textured mesh depending on the requirement. Some of the SfM packages include :

- OpenMVG (Moulon *et al.* (2016)) combined with OpenMVS ;
- COLMAP pipeline (Schönberger and Frahm (2016)).

(a) Open-MVG + Open –MVS

OpenMVG provides a SfM pipeline based on multiple view geometry principles. Feature descriptors are extracted using SIFT (Lowe (2004)) and AKAZE (Alcantarilla *et al.* (2013)). Feature matching is done by using methods like ANN-kD trees (Muja and Lowe (2009)), or cascade hashing (Cheng *et al.* (2014)). Sparse reconstruction is imple-

mented based on incremental (Moisan *et al.* (2012)) or global (Moulon *et al.* (2013)). Bundle adjustment is done using Ceres solver for refining the estimates. The dense reconstruction is implemented by the OpenMVS library based on patch-based stereo method . (Shen (2013)).

(b) COLMAP

COLMAP (Schönberger and Frahm (2016)) is a pipeline that implements both Structure from Motion (SfM) and Multi-View Stereo (MVS).A graphical user interface is included along with the package. Regarding finding feature correspondences, it implements the SIFT algorithm (Lowe (2004)) , followed feature matching options such as exhaustive matching, sequential matching, vocabulary tree, spatial matching, transitive matching and custom matching. Image pairs are considered registered if a valid mapping of their geometric relation (homography, essential or fundamental matrix) exists between the two and thus the scene graph is created gradually. 3D reconstruction is performed by using incremental SfM starting from a carefully selected initial image pair and applying a robust next best view selection algorithm and subsequently multi-view triangulation. The bundle adjustment uses Ceres solver and global BA every certain steps to improve camera and point estimations and avoid drifting. Multi-view stereo reconstruction is implemented based on the framework of (Zheng *et al.* (2014)) using a probabilistic patch-based stereo approach

1.2.5 Convolutional Neural Networks (CNNs/ConvNets)

Convolutional Neural Networks (CNNs/ ConvNets) have made breakthrough performances in the field of computer vision and it has been successfully deployed for recognition, classification and tracking of everyday objects, faces, vehicles etc in images. CNN takes images as inputs, and passes it through various convolutional layers, pooling layers and fully connected layers thus learning the weights and biases required for accomplishing the assigned task. Convolution layers act as a feature extractor whereas the fully connected layers classify the image based on the extracted features. The various

layers of CNNs act as relevant filters and learn the spatial and temporal dependencies of the images for extracting its salient features. In this work the application of CNNs are extended to two category of problems:

(a) Enhancement of underwater optical images for 3D reconstruction using SfM

(b) Tracking underwater objects in sonar images.

(a) Image Enhancement and 3D reconstruction

Most of the approaches for 3D reconstruction aim to recover the structure, shape and appearance of real 3D objects from stereo vision, motion and monocular cues like texture, defocus, shadow etc. Among them Structure from motion techniques are particularly suited for underwater tasks where the optical cameras can be fitted on AUVs/ROVs and deployed for capturing the details of the scene. SfM algorithm takes in images of multiple views and performs a feature matching across images to register them to a common view and then triangulate to find the depth of the points.

For Sfm algorithms to work, the images should be of good resolution. But in underwater scenario, the image quality is not good hence it is required to enhance the images before performing SfM. A CNN model based on Water-Net (Li *et al.* (2020*a*)) is implemented which first generates outputs by passing the input image through 3 algorithms i.e White Balance (WB), Histogram Equalization (HE) and Gamma Correction (GC) algorithms and then fuses it to generate the final enhanced output.

Structure from motion (SfM) algorithms and dense image matching Multi-View Stereo (MVS) algorithms have achieved remarkable success in 3D reconstruction of land-based objects, buildings and scenes from a sequence of images taken at different viewpoints .The Photo Tourism project (Agarwal *et al.* (2009)) investigated the problem of taking extremely large number of unstructured collections of photographs from internet and computing 3D model of the scene to enable browsing of the photo collection in 3D. Figure 1.5 shows the example of 3D reconstruction of the famous Colosseum in Rome which has been reconstructed from a huge collection of photographs downloaded from Flickr.

Fig. 1.5: 3D reconstruction of Colosseum, Rome (Agarwal *et al.* (2009))

(b) Tracking Underwater Objects in Sonar Image

In tracking, the goal is to find position of the object of interest in all subsequent frames given that Ground Truth is available on the first frame. The tracked target is marked by a bounding box in every frame. Sonar images are quite different from their optical counterparts in terms of data perception.

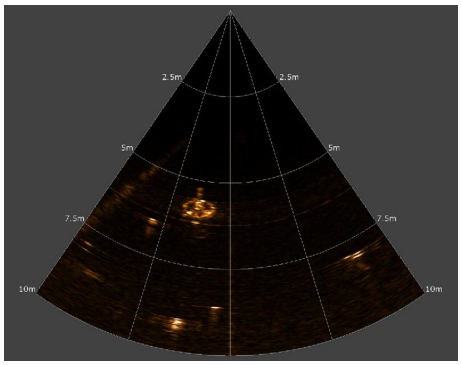

Fig. 1.6: Sonar Image for tracking

Sonar displays the reflected acoustic energy from the objects in the form of highlights and shadows which makes the object visualization very difficult even for humans. Extracting useful information from the sonar images is a challenging task because of its inherent imperfections like low resolution, lack of color information and foreground objects becoming indistinguishable from background clutter. The various object tracking algorithms that work well for visual objects, may not give expected results with sonar images especially in the scenarios where the underwater objects frequently move out of the view.

The objective here is to find a CNN model which will be well suited to the task of

tracking objects in Sonar images. The proposed method MDNet-UW (Multi-Domain Net Under Water) consists of a small CNN model for object tracking which is based on MDNet architecture Nam and Han (2016). The MDNet-UW training is done in two phases: (1) offline pre-training using images from a self-curated Sonar dataset (2) online training based on the initial ground truth. The model is trained to discriminate between the target and the background samples and to predict the most probable target candidate in each frame. A method based on Maximally Stable Extremal Regions (MSER) by Matas *et al.* (2004)and Variance Weighted Information Entropy by (VWIE) Wang and Chen (2017) is proposed for generating candidate regions while the track failure occurs. This helps to reduce the actual area and helps to improve speed and accuracy.

1.2.6 Image Entropy

Image entropy is a statistical measure which represents the average image gray level distribution and can be used as a metric for finding the similarity between acoustic images. For an image with m gray levels, the entropy at a coordinate (x,y) is given by :

$$\mathrm{E}(x, y) = -\sum_{i=1}^{m} p_i * \log p_i$$

p_i is the probability of i^{th} gray level such that $\sum_{i=1}^{m} p_i = 1$. When $p_i = 0$, then it is stipulated that $p_i * \log p_i = 0$. Smaller the entropy, the more uniform will be the gray level distribution and difference in gray level distribution within that image is less. Whereas larger the entropy, the difference in the gray level distribution is more prominent and it can be indication of presence of target objects. However image entropy only reflects the gray level distribution of image and does not reflect the complexity of the image background as it ignores the importance of gray levels.

CHAPTER 2

UNDERWATER IMAGE ENHANCEMENT

Clear underwater images can provide crucial information about the underwater world. It indicates the presence of different natural and man-made objects thus strengthening the underwater surveillance and underwater archaeology capabilities. The images of objects captured by underwater cameras is faced with different set of challenges such as poor visibility due to haze, non-uniform illumination, color cast etc. The underwater medium constraints the visibility of the scene due to scattering and absorption effects. This causes attenuation of the irradiance reaching the imaging device thus creating a hazy effect in the image. The quality of the images further deteriorates because of non-uniform illumination. The attenuation in underwater environment is also wavelength dependent thus introducing undesirable color cast along with haze. The red color light, having the longest wavelength gets attenuated first followed by green light and the blue light. Due to this frequency selective nature of the attenuation, the underwater images mostly have greenish or bluish tint which seriously affects the visual quality of the images.

