COLLEGE MATHSTIC
DOCTORAL BRETAGNE
BRETAGNE OCEANE

ÉCOLE NATIONALE D'INGÉNIEURS DE BRIEST

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE
D'INGÉNIEURS DE BREST

ÉCOLE DOCTORALE N° 644
Mathématiques et Sciences et Technologies
de l'Information et de la Communication en Bretagne Océane
Spécialité : *Human-Robot Interaction*

Par

Natnael Argaw Wondimu

« Application of Interactive Machine Learning Models For Human-Robot Interaction»

« Human Attention Based Approach »

Thèse présentée et soutenue à « Brest », le « June 30, 2023 »
Unité de recherche : «Lab-STICC»

Rapporteurs avant soutenance :

Denis KALKOFEN Flinders University, Australia
Alexandre PAUCHET INSA Rouen Normandy

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Sophie SAKKA	INSHEA
Examinateurs :	Denis KALKOFEN	Flinders University, Australia
	Alexandre PAUCHET	INSA Rouen Normandy
	Sophie SAKKA	INSHEA
	Jean-Philippe DIGUET	CNRS, IRL CROSSING, Australia
	Florian RICHOUX	AIST, Japan
Dir. de thèse :	Cedric BUCHE	CROSSING, CNRS IRL 2010, SA, Australia

Invité(s) :

Co-dir. de thèse : Ubbo VISSER University of Miami, USA

Acknowledgement

I would like to express my heartfelt gratitude and appreciation to several individuals and institutions who have contributed to the completion of my PhD dissertation.

Firstly, I extend my sincere thanks to my PhD supervisor, Prof. Cedric Buche, whose invaluable guidance, support, and encouragement have been critical throughout my doctoral journey. I am also deeply grateful to my co-supervisor, Prof. Ubbo Visser, for his insightful comments and expert advice that have enhanced the quality of my research.

I would like to acknowledge the members of my CSI committee, namely Prof. Florian RICHOUX and Prof. Jean-Philippe DIGUET, for their valuable feedback and suggestions that have helped me to improve my research work.

Furthermore, I am grateful to my wife Bizuayehu Tesfaye, my mother Amarech Shnekute, and my two kids, Kal Natnael and Lois Natnael, for their unwavering support and understanding during my PhD studies.

I also wish to express my gratitude to the French Embassy in Ethiopia, Brittany region administration, and the Ethiopia Ministry of Education (MoE) for providing the financial support that made this research possible. Finally, I am indebted to the Brest National School of Engineering (ENIB) and specifically LAB-STICC for creating a conducive research environment that enabled me to undertake this research project.

Table of Contents

INTRODUCTION

Human-robot interaction (HRI) is an emerging field that has been receiving growing attention in recent years. HRI involves the design and development of robots that can interact with humans in a natural and intuitive way. However, designing robots that can effectively interact with humans remains a challenge due to the complexity and variability of human behavior [1]. One critical aspect of HRI is the ability of robots to understand and respond to human attention cues, such as gaze direction and body orientation [2, 3]. The application of attention models to HRI is still in its early stages, and there is a need for research on how to develop and integrate these models into HRI systems [4].

Several studies have shown the importance of human attention models for anthropomorphic, intuitive and efficient HRI. For instance, research has demonstrated that incorporating attention models into HRI systems can improve the robot's ability to engage and interact with humans in a more natural and intuitive way [5]. Moreover, attention models have been shown to be effective heuristic functions for other computer vision models, such as object detection and segmentation [6].

Background

Attention is the process of selectively focusing on certain stimuli while ignoring others [7]. It plays a crucial role in human perception, memory, learning, and decision-making. Attention has been studied extensively in cognitive and neuroscience research, which has identified several factors that affect attention, including top-down attention, bottom-up attention, and attentional control.

According to Chun and Jiang (1998), top-down attention refers to attention that is driven by the goals, expectations, and knowledge of the observer [8]. For example, if someone is searching for a specific object, they will selectively attend to stimuli that are relevant to their search while ignoring irrelevant stimuli. Bottom-up attention, on the other hand, refers to attention that is driven by the sensory features of the stimuli, such as color, shape, or motion. For example, a sudden loud noise may capture someone's attention, even if they were not actively searching for it. Attentional control refers to

the ability to selectively allocate attentional resources to certain stimuli while inhibiting others [9].

Attention modeling plays a vital role in the development of robots that interact with humans in a seamless and intuitive way. Research on human attention enables the creation of models for robots to identify and respond to human attention cues, thereby enhancing their performance and usefulness in real-world applications. For instance, in social robotics, robots can use attention cues to engage in nonverbal communication with humans, such as maintaining eye contact, responding to gestures, and anticipating actions [10]. Moreover, attention modeling has extensive applications in HRI, including assistive robotics, entertainment robotics, and industrial robotics. For example, assistive robots for elderly people can use attention cues to assist in daily tasks such as fetching items or providing medication reminders [10]. Entertainment robots, on the other hand, can benefit from attention modeling to enhance their ability to interact with humans in games and social activities [11]. In industrial robotics, attention modeling can be applied to improve the safety of human-robot collaborations by enabling robots to detect and respond to human attention cues during task execution [12]. Therefore, the importance of understanding human attention and developing attention-based models for robots cannot be overemphasized as it offers numerous opportunities to enhance the performance and capabilities of robots in various HRI domains.

There are several methods of modeling human attention, mainly saliency video prediction and moving object detection models. Saliency video prediction models use computer vision techniques to predict the most salient regions of a video frame, which are the regions that are most likely to capture human attention. These models typically use features such as color, intensity, and motion to predict saliency. According to Borji and Itti (2012), saliency prediction models have been shown to be effective in predicting human fixations in natural scenes [13].

Moving object detection and segmentation models use computer vision techniques to detect and segment moving objects in a scene, which are likely to capture human attention. These models typically use motion analysis techniques to detect moving objects. According to Wang et al. (2003), moving object detection models have been shown to be effective in detecting and tracking human movements in HRI applications [14].

Research Problem and Questions

HRI is a rapidly evolving field that aims to improve the interactions between humans and robots. With the increasing prevalence of robots in our daily lives, it is becoming increasingly important to develop robots that can effectively interact with humans in various contexts. While traditional approaches to HRI have focused on designing robots with pre-programmed behaviors and responses, recent advances in machine learning have enabled the development of more sophisticated HRI systems that can learn from and adapt to human behavior [2, 15, 16].

One critical aspect of HRI is the ability of robots to understand and respond to human attention cues. Attention is a fundamental cognitive process that allows humans to selectively process information and focus on relevant stimuli. By understanding human attention cues, robots can adapt their behavior and responses to better meet human needs and preferences [17]. For example, a robot that can accurately predict where a human is looking can better anticipate the human's needs and preferences, leading to a more intuitive and efficient interaction.

However, the application of attention models to HRI is still in its early stages, and there is a need for research on how to develop and integrate these models into HRI systems. Current attention models are often developed and tested in controlled laboratory settings, and there is a lack of research on how to apply these models in real-world HRI scenarios. Furthermore, there is a need to develop attention models that can be integrated into existing HRI frameworks, such as the Robot Operating System (ROS) [18], to enable more intuitive and adaptable HRI.

To address these challenges, this thesis aims to apply interactive machine learning models to HRI by developing and testing attention models that can be used as heuristic functions in real-world HRI applications. By developing and testing these models, we can gain a better understanding of the underlying principles of attention in HRI and how they can be leveraged to create more intuitive and adaptable robots. The goal is to develop attention models that can be integrated into existing HRI frameworks, enabling more effective and efficient HRI in various application domains.

These are the main research questions this thesis is going to address.

— What are the key computer science principles underlying attention models in HRI, and how can these principles be translated into interactive machine learning models ?

— How effective are the proposed attention models, including the Video Saliency Prediction Model and the Moving Object Detection and Segmentation Model, in predicting and responding to human attention cues in real-world HRI scenarios ?
— How can the proposed attention models be integrated into existing HRI frameworks, such as the ROS, to enable more intuitive and adaptable HRI ?
— How do the proposed attention models compare to other state-of-the-art machine learning models in terms of accuracy, efficiency, and robustness, and what are the implications for future research in this area ?
— What are the potential applications of the proposed attention models, both within and beyond the field of HRI, and how can they be used to improve the performance and usability of interactive systems in various domains ?

Objectives and hypotheses

The main objectives of this thesis are :

— Develop and evaluate human attention models, namely Video Saliency Prediction Model and Moving Object Detection and Segmentation Model, for HRI applications.
— Develop an integrated framework for applying these attention models to human-robot interactions on the Pepper Humanoid Robot using the ROS.
— Test the efficacy of the developed attention models and the integrated framework in various HRI scenarios.

Moreover, we hypothesize that

— The Video Saliency Prediction Model and Moving Object Detection and Segmentation Model will accurately predict human attention cues in various HRI scenarios, leading to more efficient and intuitive interactions.
— The integrated framework for applying these attention models to HRI will improve the overall performance of Social Robots like Pepper Humanoid Robot in various HRI scenarios.

Methodology

The methodology for this thesis will consist of four main phases : data collection, model development, model evaluation, and integrated framework development.

The data collection methodology for this thesis will involve using existing large dynamic gaze fixation datasets, such as DHF1K [19], solely for the development and evaluation of the Video Saliency Prediction Model. This model aims to predict human attention cues in video frames accurately. On the other hand, for the Moving Object Detection and Segmentation Model, we will be using activity recognition datasets such as CDNet2014 [20, 21]. These datasets contain a wide range of activities and will be used to evaluate the model ability to detect and segment moving objects in dynamic environments. By utilizing existing datasets, we can reduce the cost and time associated with data collection, allowing us to focus on the development and evaluation of the attention models for efficient HRI.

In the model development phase, two attention models, namely the Video Saliency Prediction Model and Moving Object Detection and Segmentation Model, will be developed using state-of-the-art machine learning techniques. The models will be trained using the collected data from the previous phase. The Video Saliency Prediction Model will predict regions of the video that are most likely to attract human attention, while the Moving Object Detection and Segmentation Model will identify and segment moving objects in the video, which are also likely to attract human attention.

In the model evaluation phase, the developed models will be evaluated based on their ability to predict human attention cues accurately. The evaluation will be conducted using various performance metrics, including but not limited to Area Under the Curve - Judd (AUC Judd), Shuffled Area Under the Curve (s-AUC), Correlation Coefficient (CC), Similarity (SIM), Normalized Scanpath Saliency (NSS), precision, recall, and F-measure (F1-score). The models will be evaluated using a set of test data that were not used for model development.

In the integrated framework development phase, an integrated framework will be developed for intuitive and anthropomorphic HRI. The integrated framework will be based on the Robot Operating System framework and will be designed to work with the Pepper Humanoid Robot. To evaluate the performance of the integrated framework for intuitive HRI, subjective rating will be conducted. These studies will involve human participants interacting with the Pepper Robot both in virtual and real environments, in various scenarios, while their responses will be evaluated to assess the framework from various perspectives under four scenarios.

