

TRACKING OF MULTIPLE MOVING FISHES IN MEDIUM AND HIGH DENSITY
UNDERWATER SCENARIO

SHILPI GUPTA



BHARTI SCHOOL OF TELECOMMUNICATION TECHNOLOGY & MANAGEMENT

INDIAN INSTITUTE OF TECHNOLOGY DELHI

SEPTEMBER, 2024

© Indian Institute of Technology Delhi (IITD), New Delhi, 2024

TRACKING OF MULTIPLE MOVING FISHES IN MEDIUM AND HIGH DENSITY
UNDERWATER SCENARIO

by

SHILPI GUPTA

BHARTI SCHOOL OF TELECOMMUNICATION TECHNOLOGY AND
MANAGEMENT

Submitted

In fulfilment of the requirements of the degree of

Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

SEPTEMBER, 2024

Certificate

This is to certify that the thesis titled Tracking of Multiple Moving Fishes in Medium and High Density Underwater Scenario, submitted by Shilpi Gupta, to the Indian Institute of Technology, Delhi, for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work done by her under our 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.

Professor Santanu Chaudhury
Dept. of Electrical Engineering
IIT-Delhi

Professor Brejesh Lall
Dept. of Electrical Engineering
IIT-Delhi

Acknowledgements

First and foremost, I would like to thank God, the Almighty, who has granted me countless blessings, so that I have been finally able to accomplish the thesis. I would like to express my sincere gratitude to Prof. Santanu Chaudhury, Director, IIT Jodhpur for giving me an opportunity to work in the area of underwater computer vision. I am extremely grateful to Prof. Brejesh Lall, IIT Delhi for his valuable guidance with understanding, continual encouragement,

and intellectual support during the completion of this thesis. There were moments when I gave up it was Brejesh Sir who motivated me. Sir always used to say “Life is too short to be sad”, this is going to be lifelong learning for me. I am fortunate to work under his supervision.

I would like to thank my committee members, Prof. Prem Kalra, Prof. Tapan Gandhi, and Prof. Ranjan Bose for their precious time and valuable feedback. I am thankful to Dr. Prerana Mukherjee, JNU for her unwavering efforts, helping me during the late night working hours and inspiring me to realize my goals. Without her support, this thesis would not have been possible.

My Special thanks are extended to Dr. Chetan Ralekar, Postdoctoral Research Associate at Massachusetts Institute of Technology for always supporting me when I needed help. I am thankful to my husband Mr. Praval Gupta for enriching my life and widening my horizons with his patience, care, companionship, empathy, and love. I feel indebted to my daughter Prashi for giving me the necessary distractions from work, bringing balance and richness to my doctoral experience.

My Ph.D. journey will be incomplete if I do not mention my friends Parul, Jaya, Shweta, Devanshu, Preeti, and Anand, who always stood by me through thick and thin. I would like to express my deepest gratitude to my parents and sister for making me what I am today. I can never thank them enough. Last but not least, I am thankful for the support received from all my labmates and lab staff.

Shilpi Gupta

Abstract

Due to remarkable progress in deep learning techniques, computer vision has gained increasing attention in research. The maximum work of this area is for on-ground scenarios which are either human-specific or vehicle specific. While water covers around 70 percent of the earth's surface and Fishes are an inseparable part of water. In addition to that, Fish is a major food source for humans, a living hood for many people belonging to the fishing industry, a contributor to the economy with fish farming, and an indicator of environmental change. Fish is of great importance for both the fish farming industry and marine researchers. Sensor-based techniques and underwater acoustic information is used for underwater research. these methods are invasive, expensive, and difficult to deploy. However, underwater cameras are non-invasive, cost-effective, and can record continuous data. In view of this, I investigate in this thesis, computer vision for underwater videos focusing on fish.

In the first approach, we propose a refined optical flow estimation method. Estimating motion in underwater videos is challenging due to low contrast, highly cluttered background, dynamic background, and illumination change. Central to our approach is exploiting contour information as most of the motion lies on the edges. Further, we have formulated it as sparse to dense motion estimation. The proposed method has been evaluated on challenging real-life image sequences of the Fish4Knowledge database. In the next approach, we solved a highly challenging task of multiple fish tracking as fish movements are highly erratic in nature. Trajectories of fishes convey critical information for the analysis of fish behaviour. We have proposed DFTNet which incorporates Siamese network for encoding the appearance similarity and attention-long short term memory network to capture the motion similarity across subsequent frames. Finally, intersection over union matching score is computed to amalgamate spatial similarity cue in the final score. The proposed framework provides a joint optimization score to maintain the tracklet information encoding appearance, motion, and spatial similarity cues. Motion detection and multiple fish tracking are mid-level computer vision tasks, which can provide the basis for other high-level computer vision tasks like activity, intention, and behavior recognition of fish. Fishes often tend to show collective behavior. They prefer to move into a group. Identifying fish groups and their profiling is a major problem in underwater computer vision areas. In the existing literature, fish grouping behavior has been explained in many ways, but there is a lack of explanations supported by empirical research using video analytics methods. Analysis and comparison of work in this area is made difficult in the absence of data sets with ground truth of fish groupings. To