2.1 Related Works

Some of the earlier works were based on modifying the pixel intensities to improve the quality. Fusion of different processed images have shown good results. Ancuti *et al.* (2012) proposed a method for blending color correct and contrast enhanced image using a multiscale fusion technique. Ghani and Isa (2014) used contrast stretching based on rayleigh distribution and color correction methods for enhancement. A novel Retinex based method was proposed by Fu *et al.* (2014) which consists of a three stage process which includes color correction, variation framework for decomposing reflectance and illumination and enhancement.

A model based approach for learning the latent parameters of the image formation was adopted by Dark channel Prior (Zhang *et al.* (2017)). It introduces human visual attention mechanism for removing haze from images. Generalization of the Dark Channel Prior (GDCP) was proposed by Peng *et al.* (2018) which estimates the ambient light by depth dependent colour differential and performs adaptive color correction.The local proximity based method by Mandal and Rajagopalan (2020) uses patch similarity assessment in the outdoor images to arrive at a transmission depthmap and a non local means filtering for removing haziness.

Many CNNs and Generative Adversarial Networks(GAN) based approaches are also being deployed.A pixel-to-pixel (P2P) network was proposed by Xin Sun (2019) to design an encoding–decoding framework for enhance underwater images.This model is similar to REDNet proposed by Peng *et al.* (2018).A UWCNN network was introduced by Saeed Anwar (2019) which is an end-to-end model containing three densely connected building blocks and is trained by the synthetic underwater image datasets.To enhance the underwater images,Guo *et al.* (2020) introduced a multiscale dense block (MSDB) algorithm, namely, DenseGAN1 which employs the use of dense connections, residual learning, and multi-scale network for underwater image enhancement.

2.2 Underwater Image Formation Models

The underwater image formation model is very complex due to the frequency selective attenuation of light underwater. A simplified underwater image formation model (Chiang and Chen (2012)) is used in most of the cases, which is similar to atmospheric scattering except that it captures the wavelength dependent attenuation of light underwater. It is suitable for scenarios such as shallow waters where there is lesser backscattering of light . The simplified model is expressed below:

$$\mathrm{U}_\lambda(x) = \mathrm{I}_\lambda(x).\mathrm{T}_\lambda(x) + \mathrm{B}_\lambda(x).(1 - \mathrm{T}_\lambda(x))$$

where $\mathrm{U}_\lambda(x)$ is the captured underwater image ; $\mathrm{I}_\lambda(x)$ is clear latent image or scene radiance; $\mathrm{B}_\lambda(x)$ is homogeneous global background light; λ is the wavelength of RGB

light; x is the location of a scene point; $\mathrm{T}_\lambda(x)$ is the medium energy ratio which is the percentage of scene radiance reflected from a point x given by

$$\mathrm{T}_\lambda(x) = 10^{-\beta_\lambda d(x)} = \frac{\mathrm{E}_\lambda(x,d(x))}{\mathrm{E}_\lambda(0,d(x))} = \mathrm{N}_\lambda(d(x))$$

where β_λ is wavelength dependent medium attenuation coefficient; $\mathrm{E}_\lambda(0, d(x))$ is energy of light from submerged scene before it passes through the transmission medium from a distance d(x); $\mathrm{E}_\lambda(x, d(x))$ is strength of the light after absorption bt medium; $\mathrm{N}_\lambda(d(x))$ is the normalised energy residual which is the ratio of residual energy to initial energy per unit of distance and is dependent on wavelength of light.

A revised model was proposed by Akkaynak and Treibitz (2018) which takes into consideration the key factors such as dependency of attenuation coefficient on veiling light, different attenuation coefficients for direct and backscattered light, light absorption etc. The image formation model can be expressed as :

$$\mathrm{U}_\lambda(x) = \mathrm{I}_\lambda(x).\exp -\beta_\lambda^D.(\mathrm{v}_D).z + \mathrm{B}_\lambda^\infty(1 - \exp -\beta_\lambda^B.(\mathrm{v}_B).z).$$

where $\mathrm{U}_\lambda(x)$ is the captured underwater image ; $\mathrm{I}_\lambda(x)$ is clear latent image ; $\mathrm{B}_\lambda^\infty$ is the veiling light; β_λ is the beam attenuation coefficient, D is the direct transmitted light and B is the backscattered light. the vectors v_D and v_B represent coefficient dependencies. $\mathrm{v}_{d(x)} = \{z, \rho, \mathrm{E}, \mathrm{S}_\lambda, \beta\}$ and $\mathrm{v}_{d(x)} = \{\mathrm{E}, \mathrm{S}_\lambda, b, \beta\}$ where z is range along LOC; ρ is reflectance; E is the irradiance ;S_λ is the sensor spectral response and b is the beam scattering coeff. More details can be found in Akkaynak and Treibitz (2018).

2.3 Underwater Image Datasets

One of the issues with applying deep learning techniques to the underwater image enhancement is the non-availability of a large-scale underwater images with references from the real world.

Real world Underwater Datasets

TURBID[1] consists of degraded images with their corresponding groundtruth images. TURBID has different categories of images which includes haze introduced by milk, chlorophyll, deepblue etc each with 20-50 images.

SQUID (Stereo Quantitative Underwater Image Dataset) : [2] This dataset consists of 57 stereo pairs of underwater images from four different ocean locations containing varying water properties and color charts shown in the scenes.

ULFID[3] : Underwater Light Field Image Dataset by Skinner and Johnson-Roberson (2017) contains several underwater light field images in pure water and hazy conditions, as well as images taken in the air for reference.

UIEB [4](Li *et al.* (2020*a*)) consists of 890 underwater images along with reference images and a challenging set of 60 images without reference. It is practically difficult to obtain simultaneously the underwater image of a deep ocean scene and the ground truth image of the same scene. In UIEB , the reference images are generated by fusing together the output of 12 different image enhancement techniques thus extracting the best output than what any single method would provide.

Synthetic Underwater Datasets

WaterGAN is a deep learning approach to compensate the non-availability of images by generating synthetic underwater images using the in-air images and corresponding depthmaps. These images can be used to train deep learning networks for underwater image restoration tasks. But most of the times the synthetic images fail to capture the real world scenario.

In this work,a CNN model is trained on benchmark dataset UIEB for image enhancement.

[1] http://amandaduarte.com.br/turbid/

[2] http://csms.haifa.ac.il/profiles/tTreibitz/datasets/ambient_forwardlooking/index.html

[3] https://github.com/kskin/data

[4] https://li-chongyi.github.io/proj_benchmark.html

2.4 CNN Model for Underwater Image Enhancement: WaterNet

A CNN model based on the Water-Net (Li *et al.* (2020*a*)) is implemented for image enhancement. Water-Net is trained on Underwater Image Enhancement Benchmark (UIEB) dataset (Li *et al.* (2020*a*)). The underwater environment is very complex with different water types, poor lighting and dynamic behavior. There doesnot exist a single image enhancement method that works against all adverse effects. Fusion based methods (Ancuti *et al.* (2012)) generally give decent results where inputs derived from various enhancement methods are fused together in a ratio to get desired results. The Water-net is also based on the fusion of various inputs and multiplying those input images with the learnt weights to obtain an enhanced output. The Water-Net takes underwater images as inputs and generates 3 images based on White Balance (WB), Histogram Equalization (HE) and Gamma Correction (GC) algorithms. WB is used to restore the color deviation whereas HE is used to improve the contrast of the image and GC is used to improve the brightness of the image. During training Water-Net learns confidence maps for each of the 3 inputs. During runtime, generated input images are fused together after multiplying it with the confidence maps in order to obtain an enhanced result.