Significance of the research

This thesis is important as it addresses the critical need for more intuitive and responsive HRI systems through the development and integration of human attention models. The contribution of this thesis to the field of saliency prediction, moving object detection and segmentation, and HRI lies in the development and evaluation of two novel human attention models, namely the Video Saliency Prediction Model and the Moving Object Detection and Segmentation Model, and their application to HRI scenarios. In video saliency prediction, we explore the use of spatio-temporal features to efficiently identify visually salient regions in videos. Our approach involves an interactive model that combines a stacked-ConvLSTM architecture with a custom XY-shift frame differencing layer. On the other hand, in the domain of moving object detection and segmentation, we emphasize the importance of motion in capturing human attention and identifying areas of interest. To address this, we propose an innovative method that combines XY-shift frame differencing and three-frame differencing techniques with a three-stream encoder-decoder architecture. The model integrates feature maps from both the original frame and the frame differencing component, utilizing transfer learning with VGG-16 as the convolutional base. The resulting segmentation map is generated through deconvolution.

The proposed integrated framework developed in this thesis can significantly improve the overall performance of robots in various HRI scenarios, which has practical implications in many fields. For example, in healthcare [10], robots can be used to assist with tasks such as patient monitoring and care, while in education [22], robots can be used to assist in teaching and learning. Moreover, this thesis can have potential impacts on policy or practice, particularly in the context of RoboCup@Home Social Standard Robotics. The RoboCup@Home Social Standard Robotics aims to create a benchmark for the development of socially intelligent robots that can interact with humans in various real-world scenarios. The findings of this thesis can contribute to the development of robots that meet the standards set by RoboCup@Home Social Standard Robotics.

In summary, the contribution of this thesis to the field of saliency prediction, moving object detection and segmentation, and HRI lies in the development and evaluation of two novel human attention models and the proposed integrated framework. The potential practical applications and impacts of this thesis on policy or practice are significant, particularly in the context of RoboCup@Home Social Standard Robotics.

Structure of the thesis

In this thesis, we present findings at this research interface of machine learning, human-robot interaction, computer vision and human attention models. Specifically, we develop and employ human attention models and reconfigure the standard human-robot interaction frameworks in a way they consume human attention model to achieve intuitive, anthropomorphic, and efficient HRI. Hence, this thesis is structured under the following chapters :

The Introduction part, the current section, provides an overview of the research problem and objectives, as well as the research questions that this thesis aims to answer. It also explains the significance of the research in the context of HRI.

Chapter 1 reviews the existing literature on saliency prediction, moving object detection and segmentation, and HRI. It also discusses the strengths and limitations of the current approaches and identifies research gaps that this thesis aims to fill.

Chapter 2 presents the proposed video saliency prediction model and the methodology used to develop and evaluate it. It also presents the experimental results and discusses the strengths and limitations of the proposed approach.

Chapter 3 presents the proposed moving object detection and segmentation model and the methodology used to develop and evaluate it. It also presents the experimental results and discusses the strengths and limitations of the proposed approach.

Chapter 4 presents the integrated framework that combines the attention models with control algorithms. It also describes the methodology used to evaluate the framework and presents the experimental results.

The Conclusion summarizes the main findings of the thesis and highlights the contributions to the field of HRI and other related fields. It also discusses the limitations of the proposed approaches and suggests potential future works to address these limitations.

Finally, in the Publications section, we have incorporated a chapter enclosing list of our publications constituting this thesis.

Chapitre 1

State of the Art

In recent years, there has been a growing interest in developing intelligent robotic systems that can interact with humans in a natural and intuitive way. The development of such systems requires a deep understanding of both the technical and social aspects of human-robot interaction (HRI) [23].

To this end, the development of interactive machine learning models has opened up new possibilities for HRI. These models allow robots to learn from human interactions and improve their ability to understand and respond to human behavior. These models allow robots to adapt to different situations and respond to human behavior in a more natural and intuitive way [24].

Interactive machine learning models (IMLM) refer to a class of machine learning models that are designed to engage with users to improve their performance [25]. IML models can take a variety of forms, but they generally involve some degree of interaction between the model and the user, with the aim of improving the accuracy and usefulness of the model. One of the key contributions of IMLM to HRI is their ability to improve the efficiency and accuracy of human-robot interactions. Robots equipped with these models can quickly adapt to changes in human behavior and respond appropriately. This makes it easier for humans to interact with robots, which can lead to increased productivity and efficiency in a variety of settings. Another contribution of interactive machine learning models to HRI is their ability to improve the safety of human-robot interactions. Robots equipped with these models can learn to recognize and respond to potentially dangerous situations, reducing the risk of accidents and injuries. Interactive machine learning models also have the potential to improve the overall user experience of interacting with robots. Robots equipped with these models can learn to recognize and respond to human emotions, making interactions more natural and intuitive. This can improve the overall user experience and make it easier for humans to interact with robots in a variety of settings.

A more sophisticated and nuanced understanding on human attention can be acquired by incorporating data on human attention into the machine learning process. Human-

attention models provide a more focused and specialized approach to understanding human attention, allowing robots to more accurately detect and respond to human behavior.

In this thesis, we present our research contributions that are inspired by the synergy among interactive machine learning models, human-robot interaction interfaces and perception mechanisms, and human attention models (see Figure 1.1). We have undertaken various researches towards improving the state-of-the-art human attention models (such as saliency prediction models and moving object detection and segmentation models) and implemented them to enhance the anthropomorphic capability and efficiency of robots in human-robot interaction setting.

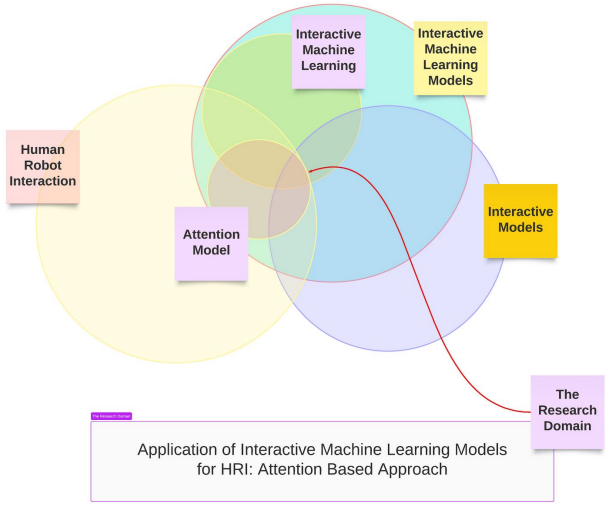


FIGURE 1.1 – Diagrammatic Representation Domain of our Thesis

Consequently, in the following sections, we will present the current state of the art in these areas, highlighting recent advancements and key research findings. By examining the latest research in interactive machine learning, interactive models, and human-attention models, we can gain a deeper understanding of the opportunities and challenges in human-robot interaction and pave the way for the development of more effective and engaging robotic systems.

1.1 Human-Robot Social Interaction

Over the past few decades, there have been significant advancements in the field of human-robot social interaction [26]. These advancements have led to the development of robots that are capable of interacting with humans in natural and meaningful ways,

enhancing collaboration and improving the overall user experience [27, 28, 29, 30, 31, 32] .

One of the key areas of research in human-robot social interaction is developing algorithms that enable robots to interpret human gestures, facial expressions, and vocal cues [33, 15, 34, 35, 36, 37, 38, 39]. These algorithms often rely on machine learning and computer vision techniques to recognize and interpret human actions and emotions. For example, researchers have developed algorithms that allow robots to recognize different facial expressions and respond accordingly. This enables robots to interact with humans more effectively, improving the overall user experience. Another area of research in human-robot social interaction is developing social robots that can understand and respond to human emotions [15]. These robots are designed to be empathetic and to provide emotional support to humans. Researchers have developed algorithms that enable robots to recognize human emotions based on vocal cues, facial expressions, and body language. These robots can then respond with appropriate emotional cues, such as facial expressions or vocal intonations, to provide emotional support to humans.

Advancements in interfacing mechanisms have also played a crucial role in improving human-robot social interaction. Researchers have developed various interfacing mechanisms, such as speech recognition [40, 41, 42], haptic interfaces [43, 44], and voice assistants [33], to enable humans to interact with robots more easily. These mechanisms allow humans to communicate with robots in natural ways, improving the overall user experience. Moreover, perception mechanisms have also seen significant advancements in recent years. Researchers have developed sensors that allow robots to perceive their environment more effectively, enabling them to navigate and interact with their surroundings more effectively. For example, robots can use depth sensors [45] to perceive objects in their environment and avoid obstacles while navigating. These advancements in human-robot social interaction, interfacing, and perception mechanisms have significant implications for the development of future robotic technologies. It is enabling the development of robots that are capable of interacting with humans in natural and meaningful ways, improving collaboration and enhancing the overall user experience. As robots become increasingly present in our lives, it is crucial to continue investing in these areas of research to ensure that robots can effectively support human needs and enhance human experiences.

In the following parts of this section, we discuss the state-of-the-art on the key enabling technologies to achieve a seamless and intuitive HRI. We discuss the state-of-the-art of intermediate human–robot interfaces (bi-directional) and state-of-the-art perception me-

chanisms.

1.1.1 Human-Robot Interaction Interfaces

Human beings are a social species that relies on cooperation to survive and thrive. Humans embrace a diversity of experiences working together. Such cooperative working environments enabled the development of implicit and explicit communication standards for intuitive perception and communication of task-oriented information flow [16]. Design of such communication standards where the robot has information of human intentions and needs has been the main objectives of human-robot interaction researches [26, 46, 47]. This is due to the fact that understanding communication principles can potentially lead to an enhanced physical human–robot interaction performance [48].

A common type of communication interfaces are built on the use of visual or language commands. In these interfaces, humans employ such user-friendly techniques to communicate with the robot. The effort to integrate such visual [49], language commands [33], or their combination [50, 51], started in early of robotics. The use of head [52], body [53] or arm gestures [35, 36, 37, 38, 39] are common examples in the areas of human–robot interaction and collaboration. In this direction, a method to interpret the human intention from the latest history of the gaze movements and to generate an appropriate reactive response in a collaborative setup was proposed in Sakita et al. (2004) [54]. Authors in Hawkins et al. (2013) developed a vision-based interface to predict in a probabilistic manner when the human will perform different sub-tasks that may require robot assistance [55]. The developed technique allows for the tracking of the human variability, environmental constraints, and task structure to accurately analyse the timings of the human partner's actions.

Such audio-visual features appear natural to humans. However, their usage is limited to activating high-level robot operations. This is mainly due to the complexity of deriving the desired sensorimotor behaviour from these higher-level features. Hence the derivation of such complex sensorimotor behaviour from audio-visual features require better capabilities of autonomous robots. On the other hand, the use of robots for a wider range of applications require these vision or auditory based interfaces.

Human attention has been analysed through interfaces that use force/pressure sensors. This has been used an alternative human-robot interface mechanism to audio-visual features due to its simplicity. The application of such interfaces has been observed in collaborative object transportation [43, 56, 57, 58, 59, 60, 61], object lifting [62, 63], ob-

ject placing [61, 64], object swinging [65, 66], posture assistance [67, 68], and industrial complex assembly processes [69, 70]

The interaction forces/torques are used to regulate the robot control parameters and trajectories in the aforementioned research works. This is following the admittance [71] or impedance [61, 57] causality. Despite the large margin of applications, tasks in a shared and non-deterministic environment can induce various unpredictable force components to the sensor readings [72]. This can significantly reduce the suitability of such an interface in more complex interaction scenarios since it can be difficult to distinguish the components related to the active counterpart(s) behaviour from the ones generated from the interaction with the environment.