bridge this gap, we have developed an Underwater Crowd Analysis Dataset UWCAD and group annotation tool GAT in chapter 5. We also propose a motion feature extraction method GGKLT and underwater temporal autoencoder for group detection UTAG by clustering the learned trajectory features by temporal autoencoder and global trajectory features.

सारांश

गहन शिक्षण तकनीकों में उल्लेखनीय प्रगति के कारण, कंप्यूटर विज्ञान ने अनुसंधान में अधिक ध्यान आकर्षित किया है। इस क्षेत्र का अधिकतम कार्य जमीनी परिदृश्यों पर है जो या तो मानव-विशिष्ट या वाहन-विशिष्ट हैं। जबकि पानी पृथ्वी की सतह का लगभग 70 प्रतिशत भाग कवर करता है और मछलियाँ पानी का एक अविभाज्य हिस्सा हैं। इसके अलावा, मछली मनुष्यों के लिए एक प्रमुख खाद्य स्रोत है, मछली पकड़ने के उद्योग से जुड़े कई लोगों के लिए आजीविका है, मछली पालन के साथ अर्थव्यवस्था में योगदानकर्ता है, और पर्यावरण परिवर्तन का एक संकेतक है। मछली पालन उद्योग और समुद्री शोधकर्ताओं दोनों के लिए मछली का बहुत महत्व है। पानी के नीचे अनुसंधान के लिए सेंसर-आधारित तकनीकों और पानी के नीचे ध्वनिक जानकारी का उपयोग किया जाता है। ये विधियाँ आक्रामक, महंगी और लागू करने में कठिन हैं। हालाँकि, पानी के नीचे के कैमरे गैर-आक्रामक, लागत प्रभावी हैं और निरंतर डेटा रिकॉर्ड कर सकते हैं। इसे ध्यान में रखते हुए, मैं इस थीसिस में मछली पर ध्यान केंद्रित करने वाले पानी के नीचे के वीडियो के लिए कंप्यूटर विज्ञान की जांच करता हूँ

पहले दृष्टिकोण में, हम एक परिष्कृत ऑप्टिकल प्रवाह अनुमान विधि का प्रस्ताव करते हैं। कम कंट्रास्ट, अत्यधिक अव्यवस्थित पृष्ठभूमि, गतिशील पृष्ठभूमि और रोशनी परिवर्तन के कारण पानी के नीचे के वीडियो में गति का अनुमान लगाना चुनौतीपूर्ण है। हमारे दृष्टिकोण का केंद्र अधिकांश गति के रूप में समोच्च जानकारी का दोहन करना है किनारों पर पड़ा है। इसके अलावा, हमने इसे विरल से सघन गति अनुमान के रूप में तैयार किया है। प्रस्तावित विधि का मूल्यांकन चुनौतीपूर्ण वास्तविक जीवन पर किया गया है

फिश4नॉलेज डेटाबेस के छवि अनुक्रम। अगले दृष्टिकोण में, हमने एकाधिक मछली ट्रैकिंग के अत्यधिक चुनौतीपूर्ण कार्य को हल किया क्योंकि मछली की गतिविधियाँ प्रकृति में अत्यधिक अनियमित होती हैं। मछलियों के प्रक्षेप पथ मछली के व्यवहार के विश्लेषण के लिए महत्वपूर्ण जानकारी देते हैं। हमने डीएफटीनेट का प्रस्ताव दिया है जिसमें उपस्थिति समानता को एन्कोड करने

के लिए सियामी नेटवर्क और बाद के फ्रेमों में गति समानता को पकड़ने के लिए ध्यान-दीर्घकालिक मेमोरी नेटवर्क शामिल है। अंत में, संघ मिलान स्कोर पर प्रतिच्छेदन की गणना अंतिम स्कोर में स्थानिक समानता क्यू को समाहित करने के लिए की जाती है। प्रस्तावित ढांचा ट्रैकलेट सूचना एन्कोडिंग उपस्थिति, गति और स्थानिक समानता संकेतों को बनाए रखने के लिए एक संयुक्त अनुकूलन स्कोर प्रदान करता है। मोशन डिटेक्शन और मल्टीपल फिश ट्रेकिंग मध्य-स्तरीय कंप्यूटर विज्ञान कार्य हैं, जो मछली की गतिविधि, इरादे और व्यवहार की पहचान जैसे अन्य उच्च-स्तरीय कंप्यूटर विज्ञान कार्यों के लिए आधार प्रदान कर सकते हैं।