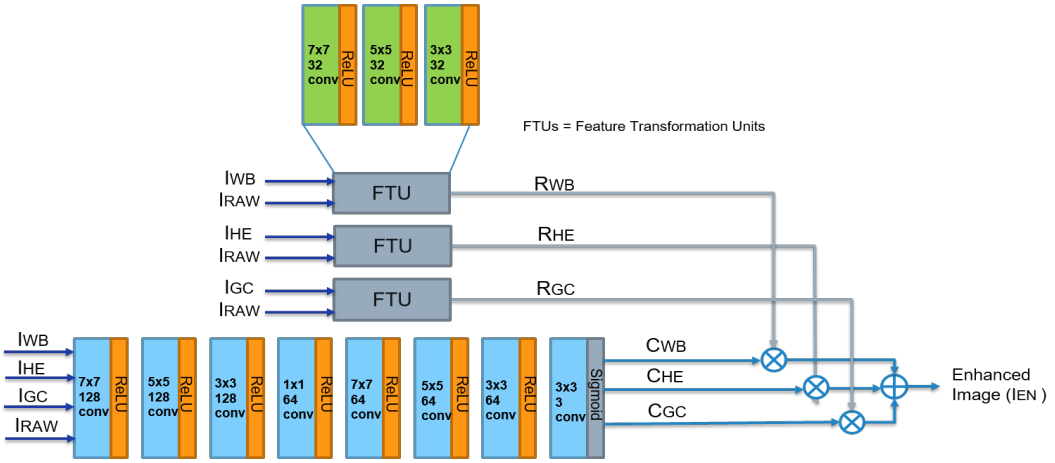

Fig. 2.1: Water-Net Architecture

$$I_{EN} = R_{WB} \odot C_{WB} + R_{HE} \odot C_{HE} + R_{GC} \odot C_{GC}$$

where I_{EN} is the enhanced result $\odot$ indicates the elementwise production of matrices; R_{WB}, R_{HE}, and R_{GC} are the refined results of input after processing by WB, HE, and GC algorithms, respectively; C_{WB}, C_{HE}, and C_{GC} are the learned confidence maps.

The 890 image pairs in UIEB were split into training and testing set. 800 image pairs were used for training and remaining 90 were used for testing. The input data was resized to 112 x 112 and data augmentation was done. This was implemented in tensorflow. Batch size of 16, learning rate of 0.001 and a momentum of 0.5 was used.

Perceptual loss is used which is based on the ReLU activation layers i.e layer relu5_4 of the pretrained VGG-19 network (Simonyan and Zisserman (2015)). Let ϕ_j denote the j^{ith} convolutional layer of the VGG network, then the perceptual loss is defined as the distance between the feature representations of reference image I_{gt} and enhanced image I_E. The expression is given by:

$$\mathrm{L}_j^{\phi} = \frac{1}{\mathrm{C_i H_i W_i}} \sum_{i=1}^{N} \left\| \phi_j(I_E^i) - \phi_j(I_{gt}^i) \right\|$$

where N is the number of each batch in the training.$\mathrm{C_i}, \mathrm{H_i}, \mathrm{W_i}$ are the number, height, and width of the feature map of the j th convolution layer within the VGG19 network. Water-Net was trained for 120 epochs with a dropout probability of 0.5. ADAM optimiser with default parameters was used for optimization

2.5 Underwater 3D reconstruction: SfM algorithm

3D reconstruction deals with estimation of the structure, shape and appearance of real objects from a sequence of 2D images or video streams. The process of imaging is to project 3D scene points from the world coordinates into 2D images on the camera plane where the process of 3D reconstruction is just the reverse of the imaging process. A typical SfM pipeline starts with image feature extraction, feature matching, incremental or global bundle adjustment and sparse 3D point cloud reconstruction. Whereas the Multi-View Stereo (MVS) pipeline consists of dense point cloud reconstruction, texturing, rendering etc. However computer vision related applications such as SfM require

high quality images so as to extract appropriate object features for performing feature matching between different views. Hence it is imperative to enhance the underwater images before performing the feature extraction.

2.6 Evaluation and Results

2.6.1 Image Enhancement Evaluation Metrics

There are two metrics:

(1) **Full-reference image quality evaluation metrics:** This is used for images with ground truth image available includes MSE, PSNR, and SSIM (Wang and Bovik (2002)).

(a) **Mean Squared Error (MSE) :** MSE provides a quantitative measure of similarity between two images. It is expressed as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

where N is the no of pixels; x_i and y_i are respectively the pixels at i^{th} location of the two images to be compared.

(b) **Peak Signal to Noise Ratio (PSNR) :** PSNR is a metric which is computed from MSE and is expressed as :

$$PSNR = 10\log_{10}\frac{L^2}{MSE}$$

where L is the range of image pixel intensities (L= 255 for image)

(c) **Structural SIMilarity (SSIM) :** Let x and y be patches taken from two different images but same locations to be compared against each other then SSIM takes 3 parameters into account i.e luminance $l(x, y)$, contrast $c(x, y)$ and local structures $s(x, y)$. SSIM is expressed as :

$$SSIM = l(x, y).c(x, y).s(x, y)$$

$$= (\frac{2\mu_x\mu_y+C1}{\mu x^2+\mu y^2+C1}).(\frac{2\sigma_x\sigma_y+C2}{\sigma_x^2+\sigma_y^2+C2}).(\frac{\sigma_{xy+C3}}{\sigma_x+\sigma_y+C3})$$

where μ_x μ_y are means and σ_x and σ_y are standard deviations of patches x and y respectively, σ_{xy} is cross correlation and constants $C1, C2, C3$ are to avoid non-zero division.

(2) **Non-reference underwater image quality metrics** include UCIQE Yang and Sowmya (2015) and UIQM (Panetta *et al.* (2016)) .

(a) **Underwater Color Image Quality Evaluation (UCIQE) :** UCIQE score quantifies the level of degradation in underwater images due color cast, blurring and low contrast. It is a linear combination of chroma, saturation and contrast. UCIQE score is expressed as :

$$UCIQE = C1 \times \sigma_c + C2 \times con_l + C3 \times \mu_s$$

where σ_c is standard deviation of chroma; con_l is contrast of luminance; μ_s is mean of saturation; $C1, C2, C3$ are constants (Yang and Sowmya (2015))

(b) **Underwater Image Quality Measurement (UIQM) :** UIQM is computed based 3 measures : Image Colorfulness Measure (UICM), Sharpness Measure (UISM) and Contrast Measure (UIConM). It is expressed as :

$$UIQM = c1 \times UICM + c2 \times UISM + c3 \times UIConM$$

where $c1, c2, c3$ are application dependent parameters.

2.6.2 Image Enhancement: Benchmark Results

A set of 12 images were selected from UIEB dataset for enhancement. The benchmark results of different image enhancement techniques are reported in table 2.1. It is to be noted that MSE the lower the better whereas PSNR,SSIM,UCIQE and UIQM, the higher the better. Platform for Underwater Image Quality Evaluation (PUIQE) by Li *et al.* (2020*b*) was used for online computation of UCIQE and UIQM scores.

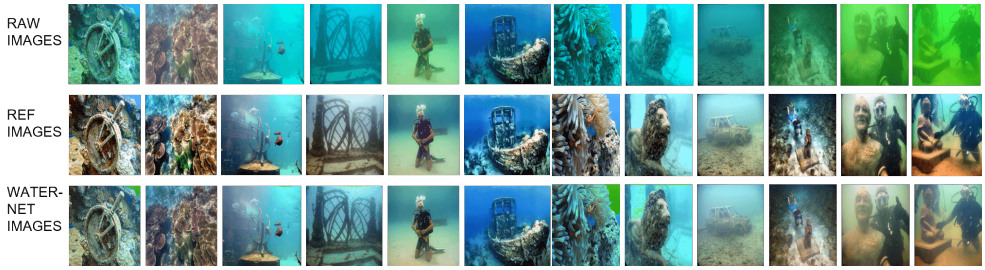

Fig. 2.2: Water-Net results from UBIE dataset

Fig. 2.3: (Left) Hazy Underwater image of Steering Wheel; (Right) Enhanced Underwater Image

...	Full Reference Metric			Non- Reference Metric	
Method	MSE ↓ ($\times 10^3$)	PSNR ↑	SSIM ↑	UCIQE ↑	UIQM ↑
Fusion Ancuti *et al.* (2012)	1.13	16.6	0.77	**0.64**	1.53
Retinex Fu *et al.* (2014)	1.35	16.87	0.62	0.60	1.43
UDCP Zhang *et al.* (2017)	5.13	11.02	0.50	0.59	**1.63**
GDCP Peng *et al.* (2018)	3.63	12.53	0.55	0.61	1.43
Local Proximity Mandal and Rajagopalan (2020)	2.52	14.26	0.48	0.58	0.71
Water-Net Li *et al.* (2020*a*)	**0.79**	**20.53**	**0.79**	0.57	0.57

Table 2.1: Image Quality Assessment Scores on UIEB test dataset. Blue color indicates the top scores in each metric

2.6.3 3D Reconstruction

The 3D reconstruction results using COLMAP are presented below. The Input to COLMAP has been a sequence of Images and the output is 3D point cloud as well

as the camera poses. Different cases include in-air images, underwater images without enhancement and with enhancement.