On the contrary, the utilization of physiological indices to comprehend the profound and intensified dynamics of the human physiological system, although valuable, presents significant challenges when applied in real-time human-robot interaction (HRI) scenarios. For instance, techniques such as electromyography (EMG) [73] and electroencephalography (EEG) [74], as well as other bio-signals like electrodermal activity [75, 76], have been employed to discern human intentions. However, their implementation in real-time HRI applications remains rare due to their weightiness and the intensive processing they require. For example, the adaptability and user-friendliness of EMG have led to exhaustive experimentation in human-in-the-loop robot controls. EMG signals have found widespread use in diverse domains, including prosthesis [77, 78], exoskeletons [79, 80], and industrial manipulator control [81, 82]. These research endeavors have utilized EMG signals to anticipate the stiffening or complying behavior of torque-controlled robotic arms in co-manipulation tasks. Notably, in the work by [82], EMG signals were employed to predict the stiffening or complying behavior of a torque-controlled robotic arm during a co-manipulation task, enabling real-time estimation of the leading and following roles of the human and robot counterparts. In another study by [83], EEG signals were employed to command a partially autonomous humanoid robot based on high-level descriptions of the task. However, the usage of EEG signals in real-time HRI applications necessitates intensive processing and is therefore infrequently applied.

In addition to human's intentions and control commands, it is crucial for the robots to estimate the emotional states of a human partner in order to be socially responsive, engage longer with users and promote natural HRI [34]. More importantly, estimation of workload, emotion and or anxiety and errors is crucial for ergonomic and safe human-robot collaboration in both domestic and industrial spaces [15, 26]. Authors in Szafir et

al. (2013) reported an interesting study in which a humanoid robot monitored students' EEG signals during storytelling and gave them attention-evoking immediacy cues (either in verbal or non-verbal form) whenever engagement drops were detected [84]. In doing so, they extracted EEG levels in alpha, beta and theta frequency bands and smoothed them into an engagement signal that would represent attention levels. Every time the attention level went below a pre-defined threshold, the robot displayed immediacy cues such as increased spoken volume, increased eye contact, and head-nodding. Similarly, Ehrlich et al. (2014) proposed an EEG-based framework for detection of social cues such as gaze by a humanoid robot as a measure for social engagement [85]. They instructed subjects to either wait for the robot to make eye-contact with them or to intentionally generate brain patterns for the robot to initiate eye-contact with them. By extracting frequency band powers as discriminating features in an offline analysis, they could find high classification performance between the two conditions. Such predictive model could be implemented in a human-robot interaction in order to enable the robot to estimate its social role and adapt its behavior to the expectations of the human partner. Another research in Rani et al. (2004) [86] developed a method to detect human anxiety in a collaborative setup by extracting features from EMG, electrocardiography (ECG) and Electrodermal responses. In a similar work, the human physical fatigue was detected and used to increase the robot's contribution to the task execution [87].

Unique sensory data based interfaces can configure a pre-defined robot behaviour in collaborative settings. However, the functionality of these interfaces is limited and cannot be easily generalized across domain scenarios. Put succinctly, force or pressure sensors outperform visual feed back based estimation of exchange amount of energy. Similarly, the use of bio-signals such as EMGs for tracking of the human limb movements may result in less accurate performances in comparison to the external optical or Inertial Measurement Unit (IMU) based tracking systems [88, 89]. To this end, multi-modal interfaces [90, 91, 82], interfaces that associate multi-modal sensory information to different robot control modalities, is proposed by various researches. For instance, the authors in Agravante et al. (2014) [57] and Rozo et al. (2016) [92] proposed a hybrid approach by merging vision and force sensing, to decouple high- and low-level interaction components in a joint transportation task where a human and humanoid robot carry a table with a freely moving ball on top. A similar work proposed a multi-modal scheme for intelligent and natural human–robot interaction [93] by merging vision-based techniques for user localisation, person localisation and person tracking and their embodiment into a multi-modal overall

interaction schema.

Likewise, auditory features has been used with other interfaces in multi-model setting. It has been used to pause, stop or resume the execution of a dynamic co-manipulation task, the control parameters of which regulated by an EMG based interface. In Yang et al. (2016), authors developed a multi-modal teaching interface on a dual-arm robotic platform [94]. The interface was built on the use of EMG on the user arm and force sensors on robot end-effector. In this setup, one robotic arm is connected to the tutee's arm providing guidance through a variable stiffness control approach, and the other to the tutor to capture the motion and to feedback the tutees performance in a haptic manner. The reference stiffness for the tutors arm stiffness was estimated in real-time and replicated by the tutee's robotic arm. In Ivaldi et al. (2017), multi-modal communication of people interacting physically with the humanoid iCub to build objects is studied [95]. Participants would naturally gaze at the robot's hands or face to communicate the focus of attention of the collaborative action, while speaking to the robot to describe each action. The authors found that individual factors of the participants influence the production of referential cues, both in speech and gaze : particularly, people with negative attitude towards robots avoid gazing at the robot, while extroverted people speak more to the robot during the collaboration.

Multi-modal interfaces improved the performance of compound robot behaviour generation. This fact attracted researchers attention towards the usage of multi-source sensory information in for improved HRI [26]. The inclusion of multiple communication channels in the development of the intermediate interfaces will potentially contribute to an increase in the human cognitive burden and the low level robot control complexity. However, the more communication channels included, the worst the intuitiveness of the HRI and the excessive human effort to operate in a specific robot modality. This has been addressed by multiple authors giving much emphasis on shared communication modalities [96, 97]. Alternative robotic learning techniques such as gradual mutual adaptation [68, 82], reinforcement learning [65] or learning from demonstration [98, 62, 99, 58], has been used to loosen the communication loop demands that come along with an increased level of robot autonomy. To mention a recent contribution in this regard, an extended version of Gaussian Mixture Model (GMM) with weighted data coupled with Gaussian Mixture Regression (GMR) is used for learning by demonstration in Legeleux et al.(2022) [98]. Their model gives the possibility to the user to impact the learning by choosing which parts of the demonstration has more importance. The performance of the model's in trajectory

generation is tested using two tasks and two robots.

1.1.2 Human-Robot Interaction Perception Mechanisms

Audio-visual systems are a source of powerful sensory inputs that contribute to a fast and accurate perception of the movement kinematics and the environment. These systems allow constant updates on internal models. Audio-visual sensory inputs have been used in various human-robot interaction applications. It has has been applied to complex HRI problems such as, human tracking [40], object weight anticipation [41], and force estimation [42].

Furthermore, dyadic interaction in human-robot interaction often implement mutual gaze and joint attention [100]. Joint attention based researches govern the robot to attend to the same object the entity in contact is looking at. For instance, [101] proposed a joint focus of attention in human-robot interaction. They used the positional information obtained from pointing gesture and saliency map obtained from biologically motivated saliency model. Similarly, Clair et al, (2011) investigated the effect of visual saliency in pointing gestures in open space [39]. The purpose of this research is to make the robot (actually or ostensibly) shift its gaze towards its intended referent using saliency. In Ivaldi et al. (2014) the robot was equipped with anticipatory gaze mechanisms and proactive behaviours, increasing the pace of the interaction and reducing the reaction time of the human to the robot's cues [102]. In this research the potential of improving mutual awareness, hence the task performance. In addition, in Saran et al. (2018), a deep learning approach which tracks a human's gaze from a robot's perspective in real time is proposed [103]. Their work uses the gaze heat map statistics to capture differences between mutual and referential gaze conditions, which they use to predict whether a person is facing the robot's camera or not. Similarly, in Shi et al. (2019), a novel approach to detect whether a human is focusing on an object in HRI application is proposed [104]. They use Earth Mover's Distance (EMD) to measure the similarity and 1 Nearest Neighbour to classify which object a human is looking at. Other researches such as, [12, 101, 105], investigated the possibility employing saliency models for joint attention.

On the other hand, several research studies have focused on haptic information for human-robot interaction. One approach is to use wearable haptic devices, such as gloves or armbands, to provide haptic feedback to users [44]. These devices can provide users with a sense of touch and pressure, allowing them to interact with robots in a more natural and intuitive manner. Researchers are also exploring the use of haptic feedback in Virtual

Reality (VR) and Augmented Reality (AR) environments, where users can interact with virtual objects using haptic devices [106, 107, 108, 109]. Another area of research is the use of haptic feedback in shared control scenarios, where both humans and robots work together to complete a task. Haptic feedback can be used to provide information to both the human and robot, enabling them to work together more effectively. Haptic information is an essential modality for human-robot interaction. It can be used to facilitate communication, convey information, and improve safety and efficiency. However, designing effective haptic interfaces for human-robot interaction presents several challenges, such as designing interfaces that are intuitive and providing accurate haptic feedback. Despite these challenges, there is ongoing research in this area, and the development of new haptic devices and interfaces is likely to enable more natural and effective communication between humans and robots in the future.

By the same token, AR technologies have also been used to enhance perception of the environment, letting the human partner observe and review a plan with the robot prior to execution [106, 107, 108, 109]. AR can enhance the interaction between humans and robots in various ways. Firstly, augmented reality can provide a more intuitive and natural way of interacting with the robot. For example, instead of using a traditional user interface, a user can use hand gestures or voice commands to control the robot. This can make the interaction more natural and user-friendly. Secondly, augmented reality can provide a better understanding of the robot's capabilities and limitations. For example, using augmented reality, a user can see the robot's field of view and understand how it perceives the environment [110]. This can help the user to understand what the robot is capable of doing and what it cannot do. Additionally, augmented reality can provide real-time feedback on the robot's performance, which can help the user to adjust the robot's behavior to better meet their needs [111]. Augmented reality can provide a better understanding of the task at hand. For example, using augmented reality, a user can see the robot's actions overlaid on the physical environment, which can help the user to understand the robot's behavior and intentions [107]. Additionally, augmented reality can provide visual cues to guide the user in completing a task. For example, augmented reality can highlight the location of an object that the robot needs to pick up or show the user the correct way to manipulate the robot's arm [112]. Despite the benefits of augmented reality for human-robot interaction, there are also several challenges that need to be addressed [113]. The accuracy and reliability of augmented reality systems can be a challenge. Augmented reality relies on accurate tracking of the user's movements and

the robot's position and orientation. If the tracking is not accurate, the augmented reality experience can be disrupted, leading to a poor user experience. Moreover, the design of the augmented reality interface needs to be carefully considered. The interface should be intuitive and easy to use, but it should also provide enough information to the user to make informed decisions about the robot's behavior. Finally, AR can potentially be prone to information overloading, limited privacy (e.g. augmenting without permission) and additional cost, that may limit the expected performance of such systems in collaborative settings.