मछलियाँ अक्सर सामूहिक व्यवहार दिखाती हैं। वे समूह में रहना पसंद करते हैं। पानी के भीतर कंप्यूटर दृष्टि क्षेत्रों में मछली समूहों की पहचान करना और उनकी रूपरेखा बनाना एक बड़ी समस्या है। मौजूदा साहित्य में, मछली समूहन व्यवहार को कई तरीकों से समझाया गया है, लेकिन वीडियो विश्लेषण विधियों का उपयोग करके अनुभवजन्य अनुसंधान द्वारा समर्थित स्पष्टीकरण की कमी है। मछली समूहों की जमीनी सच्चाई के साथ डेटा सेट के अभाव में इस क्षेत्र में काम का विश्लेषण और तुलना करना मुश्किल हो गया है। इस अंतर को पाटने के लिए, हमने अध्याय 5 में एक अंडरवाटर क्राउड एनालिसिस डेटासेट यूडब्ल्यूसीएडी और गुप एनोटेशन टूल जीएटी विकसित किया है। हम टेम्पोरलऑटोएनकोडर और वैश्विक प्रक्षेपवक्र सुविधाओं द्वारा सीखे गए प्रक्षेपवक्र सुविधाओं को क्लस्टर करके गुपडिटेक्शन यूटीएजी के लिए एक मोशन फीचरएक्सट्रैक्शन विधि जीजीकेएलटी और अंडरवाटर टेम्पोरल ऑटोएनकोडर का भी प्रस्ताव करते हैं।

Contents

Certificate	i
Acknowledgements	ii
Abstract	iii
List of Figures	x
List of Tables	xiii
1. INTRODUCTION	
1.1 Overview.....	1
1.2 Motivation.....	3
1.3 Scope.....	4
1.4 Challenges.....	6
1.5 Objectives.....	6
1.6 Thesis Layout.....	7
2. LITERATURE REVIEW	
2.1 Underwater Videos.....	12
2.2 Other Methods.....	15
3. MOTION ESTIMATION	
3.1 Introduction.....	18
3.2 Related Work.....	20
3.3 Proposed Methodology.....	22
3.3.1 Preprocessing.....	22
3.3.2 Contour Detection.....	23
3.3.3Energy minimization.....	25
3.4Experimental results and analysis.....	26

3.4.1 Dataset.....	26
3.4.2Qualitative Analysis.....	28
3.5 Summary.....	28
4. MULTIPLE FISH TRACKING	
4.1 Introduction.....	29
4.2 Related Works.....	32
4.3 Proposed Methodology.....	37
4.3.1 Appearance Similarity.....	38
4.3.2 Motion Similarity.....	40
4.3.3 Spatial Similarity.....	43
4.3.4Joint Optimization.....	43
4.4Text Explanation of Algorithm.....	44
4.5Experimental results and analysis.....	46
4.5.1 Experimental setup and parameter settings.....	46
4.5.2 Dataset Details.....	47
4.5.3 Quantitative Results.....	48
4.5.4Qualitative Results.....	52
4.6 Summary.....	53
5. GROUP DETECTION	
5.1 Introduction.....	56
5.2 Related Works.....	58
5.2.1Group Detection Methods.....	58
5.2.2Deep-Learning based Clustering Methods.....	59
5.2.3 Existing Underwater Datasets.....	60

5.3Proposed DataSet.....	62
5.3.1Data Collection & Preparation.....	62
5.3.2Ground Truth Generation.....	63
5.3.3 Human Census.....	64
5.4Proposed Approach.....	66
5.4.1Guided Generalized KLT Tracker(GGKLT).....	67
5.4.2Temporal Auto-encoder.....	68
5.5Experimental Results & Analysis.....	71
5.5.1Experimental Setup.....	71
5.5.2Quantitative Results.....	71
5.6 Approaches that didn't work.....	71
5.7 Summary.....	73
6. CONCLUDING REMARKS	
6.1 Conclusion.....	76
6.2 Future Scope.....	76
Bibliography.....	79
List of Publications.....	101
Biodata.....	102