3D Reconstruction with Sceaux castle Dataset

- The reconstruction was done using 11 images.
- A point cloud of 9500 match points were generated

Fig. 2.4: Sceaux castle Dataset (Moulon *et al.* (2016))

3D Reconstruction with Hazy Underwater Images

- The reconstruction was done using 36 images.
- Underwater Steering Wheel data was used without enhancement.

Fig. 2.5: Hazy Underwater image

3D Reconstruction with Enhanced Underwater Images

- The reconstruction was done using 36 images.
- Underwater Steering Wheel data was used after enhancing it with Water-Net

Fig. 2.6: Enhanced Underwater Image

Performance Assessment			
Image Type	UCIQE	UIQM	No Of 3D Points generated
Raw	0.34	0.18	313
Enhanced	0.45	0.36	1689

Table 2.2: Performance Assessment of 3D reconstruction using Enhanced Images

2.7 Observations and Conclusion

Point clouds and camera poses of Underwater scene were obtained using 36 images. COLMAP was used for point cloud generation.Two cases were investigated which included images without enhancement and images enhanced with Water-Net. The observations are as follows:

(a) The image quality metrics computed on both raw and enhanced images as in table 2.2 clearly indicate that enhanced underwater images are better in quality than the raw ones.

(b) From the visual inspection, the point cloud generated is good and have close correspondence with the underwater scene.

(c) Further it is observed that the point cloud after enhancement with Water-Net is better and richer in information than the one obtained with raw image. More Points are registered with the 3D model after enhancement with Water-Net.

(d) A second object closer to the steering wheel is better reconstructed after enhancement than in the raw image.

CHAPTER 3

UNDERWATER OBJECT TRACKING IN SONAR IMAGES

Visual object tracking based on images is an active area of research. The basic idea of tracking is to find the position of the object in all frames of an image sequence given that its position in the initial ground truth is known. Generally the challenges faced by tracking algorithm includes partial or full occlusion, scale variations, rotations, out of view, low resolution, background clutters, illumination variation etc. A single tracker may not be able to overcome all tracking challenges simultaneously.

3.1 Related Works

There are not much works that attempt tracking objects in Sonar imagery. So much of the ideas are inspired from the visual object tracking algorithms.There are different approaches for solving the tracking problem. Most of the approaches rely on two things: (1) Motion Model (2) Appearance Model. A good tracker should have the ability to understand the motion of an object thus learning its dynamic behavior. With motion model, the tracking algorithm can predict the future positions of the target in the upcoming frames and thus reduce the area of search. Optical flow, Kalman filtering, Kanade-Lucas-Tomashi (KLT) feature tracker, mean shift tracking etc are some examples. The appearance models try to learn the features of objects based on its appearance. These algorithms can be classified into two types namely generative and discriminative. In generative, an object model is determined using the target regions and this model is used to find the target locations in upcoming frames based on any one of the minimum error criteria. In discriminative methods, the model is build using both the target and background samples.

MIL tracker (Babenko *et al.* (2009)) trains a discriminative model online to diffentiate between target and background samples.The Kernelized CFs (KCF) achieves high computational speed via kernel trick and circulant matrices (Henriques *et al.* (2015)). DCF-CSR (Lukezic *et al.* (2017)) imposes a spatial reliability constraint on correlation filters learning the target features and a channel reliability score for weighting per-channel filters. ECO (Danelljan *et al.* (2017)) introduces a factorized convolution operator to reduce the no of parameters characterising the target model. STRCF (Li *et al.* (2018)) uses a spatial regularisation and a temporal regularisation term on filter coefficients.

CNN trackers have shown good performance in object tracking than the conventional trackers. GOTURN (Held *et al.* (2016)) is an offline CNN based tracker which is trained before they are deployed. On the other hand, Multi-Domain Network (MDNet) (Nam and Han (2016)) is an online CNN tracker that uses a smaller CNN model to learn a generic representation of the target during runtime. TADT (Li *et al.* (2019)) learns target aware features for efficient tracking of targets with arbitrary forms.

3.2 Dataset preparation

One of the problems faced with training CNN on Sonar images for track application, is the lack of publicly available Sonar datasets with Ground Truth annotations. For tracking task, each frame has to be marked with a bounding box indicating the region of interest in that frame.

A new dataset was established by collecting Sonar images from experiments in the water testbed as mentioned in subsection 1.2.3 with different types of objects using blueprint subsea oculus 750d which is a multibeam forward looking sonar operating at dual frequency of 750 MHz and 1200 MHz. Also some images were collected from online sources [1]. Matlab Ground Truth Labeler app was used to annotate the objects of interest in each of the frame.

[1] `https://www.blueprintsubsea.com/oculus/`

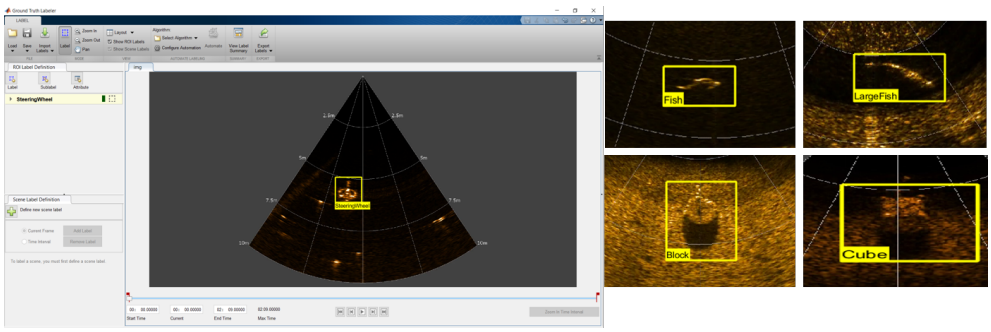

Fig. 3.1: (a)Matlab Ground Truth Labeler App (b) Different datasets

3.3 CNN model for Tracking in Sonar Images

Image Classification deep networks like VGG-Nets, GoogleNet etc have done remarkably well on the object classification task and can be tuned to Sonar images by transfer learning. Transfer learning approach offers two advantages (1) Can train networks whenever the training data available is less (2) It is faster than training a network from the scratch. There are different approaches for finding a CNN model suitable for tracking underwater objects

a) Transfer learning via feature extraction - For performing transfer learning by feature extraction, only a specified number of top layers of a pre-trained network are used and the bottom layers are removed. The top layers that have been retained in network act as an arbitrary feature extractor. The input image is propagated forward and output of the final layer of the truncated network is taken as features. Features of both target and background are extracted from the initial ground truth and a new sample can be identified as target or background based on the distance from the target space and background space. **b) Transfer learning via fine tuning** In fine-tuning, the network architecture is modified. The final fully connected layers of the network are removed and replaced with a newly initialized fully connected layer. Then the network is trained again to predict new input classes.

The disadvantage of using the above mentioned approaches is that it uses a deep network. The Networks like VGG, GoogleNet etc are trained to learn rich discriminatory features to perform tasks like 1000 class classification. A smaller CNN can perform

target/background discrimination. Secondly online training would require a lot of time as it uses deep networks. Thirdly these networks are pre-trained on ImageNet and may not give expected results with Sonar images.

MDNet(Nam and Han (2016)) is an architecture that uses a smaller CNN model to track objects. This architecture could be adapted to underwater object tracking task in sonar images. MDNet-UW (Multi-Domain Net Under Water) is the modified version of MDNet especially designed for the sonar imagery based tracking. The main objectives of this work has been :

(a) Training MDNet-UW on Sonar images for underwater object tracking.