In conclusion, various HRI interfaces and perception mechanisms have been analysed. One of the significant features that most HRI systems that define the state-of-the-art share is their powerful cognitive models. Machine learning plays a critical role in building such cognitive models because it provides a powerful framework for analyzing and learning from large and complex datasets. By training machine learning models on large datasets of human behavior, researchers can identify patterns and regularities in cognitive processes, and develop computational models that can explain and predict these patterns. These models can be built following a classic machine learning processes or employing the human-in-the-loop principle of interactive machine learning. Consequently, we believe, an extensive state-of-the-art review on human-robot interaction should emphasize on machine learning techniques for HRI in general and IML in particular. In the next consecutive sections, we discuss the state-of-the-art of machine learning models built for human-robot interaction setting and a significant model building technique that has a special use to HRI, interactive machine learning.

1.2 Interactive Machine Learning Models

Human-robot interaction is nurtured both at physical and cognitive level. Cognitive models are typically built to collect inputs from the environment and from the user, elaborate and translate these into information that can be used by the robot itself. Machine learning has been the recent approach to build the cognitive models and behavioural block, with high potential in HRI. Consequently, it is assuming an important role in human–robot interaction. One of the main challenges in HRI is creating naturalistic interactions that are intuitive and easy for humans to understand. To this end, the synergy among machine learning, Interactive Machine Learning (IML), and human-robot interaction has been enabling the development of advanced and capable autonomous systems.

The connections between these three fields are manifold. Machine learning algorithms can be used to improve the performance of robots in HRI scenarios, by allowing them to learn from human feedback. For example, a robot designed to assist people with mobility impairments can learn from its interactions with users to improve its ability to anticipate their needs and provide effective assistance. Conversely, HRI can also be used to improve the performance of machine learning algorithms. For example, by providing users with an interface to provide feedback and corrections to machine learning models, we can improve their accuracy and effectiveness. This approach is particularly useful in situations where the data is noisy or difficult to label, as it allows humans to provide additional context and information to the system. Interactive machine learning can also be used to improve the performance of robots in HRI scenarios. By allowing users to provide feedback and suggestions to the system, we can create more naturalistic interactions that are intuitive and easy for humans to understand. For example, a robot designed to assist with cooking can learn from its interactions with users to improve its ability to understand and follow verbal instructions.

In conclusion, leveraging the connections between these fields, we can develop robots that are more effective, more naturalistic, and better able to collaborate with humans in a wide range of settings. Hence, in this part of the chapter, we discuss about the state-of-the-art on the significant contributions under interactive models and interactive machine learning. We start by discussing the state-of-the-art of a general interactive models for human-robot interaction. Then we extend our discussion to the most significant model building technique for human-robot interaction, IML. We close this section of the chapter by laying foundation to extended discussions on special forms of interactive models, human attention models.

1.2.1 Interactive Models

With the advent of technology, robots have become increasingly advanced and capable of interacting with humans in various settings. The field of human-robot interaction has gained significant importance, and researchers are exploring different interactive models to enhance the effectiveness of human-robot interaction. In this part of the chapter, we will discuss the state-of-the-art of interactive models used in HRI and how they are shaping the future of robotics. Specifically, we subdivide our discussion using the most common taxonomy of machine learning algorithms [114].

Unsupervised learning

The process of human-robot interaction involves modeling the relationship between a human and a robot in order to achieve a shared objective. However, designing such interactions using predetermined rules can be very difficult due to their complex nature. Therefore, many researchers have turned to probability theory to develop human-robot collaborations. In the field of machine learning, graphical models are often used to represent probability, with nodes connecting variables that are conditionally dependent on each other. To train ML models using probability, unsupervised learning algorithms are commonly utilized, which rely on unlabelled data during the learning process.

The Gaussian Mixture model (GMM) is the most commonly used type of unsupervised learning (UL) model. During the learning phase, a fixed number of multi-variable Gaussian distributions are fit to the training dataset, which is called Gaussian Mixture Regression (GMR) [98]. In the related graphical model, the observation is assumed to be conditionally dependent on the parameters that model the distributions and a latent variable vector that indicates which distribution the data point is likely to belong to [115]. These latent variables are also conditionally dependent on the distribution parameters. This ML model is trained using the Expectation Maximization (EM) algorithm in an unsupervised manner. Subsequently, the ML model categorizes a new data point based on the distribution it is most likely to belong to.

The TP-GMM, which is a modified version of the GMM, has significant applications in robotics [116, 58]. It enables the performance of GMR while considering various observation frames, which are incorporated into the model using task parameters. This algorithm yields a mixture of Gaussian distributions that fit the best performing frame for the specific task, which is advantageous for adaptive trajectories in robotic manipulators.

Another ML model commonly trained with an UL algorithm is the Hidden Markov model (HMM). The Hidden Markov model (HMM) is a machine learning model that is also trained using an unsupervised learning algorithm. It assumes that the current value of a random variable is dependent on its previous value, making it capable of considering time dependence. The variables in HMM are not directly observed but can only be inferred through indirect observations. These hidden variables are represented as nodes that are connected only to the relevant timestamp being considered [117]. In a HRC scenario, this can be seen as a robot trying to understand user's intentions by looking at its movements. HMMs are trained through a variation of the EM algorithm for HMMs, the Baum-Welch algorithm [118]. Authors in Rozo et al. (2016) [119]used it to perform a pouring task

while the human user is holding a cup, while Vogt et al. (2016) [120] employed a HMM to handle the collaborative assembly between a human and a robot.

During the learning phase of the algorithm, the HMM structure takes into account time dependence, which is a crucial factor in HRI. In most interactions, the sequence of actions taken before the current interaction influences the outcome. Therefore, it is crucial for a machine learning algorithm and model to consider this time-dependent aspect.

The last ML model in the selection is the variational autoencoder (VAE), which belongs to the deep learning (DL) family of models. These models consist of a neural network with two or more hidden layers capable of mapping highly complex nonlinear functions. In recent years, DL models have gained popularity in HRC. VAEs learn the parameters of a latent multivariate distribution as the output of a DL model. This output is then used to model the distribution, and a different deep neural network is used to model a distribution of the original training dataset. The model is trained by minimizing the ELBO bound [121]. VAEs can also be used for classification, without using any labels, through self-supervised learning. However, these models cannot take time dependence into account, and they are typically used as the first stage of a composite ML system. This is similar to the case of GMM as neither model can consider time dependence.

Reinforcement learning

Markov decision processes (MDP) are a type of graphical model that involves an agent interacting with an environment according to a particular policy, receiving a reward and an observation about the state of the environment. This scenario is highly relevant to HRC applications, where the robot acts as the agent and the collaborative environment represents the environment. MDPs can be fully observable [122], partially observable [123], or have mixed observability [124]. A wide range of learning algorithms are built around the MDP model, constituting the reinforcement learning (RL) family. In this case, the primary objective is to maximize a cumulative function of the reward over time, known as the value function. This highlights the inherent ability of RL algorithms to consider time dependence.

Such result is achieved in these works in different manners. Model-free is the most used family of RL algorithms, simply because they do not require a model of the transition of the environment from a state to the next one, which is hard to design beforehand. The most frequent algorithm of this kind is Q-learning [125, 126, 127, 124]. It is a model-free RL algorithm that updates the value function Q arbitrarily initialized. During the

learning phase, its values are then updated with a percentage of the current reward and the highest value depending on the action and possible future states. Most of the works that use Q-learning sticks to the traditional formulation. However, others try variations of it. Authors in Lu et al. (2020) used Q-learning with eligibility traces and fuzzy logic for their object handling task [128].

Q-learning is just one model-free reinforcement learning algorithm. In contrast, inverse reinforcement learning tackles a different problem by attempting to construct a reward function based on a history of past actions and observations, without using a pre-defined reward function during the training phase. This is because starting the training with a reward function can be difficult due to the complexity of the problem. In a recent paper [129], an inverse RL algorithm was used which incorporates robotic learning from demonstration during the learning phase of the algorithm when processing the received history of observations and actions.

In a distinct approach, [122] utilized interactive RL that involves the agent learning from two sources : environmental observations and a supplementary source such as feedback from a teacher or sensor feedback. A comparable technique is observed in the field of robotics research, where domain experts evaluate a robot's performance in achieving a particular task.

While model-free algorithms are more prevalent, there are some works that employ model-based RL algorithms, although in smaller numbers. For example, [130] utilized a model-based RL algorithm that models the dynamics between humans and robots through a neural network. In Huang et al. (2018), a PI (policy iteration) algorithm was utilized [116]. In this case, the optimal policy is sought by exploring noisy variations of the trajectories generated by the robot and updating the task parameters based on the cumulative value of a cost function, which can be considered the inverse of a reward function.

In machine learning systems that are composed of multiple components, reinforcement learning (RL) relies on information provided by other ML models, such as neural networks that are trained through supervised learning methods [128, 130]. Deep reinforcement learning is a particular type of composite ML system, in which a deep learning (DL) model is used to map a complex state space to a value space, providing RL with an output value given the current state. The DL model is trained together with the RL algorithm, so it becomes a part of it. Researchers in Ghadirzadeh et al. (2020) have successfully used deep Q-networks in HRC scenarios, but despite its potential in robotics, deep RL is not commonly used in HRC and should be further explored [123].

Supervised learning

As we have discussed in the above sub-sections, RL involves an agent interacting with the environment to learn a policy that maximizes cumulative rewards. While RL is a suitable approach for representing a robot's actions in the real world, some researchers still prefer to use supervised learning (SL) to construct a cognitive model.

Supervised learning differ from RL in its data and process management. In SL, the model is trained on labeled data, where the input and output pairs are known. The model tries to learn the relationship between input and output by minimizing the difference between its predictions and the actual labels. In terms of data, SL uses labeled data, while reinforcement learning generates its own data through interactions with the environment. Overall, these differences make supervised learning suitable for scenarios with labeled data, while reinforcement learning is more suitable for scenarios where the agent must learn to interact with an environment to achieve a goal.

To this end, Supervised training of models based on probability distributions is possible, as demonstrated by various examples. One common approach is Naive Bayes classification, which utilizes Bayes theorem for supervised training. For instance, in Vinanzi et al. (2020), Naive Bayes classification was employed to recognize human intention in a collaborative assembly [131]. Another example is Peternel et al. (2019), where Gaussian Process Regression was used to predict force values during a collaborative manufacturing task [132]. In addition, Grigore et al. (2018) trained a Hidden Markov Model (HMM), a model typically trained using unsupervised learning algorithms, in a supervised manner [118].

Although supervised learning (SL) is a viable approach for developing models based on probability distributions, most research in this area focuses on deterministic models. One widely used deterministic model is the artificial neural network (ANN), which consists of feedforward layers of perceptrons that compute a function of weighted input sums. The connections between neurons in ANNs are trained through the backpropagation algorithm [133]. ANNs are typically composed of an input layer, a hidden layer, and an output layer, which can limit their ability to account for time dependencies, as their memory is primarily based on input dimensionality. Interestingly, in the context of Human-Robot Collaboration (HRC), ANNs are frequently used to produce output variables related to platform actuation in the continuous time domain from a control system perspective [134]. ANNs have been employed to produce coefficients of a Lagrangian control system in Chen et al. (2020) [133] and the variable to be minimized in Lorenzini et al. (2018)

[135]. Additionally, ANNs are sometimes combined with fuzzy logic [136, 128, 134]. In a less conventional approach, [129] used an extreme learning machine to train ANNs, which includes neurons that do not require training.