List of Figures

1.1 (a) On-surface Data : MOT2016 (b) & (c) Underwater Data.....	2
1.2 Underwater Drones (a) Soft Robotic Fish (SoFi) (b) BionicRobotic Fish (BIKI)	5
1.3 Thesis Layout	9
2.1 Existing Databases.....	13
3.1 Challenges in marine environment (a) Camouflage (b)Hazy environment (c) Clutter.....	20
3.2 Various motion estimation techniques in video.....	21
3.3 Flowchart of the proposed method.....	22
3.4 Pipeline of the proposed flow method. Given two frames of the video, we first perform enhancement using Light scattering technique as proposed in [1]. Next, we use DeepMatching technique to get correspondence between two frames[2] and contours of the first frame (t) is computed using Holistic Nested Edges (HED) [3]. Finally, we combine these two cues to get dense interpolation image which enables computation of dense correspondence field. this is used to initialize the energy minimization framework of optical flow generation.....	23
3.5 (a) Original Image. Enhanced image by (b) DehazeNet[4] and (c) Light Scattering Model[1].....	24
3.6 (a) Original Image. Edges detected by (b) Canny[5], (c) SED [6] and (d) HED[3].....	24
3.7 Success cases: (a) Frame t (b) Frame t+1 (c) Proposed (d) DDFlow (e) EF (f) LDOF (g) SIFTFlow.....	27
3.8 Failure cases: (a) Frame t (b) Frame t+1 (c) Proposed (d)DDFlow (e) EF (f) LDOF (g) SIFTFlow.....	27
4.1 Left: Constrained Scenario, Right: Unconstrained Scenario.....	30
4.2 Issues in Multiple object tracking	31
4.3 (a) Positive Pair (b) Negative Pair for Siamese Network training	37
4.4 Overview of the proposed architecture: DFTNet	38
4.5 Sample trajectories in (a) Frame t (b) Frame t+1 (c) Frame t+2 (d) Frame t+3 (e) Frame t+4 of different videos of Fish4knowledge database	51

4.6 3D Figure of (a) Ground Truth Trajectories and Trajectories generated by (b) DFTNet(Proposed) (c) V-IOU Tracker (d) IOU Tracker (e) DeepSORT (f) MDP Tracker....	52
4.7 3D Figure of an example video having low density of fishes (a)Ground Truth Trajectories and Trajectories generated by (b) DFTNET (Proposed) (c) V-IOU Tracker (d) IOU Tracker (e) DeepSORT (f) MDP Tracker.....	54
4.8 3D Figure of an example video having medium density of fishes (a) Ground Truth Trajectories and Trajectories generated by (b) DFTNET (Proposed) (c) V-IOU Tracker (d) IOU Tracker (e) DeepSORT (f) MDP Tracker.....	54
4.9 3D Figure of an example video having high density of fishes(a) Ground Truth Trajectories and Trajectories generated by (b) DFTNET (Proposed) (c) V-IOU Tracker (d) IOU Tracker (e) DeepSORT (f) MDP Tracker.....	55
5.1 Comparison between Proposed UWCAD: Underwater Crowd Analysis Dataset and existing pedestrians dataset for crowd analysis.....	62
5.2 Statistics of Crowd density in UWCAD: Extreme, High, Medium and low.....	63
5.3 Statistics of Bycatch, Collectivenss and Cohesiveness in UW-CAD.....	63
5.4 Few sample frames of UWCAD for each group category (a) School (b) Shoal (c) Bycatch (d) more than one group in a video.....	64
5.5 Instruction set for crowd density, minimum and maximum number of group in video....	65
5.6 Instruction set for group detailing of each group for Collectiveness, Cohesiveness and Bycatch.....	66
5.7 Proposed framework for Underwater temporal auto-encoder for group detection.....	67
5.8 Number of trajectories generated by GKLT vs Number of trajectories after removing redundant trajectories by GGKLT for each video of UWCAD.....	68
5.9 (a) Motion vectors generated by gKLT (b)Nearest Neighbour-hood Selection with in specific radius(c) Motion vectors generated by GGKLT.....	69
5.10 Temporal Auto-encoder to learn latent representation for	70

5.11 estimated number of groups with Handcrafted features and learned features using various clustering methods.....	72
5.12 Metrics obtained for group detection using proposed UTAG, Collective Density Clustering(CDC), Crowd Collectiveness(CS), Handcrafted feature(HF-GD) and Learned Features (LF-GD)	72
5.13 Precision, Recall, F1 Score obtained for group detection using proposed UTAG, Collective Density Clustering(CDC), Crowd Collectiveness(CS), Handcrafted feature(HF-GD) and Learned Features (LF-GD).....	73
5.14 Framework of Co-motion Saliency for foreground region.....	74
5.15 frame with output	74

List of Tables

4.1 Ablation Study.....	48
4.2 Quantitative Results on Fish4knowledge videos.....	49