(b) Handling Track failures by re-initialization of track using Maximally Stable Extremal Regions (MSER) features.

3.3.1 MDNet-UW Architecture

MDNet-UW (Multi-Domain Net Under Water) is adapted from MDNet architecture (Nam and Han (2016)). The architecture is shown in Figure 3.2. This CNN model consists of 6 layers: convolutional layers conv1-3 and fully connected layers fc4-6. MDNet is trained in a specific way such that it can learn a generic representation of the target and background.The MDNet separates the network into two parts: first part is the shared part consisting of convolutional layers which act as a feature extractor and is common to all domains. Then the second part consisting of the fully connected layers which is independent for each domain.

3.3.2 Training MDNet-UW on Sonar Images

The key task in developing tracking framework is to train MDNet-UW to discriminate between the target samples (positive samples) and background samples (negative samples). Here, the MDNet-UW tracker is first trained offline with self curated Sonar Image Dataset. More details about the dataset is mentioned in section 3.2. The training is done in two phases: (1) offline pre-training (2) online training based on the initial ground truth.

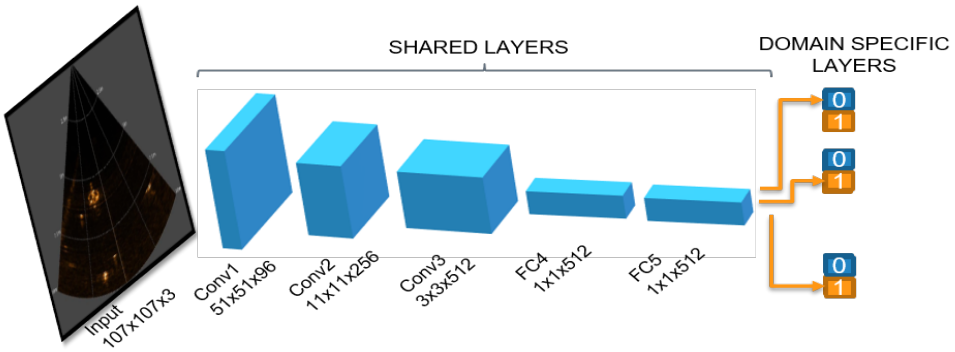

Fig. 3.2: Block diagram of MDNet architecture by Nam and Han (2016)

During offline training, each image sequence is considered separate domain and has a dedicated fully connected layer. K-domains have K-independent fully connected layers which classifies between target and background in that particular domain and the network is trained offline over K-domains iteratively. Each domain has image independent features and image dependent features. The idea behind offline training is to learn the image independent features before deployment and reuse it during runtime.

During the online training the network is made to learn the image dependent features. The bounding box is initialized in the first frame which provides the ground truth for training. Region around the bounding box is sampled to obtain the target and background samples. During online training the weights of the top layers are being learnt.

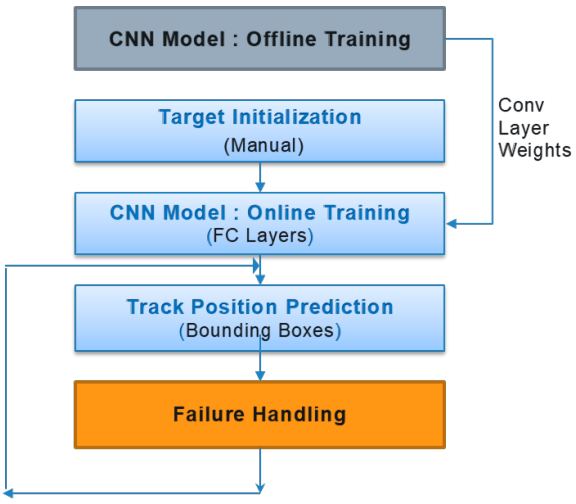

Fig. 3.3: MDNet-UW Tracking Flowchart

The weights of the convolutional layers remains fixed throughout the tracking process whereas the weights of the fully connected layers are update in regular time intervals.

3.3.3 MDNet-UW Tracking framework

The main components of the tracking framework consists of

(1) Offline Training: The CNN model is pre-trained on the Sonar dataset. The MDNet model weights are used for initialisation.

(2) Target initialization: A bounding box is drawn to mark the region of interest containing the object to be tracked. This acts as the ground truth for tracking algorithm.

(3) Online training: The purpose of online training is to learn appropriate weights for the single fully connected layer using the ground truth. The bottom layers are loaded with weights learnt during the offline training process. Both positive samples and negative samples are generated for training the network online . The samples whose intersection over union (IoU) overlap with the ground truth bounding box is greater than 0.7 are considered to be positive samples. Positive samples contain more information on the target. Whereas the samples whose intersection over union (IoU) overlap with the ground truth bounding box is lesser than 0.3 are considered to be negative samples and it represents the background. The network is trained for either a predefined number of iterations or till convergence. This is done only once in the beginning of tracking process.

(4) Target position prediction: During the tracking process, for each new image frame, regions are sampled around the previous known position of the target. These samples are passed to the MDNet which predicts the score for each of the two classes. The sample with highest score for target class is selected

(5) Failure Handling (By detecting Candidate Regions): The purpose is to provide guidance during track failure. Whenever a track failure occurs, the track algorithm gradually expands the area of search. But sometimes target may have moved out of this

search area and this slows down the target position prediction. So this proposed method scans the entire image for probable target regions. It uses Maximally Stable Extremal Regions (MSER) algorithm by Matas *et al.* (2004) and Variance Weighted Information Entropy (VWIE) as a metric for finding dissimilarity between target and background regions inspired by Yang *et al.* (2020). Once these regions are identified, it's sampled finely and forwarded to the MDNet which predicts the most probable target region as before. The advantage of this method is that it can converge to the true position faster.

3.3.4 Candidate Region sampling

The candidate regions are the most probable regions where a target can be present. In tracking process, handling of the track failures is also very important. Targets moving out of view, occlusions and highly cluttered background which are the main challenges faced by Sonar image based tracking.

To overcome these challenges, a method based on Maximally Stable Extremal Regions (MSER) and Variance Weighted Information Entropy (VWIE) is proposed to be included in the tracking framework inspired by Yang *et al.* (2020). The Candidate Regions from this algorithm are forwarded to the CNN model which predicts the scores for each of the regions. This helps to reduce the actual search area and to improve the accuracy and speed.

The MSER algorithm by Matas *et al.* (2004) is an area detection algorithm, which can detect connected pixels in an image. MSER are the regions that are having almost similar intensities and are either lighter or darker than their neighborhood. MSER regions are stable across a range of thresholds of intensity function and are mathematically expressed by Matas *et al.* (2004) as follows :

Image I is a mapping $I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow \mathcal{S}$. Extremal regions are well defined on images if:

(a) $\mathcal{S}$ is totally ordered, i.e. reflexive, antisymmetric and transitive binary relation $\leq$ exists. Here only $\mathcal{S} = \{0, 1, \ldots, 255\}$ is considered, but extremal regions can be defined on e.g. real-valued images ($\mathcal{S} = \mathcal{R}$).

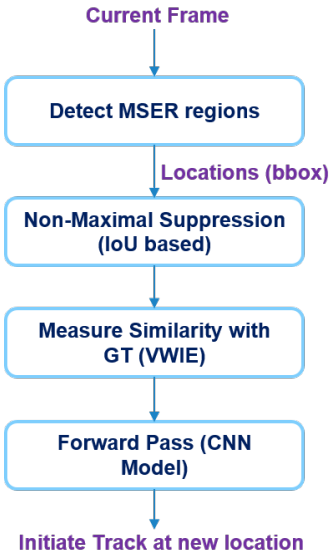

Fig. 3.4: Candidate region sampling flowchart

(b) An adjacency (neighbourhood) relation $A \subset \mathcal{D} \times \mathcal{D}$ is defined. Here 4-neighbourhoods are used, i.e. $p, q \in \mathcal{D}$ are adjacent (pAq) iff $\sum_{i=1}^{d} |p_i - q_i| \leq 1$

Region $\mathcal{Q}$ is a contiguous subset of $\mathcal{D}$, i.e. for each $p, q \in \mathcal{Q}$ there is a sequence $p, a_1, a_2, \ldots, a_n, q$ and $pAa_1, a_iAa_{i+1}, a_nAq$.