In the realm of supervised learning for human-robot collaboration (HRC), deep learning models are widely utilized, comprising nearly half of the papers within the selected set. These models possess a structural advantage in separating the hyperspaces related to distinct classes in a machine learning problem. Recurrent neural networks (RNNs) are a particular class of deep learning models capable of effectively incorporating time dependence. During the training phase, the output is fed back into the network as part of the input for the subsequent iteration, a process known as backpropagation through time. This enables the model to learn time dependence by utilizing the output from the previous input as the current input, thus retaining memory of the first input even after processing an infinite number of samples. In Zhang et al. (2020) [137], this technique was employed to process sequences of motion frames of the user during experimental validation using multiple cascaded RNNs, a typical usage of RNNs. In another case, Murata et al. (2020) implemented a time-scaled version of an RNN to account for both slow and fast dynamics of the collaboration [138].

Long short term memory (LSTM) networks belong to a specific category of RNNs that feature nonlinear elements in their structure. This characteristic sets them apart from standard RNNs and enables them to capture long-term dependencies [128]. Similarly to RNNs, cascaded LSTMs are also frequently used in the literature [139]. Additionally, there is a trend in cascading LSTMs with other machine learning models that are trained using either unsupervised learning (UL) or reinforcement learning (RL). In the former case, the LSTM network receives processed information from a model that cannot incorporate time dependencies [121]. In the latter case, the LSTM network extracts high-level information, such as the user's intention [128, 130], which is then used to adjust the agent's observation.

In addition, some studies in HRC have also explored the use of convolutional neural networks (CNNs), albeit to a lesser extent. CNNs are neural networks designed specifically for processing raw images or input data composed of measurements from multiple sensors. While each pixel of an image can be considered a separate input, capturing time dependence in CNNs can be challenging. To address this issue, multiple images must be concatenated as a single input. However, the capacity to incorporate time dependence is often limited by the input dimensionality. For example, in Ahmad et al. (2020), a CNN was utilized to track the position of a target object captured through a camera [140]. Si-

milarly, in Chen et al. (2020), a CNN was used to process electromyography data collected from multiple sensors worn by the user [141].

A subset of studies focuses on using a dynamic neural system (DNS) to implement the reasoning block of a robotic system. A DNS is a dynamic system that simulates a network of real neurons, allowing for the creation of machine learning models that closely mimic the firing dynamics of actual neurons. While traditional neural networks take inspiration from the way neurons work, they do not accurately reflect their dynamics. In contrast, DNS models can be used to incorporate time dependence as they are time-variant systems. However, DNS models are not machine learning models per se and their training methodology depends on the specific application. DNS models are inspired by neuroscience, and the most commonly used equation in HRC applications is the Amari equation [142, 143].

While general interactive models have proven to be incredibly powerful for analyzing and learning from large and complex datasets, there are still limitations to what they can achieve without human input. To address this, interactive machine learning has emerged as a field that incorporates human feedback into the learning process. With IML, humans and machines can work together to achieve a common goal, such as building more accurate and generalizable models of human cognition. By allowing humans to provide feedback on the performance of machine learning models, IML can help to identify errors and biases, and improve the overall quality of the models. Therefore, we will be presenting an extensive state-of-the-art review of IML in the next section.

1.2.2 Interactive Machine Learning

Overview

Interactive machine learning (IML) is a subfield of machine learning that involves the development of algorithms and systems that can learn from user feedback. In the context of human-robot interaction (HRI), IML can be used to improve the performance of robots by allowing them to learn from their interactions with human users [144, 25, 145, 146, 147, 148, 149, 150]. The conception of IML dates back to the emergence of query learning where queries are used to learn unknown concepts [151, 152, 153, 154, 155]. A model building component in IML framework interacts with Oracles by issuing queries for additional training data or feedback against its intermediate results. IML based methods mainly aspires to build robust [156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167,

168, 169] models by trading-off accuracy for trust [170, 171, 172, 173, 174, 171, 147, 175, 176, 177, 178, 179, 180, 181] and low resource learning [182, 183, 165, 184, 185, 145, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 144].

IML is becoming the center of machine learning research [146, 144] as the need to operate in resource constrained environments and engaging the human-in-the-loop increased over time. Various IML systems have been designed and implemented to alleviate the long-standing setbacks of standard machine learning algorithms. Its impact has been studied in health [147, 197, 153, 198, 185, 199, 200, 201], agriculture [202, 203, 204], finance [199, 205], politics and sociology [188, 206], military and robotics [171, 176, 207], cyber security [154, 156, 157, 158, 161, 162, 168, 169, 208, 209], education [210, 211, 212] and game and entertainment [183, 195, 192, 194].

Moreover, IML has been analysed and built over the predominant standard machine learning algorithms such as SVM [170, 213, 214, 215, 216, 201, 213], Genetic Algorithms [171, 217, 218, 219, 220, 221], Ant Colony [222, 223, 224] and some other algorithms [201, 183, 155, 225, 226]. Similarly, its capacity to uncover the details behind black-box models is presented in various research works [171, 170, 178, 179, 166, 181]. Most research focus ranges from extending standard machine learning models with interactive capabilities to formulating robust performance evaluation mechanisms.

Similarly, IML has an invaluable contribution to the development of human-robot interaction. As it is known, robots are designed to interact with humans in a naturalistic and intuitive way. However, designing robots that can understand human behavior and respond appropriately is a difficult task, as human behavior can be complex and difficult to predict. This is where IML comes in - by allowing users to provide feedback and corrections to the system, we can improve the robot's ability to understand and respond to human behavior. Overall, IML is a powerful tool for improving the performance of robots in HRI scenarios. By allowing robots to learn from their interactions with human users, we can create more naturalistic and intuitive interactions that are better suited to the needs of individual users.

Hence, in the following few sections, we will be discussing the state-of-the-art of interactive machine learning and thoroughly present research works by creating a merit-oriented taxonomy. While our main focus is on discussing IML contributions for robust, trustworthy and low resource machine learning, we have also presented issues related to IML performance evaluation that might implicate performance measurements in HRI. In section 1.2.2, we discuss about the overall merit-oriented architecture of IML. In sec-

tion 1.2.2, adversarial attacks against black-box machine learning models are discussed. Moreover, it presents various IML inspired contributions aiming to combat adversarial attacks. We present researches conducted to increase the explainability and interpretability as a way to achieve trustworthy machine learning in section 1.2.2. Researches that employ human-in-the-loop for data processing and model building with the intention of lowering the data and computational resources are discussed in section 1.2.2. In section 1.2.2, we discuss the subjective nature of IML and related efforts to address problems in evaluating it. Further analysis of findings and discussion is presented in 1.2.2. In the last section, a summary of the state-of-the-art review and potential research opportunities that are inspiring for future work on IML are discussed.

Merit-Oriented Taxonomy of IML

IML inspired contributions can be analysed from various perspectives. However, architectural [227, 228, 229] and application/sector oriented [146, 204, 198] analysis has been significantly used techniques in the past. However, according to the meta-analysis [229, 227, 198, 204, 228, 146, 25, 149, 230] we conducted, most survey papers lack extensiveness and inclusiveness, leaving the various researches that constitute the state-of-the-art of IML untouched. Consequently, the employment of IML has been constrained to only predefined sectors or limited scope of the architecture.

To this end, we have thoroughly analysed the recent IML-inspired research works using merit-oriented taxonomy. After an extensive review of IML-inspired literature and untapped problems, we have categorized contributions into Robust Machine Learning [156, 157, 158, 159, 160, 161, 162, 163, 164, 166, 167, 168, 169], Trustworthy Machine Learning [170, 171, 172, 173, 174, 171, 147, 175, 176, 177, 178, 179, 180, 181], and Low Resource Machine Learning [182, 183, 165, 184, 185, 145, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 144] based on their merit in the solution space. Besides, performance evaluation techniques that face bias because of the subjective nature of IML are analysed [148, 231, 166, 232, 233, 234]. A high-level notional representation of merit-oriented architecture of IML is depicted in Figure 1.2. Our state-of-the-art review is channeled under these major merits of IML.

Robust Machine Learning Deep learning algorithms are known for their vulnerability to adversarial attacks [235]. Adversaries can craftily manipulate legitimate inputs, which may be imperceptible to human eye, but can force a trained model to produce incorrect

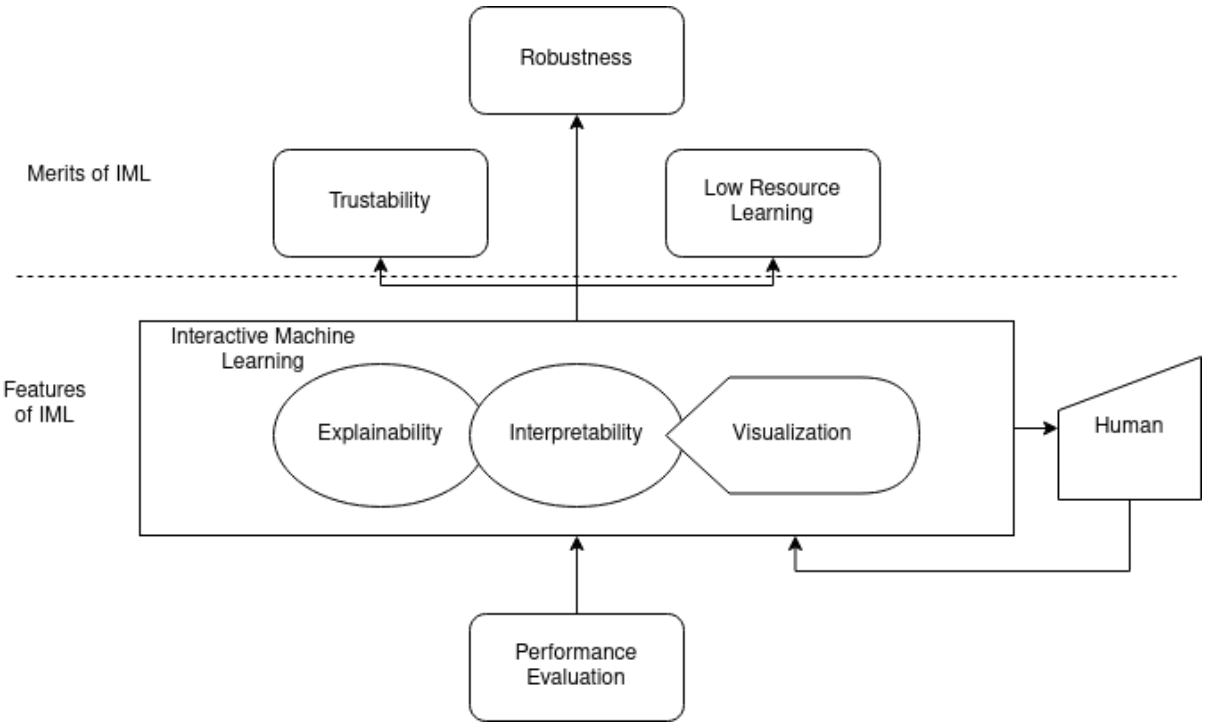


FIGURE 1.2 – Merit-Oriented Architecture of Interactive Machine Learning

outputs. This issue is directly related to the black-box and intricacy nature of deep learning [157, 158, 159, 160, 236, 237, 156].

We use [238] and [239] threat model to highlight the state-of-the-art of the adversarial attack. According to [238], adversarial threat model is comprised of the following three dimensions, namely : the attack surface, adversarial capabilities, and adversarial goals. The attack surface refers to the overall data processing pipeline of machine learning from input to output and then action. Evasion attack [240, 241, 242], poisoning attack [243, 244, 245] and exploratory attacks [157, 161, 162] are the main scenarios considered under the attack surface. The aforementioned attacks can further be dissected into training phase [243, 244, 245] and testing phase [240, 241, 242] attacks from the adversarial capabilities point of view. Data injection, data modification and logic corruption are among the strategies of training phase adversarial capability attacks. On the other hand, white-box [246, 247] and black-box [248, 249] attacks are among the testing phase capability attacks. Adversarial goal attack, on the other hand, infer adversary from the incorrectness of the target model. It is also further classified as confidence reduction, misclassification, targeted misclassification and source/target misclassification.

Adversarial attacks have real impacts on the robustness of a deep learning and other

standard machine learning methods [250]. Therefore, exploring potential adversarial attacks and building a robust machine learning has been the focus of machine learning researchers.

There are various techniques being studied to address the problem of adversarial attacks. Researchers in the IML domain derived a number of strategies both to showcase the impact of adversaries [251] and tackle [168, 252] adversarial attack strategies discussed in Papernot et al. (2016) [238]. The fundamental assumption is that assuring the explainability and interpretability of black-box models by having the human-in-the-loop reduce the vulnerability of machine learning models to adversarial attack [253].

IML has been used as a way to explain and explore model vulnerabilities to adversarial attacks as in Ma et al. (2019) [166] and Das et al. (2020) [167]. Specifically, Ma et al.(2019) enables exploration and explanation of model vulnerabilities to test-phase or poisoning adversarial attacks from the perspective of models, data instances, features, and local structures [166]. A white-box exploratory attack based approach, [167], interactively visualizes neurons and their connections inside a DNN that are strongly activated or suppressed by an adversarial attack. Massif provides both a high-level, interpretable overview of the effect of an attack on a DNN, and a low-level, detailed description of the affected neurons.

In addition to explaining potentially perturbated inputs and models, IML based methods such as [168] propose frameworks that masks the discriminatory biases of black-box classifiers. This plays a vital role to compensate the effects of perturbated inputs on a given model.

In addition to letting the user detect potential adversarial attacks and managing its impacts, IML has also been used to directly engage users in the model building process. This helps to avoid both training phase and test phase adversarial attacks as the user will be there validating inputs and intermediate results. A graph based framework, [169], where user feedback is represented as edges and nodes are the models is a good example of such applications. It proposes Angluin's equivalence query model [152, 254] and Littlestone's online learning model [255] based on a robust machine learning framework that interactively learns models like classifiers [152, 254], orderings/rankings of items [256, 257], or clusterings of data points [258, 259, 260]. In each iteration, the algorithm proposes a model, and the user either accepts it or reveals a specific mistake in the proposal. The feedback is correct only with probability $p > 1/2$ (and adversarially incorrect with probability $1 - p$), i.e., the algorithm must be able to learn in the presence of arbitrary

noise.

Interactive machine learning has also been applied to tackle adversarial attacks on automatic speech recognition (ASR) models. For instance, [251] designed a tool allowing experimentation with adversarial attacks and defenses on ASR. It implemented IML techniques and psychoacoustic principles to effectively eliminate targeted attacks. A similar and cloud-based tool to evaluate and compare state-of-the-art adversarial attacks and defenses for machine learning (ML) models is proposed in Das et al. (2019) [261].

Trustworthy Machine Learning To design and develop AI-based systems that users and the larger public can justifiably trust, one needs to understand how machine learning technologies impact trust [262]. The trustworthiness of AI-based systems is directly related to how the user is confident about the decisions made by the machine learning components. This may include its perception about both the intelligent model and knowledge [263].

The users' understanding of a given domain may have a negative impact on the trustworthiness of a model developed within that specific domain. As stated in Honeycutt et al. , providing feedback can influence users' comprehension of an intelligent system and its accuracy [264]. User trust in a particular model, as well as their perception of its accuracy, may decrease, regardless of whether the system's accuracy improves in response to their feedback. However, the trustworthiness of a model should be defined by considering its relative ability, benevolence, and integrity.

To this end, we use the Trust Antecedent (TA) framework [265] definition of trust. According to this framework a trustee is given trust if it is perceived to have the ability, benevolence, and integrity toachive the desired goal. Nevertheless, a black-box machine learning model that fails to reveal its implementation details cannot align with the principles of this framework, thereby compromising its trustworthiness.Besides, the black-box nature of these models significantly affects their applicability in in decision sensitive areas like health, finance, autonomous vehicles, criminal justice, etc.

Consequently, enhancing the trustworthiness of machine learning algorithms comes with making the model building process interactive [171, 172, 198]. Put succinctly, increasing the explainability and interpretability of machine learning by engaging user-in-the-loop increase its trustworthiness.

To this end, various explanatory frameworks that shows the implementation details of the model building process have been introduced. The essence of most exploratory algorithms is that intermediate results corresponding to a batch or even tuple are made

subject to user-feedback whenever there is a variation between the predicted and the actual label.

In important contributions like [170, 178, 174], the IML is set to interactively query labels whenever the intermediate model fails to predict the label for each data point. For every prediction, the model provides the predicted label and explanations for the prediction. However, the way the the model and feedback are could be visual or textual.

Visual explainer systems are designed considering a set of guidelines proposed in various research contributions. For instance, [181] propose a set of guidelines for integrating visualization into machine learning algorithms through a formalized description and comparison. Specific to automated iterative algorithms, which are widely used in model optimization, these researchers recommend exposing APIs for visualization developers. Using high-resolution APIs, visualization developers access the internal iterations for a tighter integration of the user in the decision loop.

Accordingly, [170] propose a framework called CAIPI is proposed. It uses a model-agnostic explainer, LIME [173], as fundamental component to compute the explainer and present them to the user as interpretable (visual) artifacts. Similarly, Van et al. (2011) propose a natural visual representation of decision tree structures where decision criterion is visualized in the tree nodes [178]. BOOSTVis [179] and iForest [180] also focus on explaining tree ensemble models through the use of multiple coordinated views to help explain and explore decision paths.

Models and feedback from users can also be represented using textual-explanatory systems. An argument based explanatory IML algorithm [174] propose a framework that exploits the usability of arguments to precisely axiomatize feedbacks both from the user and system side. It practically shows how to narrow the gap between domain experts and the machine learning model by engaging domain experts. Users provide feedbacks using a pair of reasons and outputs called arguments. Since arguments are generally presumptive and cannot be untaken for a general set of predictions, the authors introduced prediction-level argument based explanations of decisions made by the learning model. Put succinctly, the training module generates initial features to perform prediction or classification. Whenever the learner notice major deviation between the desired and actual prediction, it consults the domain expert to provide feedback on the output (both the prediction and explanation). At this stage, the domain expert would be able to see the problem either in the predicted label or the rule yielding the outcome. Then, the user provides a set of arguments for each critical example (predictions with problem) and let

the model to retrain keeping the feedback given from the user. Moreover, the paper dictated how the user should communicate with the machine learning component through a sequence of steps namely : Selection, Presentation, Argument formulation, Counter example detection, refinement, argument pruning. Another goal goal-oriented safe exploration algorithm that provably avoids unsafe decisions in real world problems is proposed in Turchetta et al. (2019) [175]. This framework takes suggested decisions as input and exploits regularity assumptions in terms of a Gaussian process prior in order to efficiently learn about their safety.

Exploratory machine learning has been applied to various decision sensitive sectors such as autonomous vehicles, health, and military. In Robert et al. (2017) used exploratory components to build trust among autonomous robots [171]. Specifically, they analyzed the importance of using interpretable machine learning (iML) with neuro-evolutionary algorithms. They conducted two separate tests on two autonomous environments : one with an iML-empowered machine learning algorithm, and the other with black-box machine learning algorithms. The major problem they selected for this experiment is Search. Participants in this research engaged in building IML based search plan models. Finally, they were allowed to see and distinguish between the plans as IML generated or black-box generated. To this end, participants were able to choose IML based search plans for its effectiveness but with less trust relative to the black-box algorithm. The authors justified the preliminary knowledge of participants as the reason for their low degree of trust on the IML algorithm. In conclusion, the researchers convey the importance of using IML based neuro-evolutionary algorithms for searching problems of autonomous robots.

The application of IML in the health sector has also alleviated major computational trust related problems [147, 266]. A biological image data analysis tool called ilastik [266] propose an IML framework that address the speed and usability requirements of machine learning by reasonably compromising model accuracy. Ilastik is able to formulate a feature space without the need to use bulk data as the other classical machine learning algorithms.

Low Resource Machine Learning Low resource machine learning is a process of building an analytical model employing optimal resource utilization techniques. However, most machine learning algorithms have tradeoff between accuracy and resource utilization. As it is stated in Preuveneers et al (2020), the most accurate model might be prohibitively expensive to computationally evaluate on a resource constrained environment [184]. Consequently, the problem of building accurate and high performance machine learning

models has been achieved at the expense of resource (data and computing) utilization. Nonetheless, sufficient amount of data and computing resources are not always at the stake. Besides, some problems may also require to be run on low resource setting. For instance, the use of pervasive devices and robots to build model is one valid scenario.

To this end, interactive machine learning help scientists and engineers exploit more specialized data within their deployed environment in less time, with greater accuracy and fewer costs Porter et al. (2013) [165]. Various IML inspired researches has been conducted in the past couple of years. In this section, we will be discussing IML inspired solutions that aim at enabling low resource machine learning and improving the speed of models.

We have categorized our analysis of IML for low resource learning into small data machine learning and pervasive machine learning. In the small data machine learning section, we discuss contributions that use small data to build models in a low resource environment. On the other hand, the application of IML in pervasive low resource environment is discussed in the pervasive machine learning section.

Small Data Machine Learning Machine learning algorithms usually requires a large volume of data in order to yield accurate result [185]. However, big data is not always at stake to be used in some problem domains like under-developed languages, clinical trials, biomedical science and etc. Achieving small data machine learning requires optimal utilization of data. Engaging domain experts in the model building process in an interactive way would result in optimal utilization of important data items in a way contributing to the development of accurate models with small data.

To this end, interactive machine learning has a critical role. Hence, it has been studied by various researchers in the area of machine translation [267, 182, 268, 269, 270, 271, 272], computer vision, search engine, social network analysis, music and games, and in data constrained sectors like health.