(Outer) Region Boundary $\partial\mathcal{Q} = \{q \in \mathcal{D} \setminus \mathcal{Q} : \exists p \in \mathcal{Q} : qAp\}$, i.e. the boundary $\partial\mathcal{Q}$ of $\mathcal{Q}$ is the set of pixels being adjacent to at least one pixel of $\mathcal{Q}$ but not belonging to $\mathcal{Q}$.

Extremal Region $\mathcal{Q} \subset \mathcal{D}$ is a region such that for all $p \in \mathcal{Q}, q \in \partial\mathcal{Q} : I(p) > I(q)$ (maximum intensity region) or $I(p) < I(q)$ (minimum intensity region).

Maximally Stable Extremal Region (MSER). Let $\mathcal{Q}_1, \ldots, \mathcal{Q}_{i-1}, \mathcal{Q}_i, \ldots$ be a sequence of nested extremal regions, i.e. $\mathcal{Q}_i \subset \mathcal{Q}_{i+1}$. Extremal region $\mathcal{Q}_{i^*}$ is maximally stable iff $q(i) = |\mathcal{Q}_{i+\Delta} \setminus \mathcal{Q}_{i-\Delta}| \,/\, |\mathcal{Q}_i|$ has a local minimum at i^* ($|\,.\,|$ denotes cardinality). $\Delta \in \mathcal{S}$ is a parameter of the method.

Variance Weighted Information Entropy (VWIE) :

To measure the complex degree of image, weighted information entropy is used. Consider a digital image I, with a local region containing m gray values $I_1, I_2, ...I_m$

and mean gray level value $\bar{I}$, then the VWIE at an image coordinate x,y is given by

$$\mathrm{E}(x,y) = -\sum_{i=1}^{m}(\mathrm{I}_i - \bar{\mathrm{I}})^2 * p_i * \log p_i$$

p_i is the probability of i^{th} gray level such that $\sum_{i=1}^{m} p_i = 1$. When $p_i = 0$, then it is stipulated that $p_i * \log p_i = 0$

The VWIE is more like a pixel by pixel comparison which is used to measure the dissimilarity between the target and the background sample regions. From the equation it can be seen that the pixels whose intensities are significantly different from their mean intensities get emphasized. VWIE is a simple and robust against speckles and heterogeneous regions and has been widely used in the SAR images for ship detection against complex backgrounds such as works done by Wang and Chen (2017) Lou *et al.* (2017). This method is independent of any prior knowledge of the target objects and backgrounds.

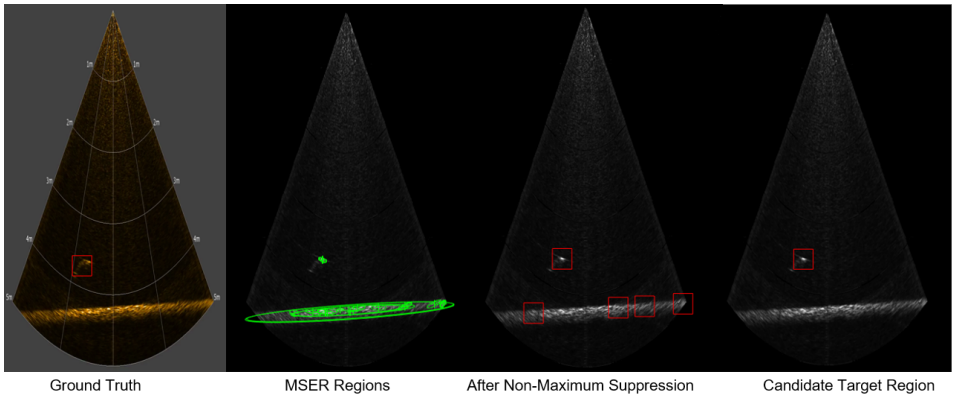

Fig. 3.5: Sonar Image with Ground Truth and MSER regions detected

Combining these two algorithms can well combine the advantages of pixel detection and area detection, and more completely and accurately detect the candidate target samples. The Candidate Regions should be repeatable and stable. The proposed techniques are integrated into the pipeline of Track.

3.3.5 Implementation

MDNet-UW is implemented in matlab and runs at a speed of 2.5 fps on a laptop withh intel i5 CPU(@1.6 GHz) with Nvidia gpu GEFORCE 250mx. For offline Training, 50 positive and 200 negative candidate regions are extracted from each frame. For target location update during runtime, 256 bounding boxes are drawn around the previous tracked location. The translation motion and scale affecting each box is sampled from a gaussian distribution. The CNN model acts like a binary classifier. Learning rate of 0.0001 is used.The momentum and weight decay of this network is always set to 0.9 and 0.0005, respectively

3.4 Evaluation And Results

3.4.1 Performance Metrics

(a) Intersection over Union (IoU) : IoU is a metric used to assess the accuracy of the predicted bounding box wrt Ground Truth box. The value ranges from 0 to 1 where 0 indicates no overlap with Ground Truth box and 1 indicates complete overlap. Generally a IoU score greater than 0.5 is considered a good overlap. It is defined as : $IoU = \frac{AreaofOverlap}{AreaofUnion}$. The bounding boxes are expressed as a 4-tuple (x,y,w,h) where (x,y) is the top leftmost coordinate of the box and (w,h) are width and height of the boxes respectively.

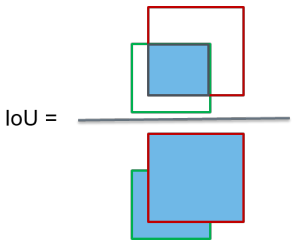

Fig. 3.6: Intersection over Union (IoU)

(b) Success Plots : A predicted bounding box is considered to be successful if the IoU is greater than the threshold. To generate a success plot, the threshold is varied from

0 to 1 and no of successfully tracked frames are counted.This is expressed as a ratio by dividing it with the total no of frames.Success Rate @ threshold =0.5 is expressed as $SuccessRate = \frac{FrameswithIoU\ 0.5}{TotalFrames}$

(c) Area Under Curve (AUC) : The area under the success plots is a metric used for ranking trackers

(d) Precision plots : A predicted bounding box is considered to be successful if the distance between the centres of predicted box and ground truth box is less than the threshold. The precision plots are obtained by plotting the percentage successful predictions against varying thresholds for centre location error.

(e) Centre Location error (CLE) : CLE refers to the avearge error in the location of centres between the predicted and ground truth boxes. The percentage successful predictions when the centre location error threshold is 20 pixels is considered as a good metric for ranking the trackers.

Further details can be found in Wu *et al.* (2013).

3.4.2 Results

(a) The success plots and precision plots of the trackers under evaluation for steering wheel dataset are shown in Figure 3.7. The performance metrics are tabulated in Table 3.1

(b) The success plots and precision plots of the trackers under evaluation for steering wheel dataset out-of-view case are are shown in Figure 3.8. The performance metrics are tabulated in Table 3.2. Here target moving out-of-view is simulated using the steering wheel dataset.