As it is mentioned above [267, 182, 268, 269, 270, 271, 272], IML based algorithms defines the state-of-the-art of machine translation systems. In such kind of systems, the knowledge of a human translator is combined with a MT system. For instance, [182] propose a resource constrained machine translation. This paper explored active learning as an efficient way to reduce costs and make best use of human resource for building low-resource machine translation systems. Specifically, the author extended the traditional active learning approach of single annotation optimization to handle cases of multiple-type annotations. Furthermore, it show further reduction of costs in building low-resource

machine translation systems. Reduction of the required annotated data inherently enabled low resource machine learning. Similarly, various classical IMT researches has been conducted in the few consecutive years. For instance, [267] propose an interactive online machine translation system that avoids the use of batch learning. In Gonzalez et al. (2012), an AL framework for interactive machine translation specially designed to process data streams with massive volumes of data is presented. Recently, a couple of researches have also began to incorporate contemporary machine learning techniques for iMT [268]. In Peris et al. (2017) [269] and Lam et al. (2018) [270], an interactive translation system, based on neural machine translation is deployed. Besides, a full-fledged iMT tools like Huang et al. (2021) has been proposed to further boost accuracy of machine translation systems [272].

Interactive machine learning also plays a significant role in image processing and computer vision in general.The use of IML for automatic feature selection is proposed in Fails et al. (2003) [145]. In this work, the researchers exploit the capacity of IML for automated feature selection. It show how to use IML to eliminate a manual feature selection. The training component of their architecture incorporates the feature selection sub component which further cooperates with the classifier, user feedback manager, and the user as a manual feedback provider. Furthermore, this research introduced an IML image processing framework called Crayons. Crayons is a tool to review and correct classifier errors by using customized decision tree as a learning algorithm and Mean Split Sub Sampled (MSSS) sampling technique.

Search engine optimization is also among the areas that has been studied by IML researchers. In a research contribution by Fogarty et al. (2008), an interactive KNN based desktop application that enables re-rankings on a keyword-based web image search engine relying on the visual characteristics of images is proposed [155].

Another important IML enabled low resource learning is employed for pattern mining research problem. As it is known, pattern mining is an important process in exploratory data analysis. Various tools have been built to mine patterns from large datasets. However, the problem of identifying patterns that are genuinely interesting to a particular user remains challenging as it requires machine learning experts in addition to the domain expert [189]. To this end, IML plays a vital role both in enabling low resource pattern mining and assuring its valuability for specific users. For instance, [188] formulated a tool called ReGroup that provides users with filters that were generated based on features in the model. Put succinctly, ReGroup leverages the IML paradigm to assist users in creating

custom contact groups in online social networks. The classifier provides a potential pattern of users (filters) in social network that likely qualifies for membership in a custom group that is being created. The users are set to provide feedback on each recommended filters where a selection of a contact for inclusion is taken as a positive sample and the remaining are skipped as negative samples. Likewise, the user refines the behaviour of the model with incremental improvements derived through iteratively applied user inputs. Participants noted that these filters provided insight in the patterns that were being exploited by the model, and thus served the dual purpose of explaining the model as well as their intended function as an interaction element. Besides, another research [189] use IML to formulate a pattern mining framework to mine user specific patterns. In this framework, the user is asked to rank small sets of patterns while a ranking function is inferred from this feedback by preference learning techniques. With a user supplemented by active learning heuristics, the resulting accurate rankings of patterns are used to mine new and more interesting patterns. They demonstrated their framework in frequent item-set and subgroup discovery pattern mining tasks. The ability of the framework to learn patterns accurately and the importance of its heuristic approach to ease users effort is presented precisely.

IML has also been applied to various game and entertainment research problems. A sketch-rnn, [226], present an interactive recurrent neural network (RNN) that helps to construct stroke-based drawings of common objects. Besides their model encode existing sketches into a latent vector, and generates similar looking sketches conditioned on the latent space. Similarly, a research work in Jain et al. (2020) emphasizes the design of a deep reinforcement learning agent that can play from feedback alone [191]. This algorithm takes advantage of the structural characteristics of text-based games. Moreover, the application of IML in motion-driven music systems is presented in [192, 193, 194]. Due to the complications to design a robust player-recognition or motion recognition system using standalone IML system, an enhanced model has been presented in Diaz et al. (2019) [195]. They proposed an IML solution for Unity3D game engine in the form of a visual node system supporting classification (with k-nearest neighbour), regression (with multi-layer perceptron neural networks) and time series analysis (with dynamic time warping) of sensor data.

The health sector is also known for data scarcity. In this sector, researchers are often confronted with only a small number of datasets or rare events, where a machine learning algorithm suffers from insufficient training samples [153, 198]. To this end IML researches has been conducted to build resource constrained model that adheres to the

trustworthiness principles. In Holzinger et al. (2018), an interactive Ant Colony Optimization (ACO) technique is applied to the Traveling Salesman Problem (TSP) [147]. The TSP is an intransigent mathematical problem to find the shortest path through a set of points and returning to the origin. It is among the problems that resembles most computational problems in the health informatics sector. They introduce two novel concepts : Human-Interaction-Matrix (HIM) and Human-Impact-Factor (HIF) to control the ants action and the variable interpretation of the HIM respectively. The authors implemented their framework using a java-script based browser solution which has a great benefit of platform independence and configuration. Although the results they found are promising, the researchers suggested further works like such as gamification and crowdsourcing to solve hard computational problems.

As we have discussed in the first section of this document, the essence of IML is to leverage the benefits of the human-in-the-loop. However, the presence of the human does not by itself guarantee the trustworthiness and performance of the model. The interaction between the user and the learning model should be supplemented with clear and in-depth visualization methods. To this end, [186] emphasize on how to design effective end-user interaction with interactive machine learning systems. They shape the human-model interaction in a way it answers these critical questions : which examples should a person provide to efficiently train the system, how should the system illustrate its current understanding and how can a person evaluate the quality of the system's current understanding in order to better guide it towards the desired behavior. Another practical contribution for human-model interaction is provided by [190]. They propose a new visual analytic approach to IML and visual data mining. Multi-dimensional data visualization techniques are employed to facilitate user interactions with the machine learning and mining process. This allows dynamic user feedback forms (data selection, data labeling and correction) to enhance the efficiency of model building. In particular, this approach can significantly reduce the amount of data required for training accurate models. Hence, it can be highly impactful for applications where large amount of data is hard to obtain. The proposed approach is tested on two application problems : the handwriting recognition (classification) problem and the human cognitive score prediction (regression) problem. Both experiments show that visualization supported IML and data mining can achieve the same accuracy as an automatic process can with much smaller training data sets. Furthermore, [196] propose a scalable client-server system for real-time IML framework called Computer-Human Interaction for Semi-Supervised Learning (CHISSL) [144]. The proposed system

is capable of incorporating user feedback incrementally and immediately without a predefined prediction task. The light-weight computation web-client and heavyweight server constitute their architecture. The server relies on representation learning and off-the-shelf agglomerative clustering to find a dendrogram, which we use to quickly approximate distances in the representation space. The client, using only this dendrogram, incorporates user feedback via transduction. This work achieves low resource learning as it updates distances and predictions for each unlabeled instance incrementally and deterministically, with O(n) space and time complexity. Ultimately, this paper solved the scalability issue of CHISSL. Another important contribution in this regard is that of [213]. In this paper, EnsembleMatrix, SVM based visualization system, to tune classifier through user in the human-in-the-loop approach to build a superior model. This product presents a graphical view of confusion matrices to help users to directly interact with the visualization in order to explore and build combination models.

Pervasive Machine Learning Interactive machine learning can also yield a better performance in pervasive computing environment where there are low computing resources. For instance, [183] present techniques to enable low resource computation for deep reinforcement learning agents complex behaviors in 3D virtual environments. Specifically, they considered an environment with high degree of aliasing, MineCraft, and conducted experiments with two reinforcement learning algorithms which enable human teachers to give advice-Feedback Arbitration, and Newtonian Action Advice under visual aliasing conditions. Similarly, Preuveneers et al. (2020) emphasizes on the importance of hyper tuning as a process to select the best performing machine learning model, its architecture and parameters for a given task [184]. The fact that hyper parameter tuning does not take into consideration resource trade-offs when selecting the best model for deployment in smart environments has been taken as their research challenge. Consequently, the authors proposed a multi-objective optimization solution to find acceptable trade-offs between model accuracy and resource consumption to enable the deployment of machine learning models in resource constrained pervasive environments. On the other hand, [273] used pathward and fieldward techniques to enable creation of gestures that can be used by a touch-enabled pervasive mobile devices reliably. They introduced a novel dynamic guide that visualizes the negative space of possible gestures as the user interacts with the system.

On the other hand, [187] emphasises on the gaps of standard machine learning algo-

rithms in dynamic sensor setting. Consequently, an IML based framework is implemented for "activity recognition" problems. This framework takes stream of data among different devices in the internet of things as an input. Three different machine learning approaches were used in the experiments : Support Vector Machine, k-Nearest Neighbor and Naıve Bayes classifier. These algorithms were customized to adapt the dynamic sensory setting. Uncertainty, Error (output from the model), State Change (what change in state urges the user to change label of data), Time (the time difference b/n queries and labeling by user), and Randomness (randomly assigning labels) are identified as factors that increase the subjectivity of users feedback. The results of this research work make it clear that the choice of interactive learning strategy has a significant effect on the performance when tested on recordings of streaming data.

Evaluation

Although there are a number of invaluable contributions towards the goal of IML algorithms, the techniques used to evaluate their performance are not that mature. The standard machine learning algorithms use statistical model performance evaluation techniques which provide only a sole measure, obfuscating details about critical instances, failures and model features [166]. However, the human-in-the-loop nature of IML systems triggers an extended need to diagnose the subjectivity of results. Therefore, IML performance evaluation should consider subjectively generated user evaluations and the cyclic nature of influence between the algorithm and users. This makes the evaluation of IML systems subjective and complex [148, 232].

Consequently, researchers proposed both customized and noble solutions to approach these critical issues. The IML Framework for Guided Visual Exploration (EvoGraphDice) [148] is among the salient ones that is produced after an extensive experimental review of existing evaluation techniques. This paper based mainly on the gaps of user and system feedback evaluation techniques to get insights on the underlying co-operation and co-adaptation mechanisms between the algorithm and the human. EvoGraphDice couples algorithm-centered and user-centered evaluations to bring forth insights on the underlying co-operation and co-adaptation mechanisms between the algorithm and the Human. Other papers focus on the input and output nature of IML models to diagnose their performance. For instance, the INFUSE system [231] supports the interactive ranking of features based on feature selection algorithms and cross-validation performances. Another work [233] also propose a performance diagnosis workflow. In this work, the instance-level diagnosis

leverages measures of "local feature relevance" to guide the visual inspection of root causes that trigger misclassification.

[232] identify diversified scenarios and subjective nature of explanations for the absence of benchmarks to evaluate explanations of IML algorithms. Consequently, they have defined the problem of evaluating explanations and systematically reviewed the existing efforts from state-of-the-arts. The authors discuss explanation as global (overall working structure) and local (particular model behaviour for individual instances) from scope perspective and as intrinsic (self-interpretable models) and posthoc (independent interpretation model) from dimension perspective. Furthermore, generalizability (generalization), fidelity (degree of exactness) and persuasibility (comprehensibility) are identified as the three properties of explanation and corresponding methods has been revised for each aspect. As a result, they designed a unified evaluation framework based on the discussed properties of explanation and to the hierarchical needs from developers and end-users. In this work, three tiered explanation evaluation architecture is proposed corresponding to the three properties of explanation.