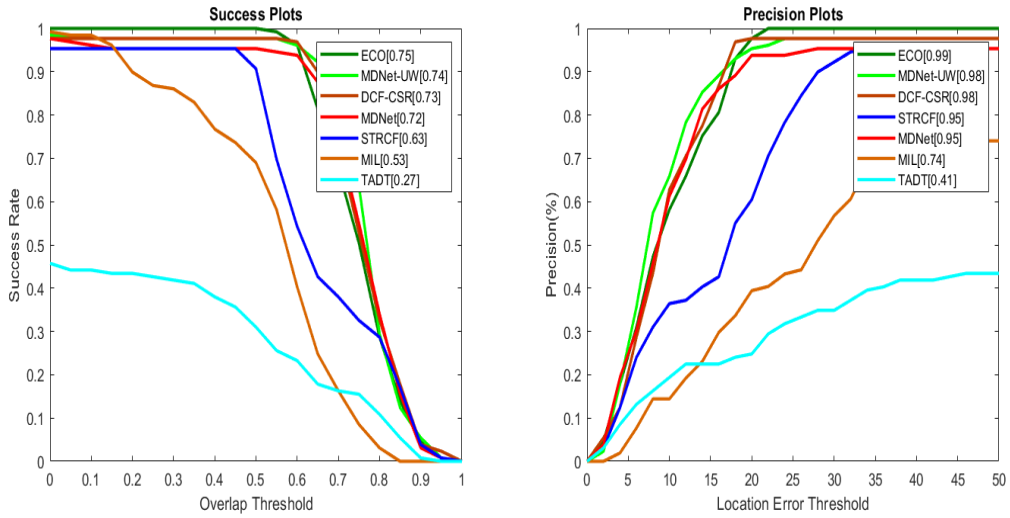

Fig. 3.7: Performance assessment of trackers on steering wheel dataset : success plot & precision Plot. The legend in the success plot shows AUC and for precision plot, it is the centre location error when threshold is 20 pixels

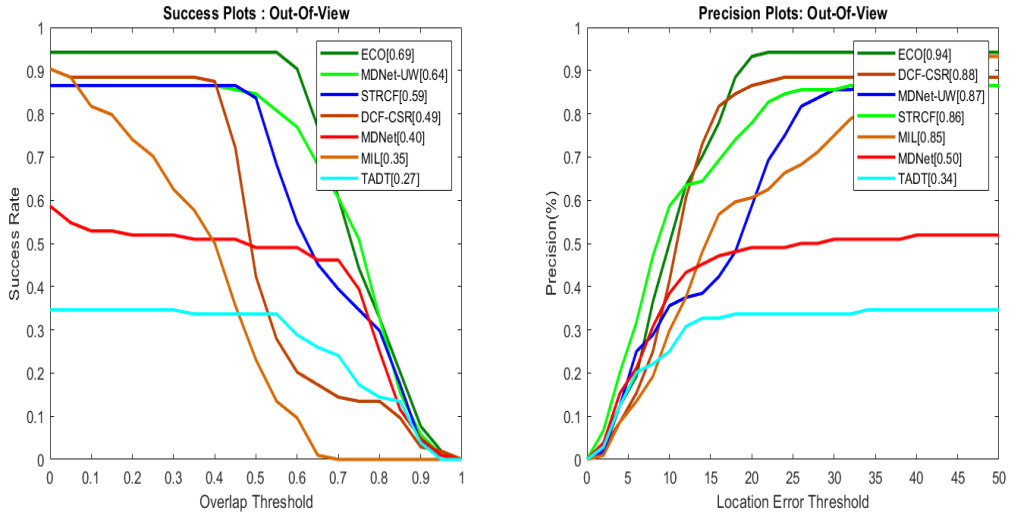

Fig. 3.8: Performance assessment of trackers on steering wheel dataset for Out-of-View case : success plot & precision Plot. The legend in the success plot shows AUC and for precision plot, it is the centre location error when threshold is 20 pixels

Tracker Performance Assessment					
Tracker Type .	AUC .	Success Rate @Thresh=0.5	CLE .	Precision @20pixels	FPS .
MIL Babenko *et al.* (2009)	0.30	0.53	24	0.57	4.2
DCF-CSR Lukezic *et al.* (2017)	0.73	0.976	12.16	0.976	4.7
STRCF Li *et al.* (2018)	0.63	0.90	22	0.6	9.79
ECO Danelljan *et al.* (2017)	0.75	0.99	9.34	0.976	0.85
TADT Li *et al.* (2019)	0.27	0.31	125.52	0.32	4.65
MDNet Nam and Han (2016)	0.72	0.95	16.6	0.92	1.5
MDNet-UW (Proposed)	0.74	0.98	13.10	0.96	1.26

Table 3.1: Performance Assessment of Trackers for Steering Wheel dataset.Blue indicates best AUC score and magenta indicates second best AUC score

Tracker Performance Assessment					
Tracker Type .	AUC .	Success Rate @Thresh=0.5	CLE .	Precision @20pixels	FPS .
MIL Babenko *et al.* (2009)	0.35	0.23	20	0.6	4.2
DCF-CSR Lukezic *et al.* (2017)	0.50	0.42	24.74	0.86	4.8
STRCF Li *et al.* (2018)	0.59	0.83	53	0.58	9.2
ECO Danelljan *et al.* (2017)	0.69	0.94	25.74	0.93	0.79
TADT Li *et al.* (2019)	0.25	0.33	172	0.33	4.8
MDNet Nam and Han (2016)	0.40	0.5	73	0.5	1.5
MDNet-UW (Proposed Method)	0.63	0.84	38	0.77	2.74

Table 3.2: Performance Assessment of Trackers for Steering Wheel dataset with simulation of Out-Of-View case. Blue indicates best AUC score and magenta indicates second best AUC score

3.5 Observations and Conclusions

The MDNet-UW trackers's performance was compared with conventional trackers like ECO Tracker (Danelljan *et al.* (2017)), DCF-CSR(Lukezic *et al.* (2017)), STRCF (Li *et al.* (2018)), MIL (Babenko *et al.* (2009)), MDNet (Nam and Han (2016)) and Deep Tracker TADT(Li *et al.* (2019)) for steering wheel dataset as well as for out-of-view simulation. The following are the observations:

(a) The MDNet-UW tracker improves the performance of MDNet tracker on the underwater Sonar dataset in all cases.

(b) The area under the curve (auc) is a criteria used for ranking the trackers. MDNet-UW is ranked second only to ECO trackers (Table 3.2).

(c) However in terms of frames processed per second (fps), it is better than ECO Tracker.

CHAPTER 4

CONCLUSION AND FUTURE WORK

Two research areas in the domain of underwater image processing were pursued in this work. The first area of the work was on enhancing underwater optical images for applications like 3D reconstruction of underwater structures. This work is mentioned in detail in chapter 2. The Underwater Image Enhancement Benchmark (UIEB) dataset was used for training Water-Net which is a CNN model for underwater image enhancement. It has been observed that with Water-Net, the image quality was improved and it manifested as higher value of the UCIQE and UIQM scores for output images that were used for reconstruction. Underwater optical images of a steering wheel submerged in water testbed was collected during experiments and the same was used for 3D reconstruction. Structures which were not not clearly seen due to haziness, were enhanced with Water-Net and this resulted in the denser reconstruction of 3D points of the scene.The No of reconstructed points were more when the images were used after enhancement. Thus there was a qualitative improvement in the 3D reconstruction after enhancement.

The Water-Net model is only a baseline model. Further GANs and Auto encoder based models can be developed for achieving better accuracy on the UIEB dataset. In the future, the image enhancement CNN models can be integrated along with the SfM pipelines. End-to-end Deep learning networks can be developed for image enhancement and 3D reconstruction on underwater images.

The second area of the work was on tracking of underwater objects in Sonar imagery This work is mentioned in detail in chapter 3. Due to non-availability of Sonar datasets with Ground Truths, Sonar images were collected during experiments using Blueprint Subsea Oculus Sonar 750d dual frequency multibeam forward looking Sonar. This data was annotated using Matlab ground Truth labeler. A CNN model MDNet-UW which is based on MDNet, was trained on Sonar data. It was observed that MDNet-UW performs well on Sonar images whereas other conventional trackers like MIL, KCF

etc do not perform as expected on the Sonar images. As an extension of this work a method based on MSER features was integrated along with track to handle tracking failures efficiently. This proposed scheme was able to predict target regions on Steering wheel dataset.

In this work single object tracking was attempted. As an advancement to this, multiple object tracking can be worked on. Instead of a binary classifier, the CNN should be trained for classifying multiple objects against the background. Motion models can also be included for multi-target tracking scenario. Track re-identification in a multi-target environment is also an interesting area to work on.

REFERENCES

1. **Agarwal, S.**, **N. Snavely**, **I. Simon**, **S. M. Sietz**, and **R. Szeliski** (2009). Building rome in a day. *In Twelfth IEEE International Conference on Computer Vision (ICCV 2009)*. IEEE. URL https://www.microsoft.com/en-us/research/publication/building-rome-in-a-day/.

2. **Akkaynak, D.** and **T. Treibitz** (2018). A revised underwater image formation model. *In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2018.00703.

3. **Alcantarilla, P.**, **J. Nuevo**, and **A. Bartoli** (2013). Fast explicit diffusion for accelerated features in nonlinear scale spaces. *In BMVC*.

4. **Ancuti, C.**, **C. O. Ancuti**, **T. Haber**, and **P. Bekaert** (2012). Enhancing underwater images and videos by fusion. *In 2012 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2012.6247661.

5. **Babenko, B.**, **M.-H. Yang**, and **S. Belongie** (2009). Visual tracking with online multiple instance learning. *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2009.5206737.

6. **Cheng, J.**, **C. Leng**, **J. Wu**, **H. Cui**, and **H. Lu** (2014). Fast and accurate image matching with cascade hashing for 3d reconstruction. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

7. **Chiang, J. Y.** and **Y.-C. Chen** (2012). Underwater image enhancement by wavelength compensation and dehazing. *IEEE Transactions on Image Processing*, **21**(4), 1756–1769, doi:10.1109/TIP.2011.2179666.

8. **Danelljan, M.**, **G. Bhat**, **F. S. Khan**, and **M. Felsberg** (2017). Eco: Efficient convolution operators for tracking. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR.2017.733.

9. **Fu, X.**, **P. Zhuang**, **Y. Huang**, **Y. Liao**, **X.-P. Zhang**, and **X. Ding** (2014). A retinex-based enhancing approach for single underwater image. *In 2014 IEEE International Conference on Image Processing (ICIP)*. doi:10.1109/ICIP.2014.7025927.

10. **Ghani, A.** and **N. Isa** (2014). Underwater image quality enhancement through composition of dual-intensity images and rayleigh-stretching. *IEEE Transactions on Image Processing*, **3**(1), doi:10.1186/2193-1801-3-757.

11. **Guo, Y.**, **H. Li**, and **P. Zhuang** (2020). Underwater image enhancement using a multiscale dense generative adversarial network. *IEEE Journal of Oceanic Engineering*, **45**(3), 862–870, doi:10.1109/JOE.2019.2911447.

12. **Held, D.**, **S. Thrun**, and **S. Savarese** (2016). Learning to track at 100 fps with deep regression networks. *In ECCV*.

13. **Henriques, J. F.**, **R. Caseiro**, **P. Martins**, and **J. Batista** (2015). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(3), 583–596, doi:10.1109/TPAMI.2014.2345390.

14. **Li, C.**, **C. Guo**, **W. Ren**, **R. Cong**, **J. Hou**, **S. Kwong**, and **D. Tao** (2020*a*). An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, **29**, 4376–4389, doi:10.1109/TIP.2019.2955241.

15. **Li, C. Y.**, **R. Mazzon**, and **A. Cavallaro** (2020*b*). Underwater image filtering: methods, datasets and evaluation.

16. **Li, F.**, **C. Tian**, **W. Zuo**, **L. Zhang**, and **M.-H. Yang** (2018). Learning spatial-temporal regularized correlation filters for visual tracking. *In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2018.00515.

17. **Li, X.**, **C. Ma**, **B. Wu**, **Z. He**, and **M.-H. Yang** (2019). Target-aware deep tracking. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR.2019.00146.

18. **Lou, J.**, **Z. Wei**, **H. Wang**, and **M. Ren** (2017). Small target detection combining regional stability and saliency in a color image. *Multimedia Tools and Applications*, **76**(13), doi:10.1007/s11042-016-4025-7.

19. **Lowe, D. G.** (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, **60**(2), 91–110. ISSN 0920-5691, doi:10.1023/B:VISI.0000029664.99615.94.

20. **Lukezic, A.**, **T. Vojír**, **L. C. Zajc**, **J. Matas**, and **M. Kristan** (2017). Discriminative correlation filter with channel and spatial reliability. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR.2017.515.

21. **Mandal, S.** and **A. N. Rajagopalan** (2020). Local proximity for enhanced visibility in haze. *IEEE Transactions on Image Processing*, **29**, 2478–2491, doi:10.1109/TIP.2019.2957931.

22. **Matas, J.**, **O. Chum**, **M. Urban**, and **T. Pajdla** (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, **22**(10), 761–767. ISSN 0262-8856, doi:https://doi.org/10.1016/j.imavis.2004.02.006. British Machine Vision Computing 2002.

23. **Moisan, L.**, **P. Moulon**, and **P. Monasse** (2012). Automatic homographic registration of a pair of images, with a contrario elimination of outliers. *Image Process. Line*, **2**, 56–73.

24. **Moulon, P.**, **P. Monasse**, and **R. Marlet** (2013). Global fusion of relative motions for robust, accurate and scalable structure from motion. *In ICCV*.

25. **Moulon, P.**, **P. Monasse**, **R. Perrot**, and **R. Marlet** (2016). Openmvg: Open multiple view geometry. *In International Workshop on Reproducible Research in Pattern Recognition*. Springer.

26. **Muja, M.** and **D. Lowe** (2009). Fast approximate nearest neighbors with automatic algorithm configuration. *In VISAPP*.

27. **Nam, H.** and **B. Han** (2016). Learning multi-domain convolutional neural networks for visual tracking. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

28. **Panetta, K.**, **C. Gao**, and **S. Agaian** (2016). Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, **41**(3), 541–551, doi:10.1109/JOE.2015.2469915.

29. **Peng, Y.-T.**, **K. Cao**, and **P. C. Cosman** (2018). Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing*, **27**(6), 2856–2868, doi:10.1109/TIP.2018.2813092.

30. **Saeed Anwar, F. P., Chongyi Li** (2019). Deep underwater image enhancement. *arXiv preprint arXiv:1807.03528 (2018)*.

31. **Schönberger, J. L.** and **J.-M. Frahm** (2016). Structure-from-Motion Revisited. *In Conference on Computer Vision and Pattern Recognition (CVPR)*.

32. **Shen, S.** (2013). Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *IEEE Transactions on Image Processing*, **22**(5), 1901–1914, doi:10.1109/TIP.2013.2237921.

33. **Simonyan, K.** and **A. Zisserman** (2015). Very deep convolutional networks for large-scale image recognition. *CoRR*, **abs/1409.1556**.

34. **Skinner, K. A.** and **M. Johnson-Roberson** (2017). Underwater image dehazing with a light field camera. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi:10.1109/CVPRW.2017.224.

35. **Wang, X.** and **C. Chen** (2017). Ship detection for complex background sar images based on a multiscale variance weighted image entropy method. *IEEE Geoscience and Remote Sensing Letters*, **14**(2), 184–187, doi:10.1109/LGRS.2016.2633548.

36. **Wang, Z.** and **A. Bovik** (2002). A universal image quality index. *IEEE Signal Processing Letters*, **9**(3), 81–84, doi:10.1109/97.995823.

37. **Wu, Y.**, **J. Lim**, and **M.-H. Yang** (2013). Online object tracking: A benchmark. *In 2013 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2013.312.

38. **Xin Sun, Q. L. J. D. E. L. R. Y., Lipeng Liu** (2019). Deep pixel-to-pixel network for underwater image enhancement and restoration. *IEEE Transactions on Image Processing*, doi:https://doi.org/10.1049/iet-ipr.2018.5237.

39. **Yang, L.**, **T. Xie**, **J. Yu**, **L. Zhai**, and **Q. Tian** (2020). Data processing based on variance weighted information entropy and maximally stable extremal regions. *In 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*. doi:10.1109/ICCCBDA49378.2020.9095615.

40. **Yang, M.** and **A. Sowmya** (2015). An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, **24**(12), 6062–6071, doi:10.1109/TIP.2015.2491020.

41. **Zhang, L.**, **X. Wang**, and **C. She** (2017). Single image haze removal based on saliency detection and dark channel prior. *In 2017 IEEE International Conference on Image Processing (ICIP)*. doi:10.1109/ICIP.2017.8297092.

42. **Zheng, E.**, **E. Dunn**, **V. Jojic**, and **J.-M. Frahm** (2014). Patchmatch based joint view selection and depthmap estimation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.