A similar work [234] discusses the importance of enhancing the quality of AI-based systems for a practical usage. They affirmed the importance of quality assurance to set benchmark for evaluations. Understandability and interpretability, defining expected outcomes as test oracles, and non-functional properties of AI-based systems of AI models are among the seven challenges they listed to assure quality of AI-based systems.

Discussion

An extensive merit-oriented state-of-the-art review of IML is presented in this section. We used a bottom-up approach to categorize researches based on their primary importance. As shown in Table 1.3, robust machine learning, trustworthy machine learning, and low resource machine learning are identified as the major themes to categorize research works contributing to the state-of-the-art of IML. Significant issues related to IML evaluation has also urged us to consider it as one of the defining factors of the state-of-the-art of IML.

For the ease of readability and document organization, we will not be presenting the discussion part of our state-of-the-art in this document. However, we advice our readers to look in to [274] for extended version of our discussion on the state-of-the-art of interactive machine learning.

In conclusion, we have seen how IML has gained significant importance in the field of

Merit-Orientation	Technical Section								Sectoral Section				Others
	HCI and Interface design	Visual Analytics	Searching and Retrieval	Security and Privacy	Information Processing	Model Explainability	Pervasive comp. and Robotics	Clustering and Optimization	Agriculture	Health	Education	Game and Entertainment	Others
Baselines and Meta Reviews	[2, 5, 31, 54, 83, 88]	[20, 64, 80]		[19, 97]	[111]	[116]			[78]	[54, 83]			[6, 36, 41, 119, 130]
Robustness	[26, 128]	[33, 82]	[39]	[14, 16, 18, 19, 21–23, 27, 28, 43, 49, 50, 60, 61, 72, 73, 89, 91, 95, 96, 106, 109, 113, 120, 121, 123, 135]	[1, 40, 48, 59, 76, 94, 98]	[107, 112]							
Trustworthiness	[13, 45, 47, 51, 57, 58, 71, 93, 131, 134]	[70, 81, 126, 136]		[67, 105, 124]	[25, 111]	[45, 92, 117, 136]	[12]	[9, 10, 81, 110, 133]		[11, 67, 103, 125]	[53, 127, 127, 137]		[86]
Low Resource Learning	[1, 3, 7, 8, 17, 24, 35, 35, 46, 56, 84, 99, 122]	[77, 114]	[39, 63, 65, 75, 104, 118, 129]	Security	[90, 100]	[38, 71]	[74, 85, 101, 115]	[9, 34, 42, 44, 55, 62, 65, 79, 87, 102, 123]	[29, 38]	[55]		[30, 37, 42, 52, 62, 66, 87, 108]	[4, 32]
Performance Evaluation	[15]	[68, 82]		[18, 132]	[69]								[129]

FIGURE 1.3 – Summary of IML Research Works

human-robot interaction by enabling intelligent systems to learn from human interactions and adapt their behavior accordingly, making them more responsive to the needs and preferences of the users. One of the key advantages of IML is that it enables robots to learn in real-time, as they interact with human users. This means that robots can adapt their behavior on-the-fly, based on the feedback they receive from the user. This is particularly important in HRI, where the robot's ability to understand and respond to human behavior is critical to the success of the interaction. Another key advantage of IML is that it enables robots to learn from a wide range of users with different backgrounds and preferences. This is important because robots are increasingly being used in a variety of settings, from homes and offices to hospitals and factories. By learning from a diverse set of users, robots can become more versatile and adaptable, making them better suited to a range of different tasks and environments.

However, for IML and other interactive models to be effective in HRI, it is important to have a good understanding of human attention. Attention models are cognitive models that describe how humans allocate and maintain their attention during interactions. These models are critical for designing effective HRI systems because they help us understand how humans perceive and process information, and how they prioritize different stimuli. These models enable robots to learn from the attentional behavior of the user. By tracking

the user's gaze, body posture, and other cues in the environment, the robot can adapt its behavior to better engage with the user and improve the quality of the interaction. Therefore, in the next section, we will be discussing about state-of-the-art of human attention models.

1.2.3 Anthropomorphic Attention Models

Interactive models that are designed to simulate human attention have become increasingly important in the field of robotics. These models have the ability to predict saliency and detect moving objects, which is particularly useful in human-robot interaction settings. The development of these models has been driven by the need for robots to be able to navigate complex environments and interact with humans in a more natural and intuitive manner.

Saliency prediction is a key component of human attention modeling. It involves identifying the most important features of an image or scene that will attract a human's attention. This is based on the concept that humans tend to focus on objects that stand out from their surroundings, such as bright colors, high contrast, or unique shapes. Saliency prediction models use various computational techniques, such as bottom-up and top-down processing, to identify these important features and prioritize them for attention.

Moving object detection and segmentation is another important aspect of human attention modeling. This involves identifying and tracking objects that are moving within a scene. It is particularly important in human-robot interaction settings, as robots need to be able to track and respond to human movements in real-time. Moving object detection and segmentation models use various techniques, such as optical flow and background subtraction, to identify and track moving objects.

In human-robot interaction settings, interactive models that simulate human attention can be used in a variety of ways. For example, they can be used to help robots navigate complex environments by identifying the most important features in a scene and prioritizing them for attention. This can help robots avoid obstacles and navigate around them more effectively. Similarly, saliency prediction models can be used to help robots identify objects of interest in a scene, such as people or specific objects, which can be useful in a variety of applications, such as search and rescue or surveillance. Moreover, moving object detection and segmentation models can also be used to help robots interact with humans more effectively. For example, they can be used to track and respond to human

movements in real-time, which can be useful in a variety of applications, such as sports coaching or rehabilitation. Similarly, these models can be used to track and respond to objects that are being manipulated by humans, which can be useful in manufacturing or assembly applications.

Overall, interactive models that simulate human attention are becoming increasingly important in the field of robotics. These models have the ability to predict saliency and detect moving objects, which is particularly useful in human-robot interaction settings. As the field of robotics continues to evolve, these models will become even more important, as they will help robots navigate complex environments and interact with humans in a more natural and intuitive manner.

Hence, in this part of the chapter, we emphasize more on the state-of-the-art of anthropomorphic attention models that empowers robots with efficiency and intuitive behaviour. We focus on the state-of-the-art of saliency prediction and moving object detection and segmentation interactive models.

Saliency Prediction Models

Saliency prediction models have emerged as a promising tool for enhancing human-robot interaction. These models can predict the areas in a visual scene that are most likely to attract human attention, allowing robots to focus on the most relevant information and interact with humans in a more natural and intuitive manner.

The application of saliency prediction models in human-robot interaction is multifaceted. In navigation tasks, these models can assist robots in identifying the most important features of a scene, such as obstacles, landmarks, and signs, which are crucial for safe and efficient navigation. By prioritizing the relevant areas in a scene, robots can avoid collisions and follow optimal paths. Saliency prediction models can also be used to enhance human-robot communication. By identifying the areas in a scene that humans are likely to attend to, robots can direct their attention towards the same areas, facilitating effective communication. For instance, in a healthcare setting, a robot may use saliency prediction to identify the body part that a doctor is examining, allowing it to provide additional information or assistance.

Moreover, saliency prediction models can be integrated into human-robot collaborative tasks. In a manufacturing setting, for example, a robot may need to identify the object that a human worker is manipulating in order to provide appropriate support. By predicting the salient areas of the scene, the robot can determine the object of interest and adjust

its actions accordingly. In addition, saliency prediction models can be used to personalize human-robot interaction. By learning the specific attentional biases of an individual user, robots can tailor their behavior to better engage the user. For instance, a robot may adjust the speed of its movements or the timing of its responses based on the user's gaze behavior.

Recent studies in saliency prediction have been continuously improving the state-of-the-art in this field. Early saliency models were primarily constructed for still images, assuming that visually striking features would naturally attract attention [13]. However, these models have been found to have limited performance, as they overlook the importance of temporal features. As a result, modern visual saliency prediction techniques incorporate dynamic features, which have been made possible by advancements in deep learning and the availability of larger video saliency datasets. This section will provide a brief review of the current visual saliency prediction models that are considered to be the best in the field. We start with various saliency prediction models and close our review by discussing the various saliency prediction datasets.

Saliency Models Researches on human gaze fixation prediction or video saliency prediction is dating back to [275, 276]. The earliest saliency prediction methods are based on various low-level manual features of still image, such as color contrast, edge, center prior and orientation to produce a "saliency map" [277, 278, 279, 280, 281, 282, 283]. A saliency map is an image that highlights the region on which human gaze could focus on a various probabilistic level.

Low-level feature based saliency models can work robustly on the simplest detection tasks. However, these models fail to perform well on a more complex image structures. To this end, various deep learning based static saliency researches are published Hou et al. [284], Lee et al. [275] and Li and Yu [276] Wang et al. [285] and Zhang et al. [286] [287, 288, 289, 290, 291, 292]. These models have achieved a remarkable result using the powerful learning ability of neural networks and growth in the size and quality of visual saliency datasets [289].

Static image saliency research is almost mature. However, subsequent trials to employ these models on video show a reduced performance [293]. These is mainly due to the frequent change in salient-goal over time in a sequence of frames. Furthermore, convolutional neural networks (CNN) have no memory function, so it is difficult to model video frames that are constantly changing in the time domain with CNN.

To this end, dynamic saliency models leverage both static and temporal features to predict human gaze fixation on videos [294, 295, 293, 296, 297, 298, 299, 300, 301]. Some of these studies [294, 293, 297] can be viewed as extensions of existing static saliency models with additional motion features. Conventionally, video saliency models pair bottom-up feature extraction with an ad-hoc motion estimation that can be performed either by means of optical flow or feature tracking. Frame-differencing [302], background subtraction [303], optical flow [304] and other methods are used to model spatial and motion information. However, these techniques are known for poor performance, especially in complex scene videos.

In contrast, deep video saliency models learn the whole process end-to-end. Some of these saliency models treat spatial and temporal features separately and fuse these features in the last few layers of the DNN architecture in certain way. Other researches simultaneously model the time and space information, directly letting the network simultaneously learn the time and space information and ensure the time and space consistency.

Research works that treat spatial and temporal information separately base on two-stream network architectures [305, 306] that accounts for color images and motion fields separately, or two-layer LSTM with object information [307, 308]

As one of the first attempts, [305] study the use of deep learning for dynamic saliency prediction and propose the so-called spatio-temporal saliency networks. They applied a two-stream (5 layer each) CNN architecture for video saliency prediction. RGB frames and motion maps were fed to the two streams. They have investigated two different fusion strategies, namely element-wise and convolutional fusion strategies, to integrate spatial and temporal information.

Jiang et al. (20170) [307] concluded that human attention is mainly drawn to objects and their movement. Hence, they propose object-to-motion convolutional neural network (OM-CNN) to learn spatio-temporal features for predicting the intra-frame saliency via exploring the information of both objectness and object motion. Inter-frame saliency is computed by means of a structure-sensitive ConvLSTM architecture.

Zhao et al. (2019) [306] proposes two modules to extract temporal saliency information and spatial information. Moreover, the saliency dynamic information in time is combined with the spatial static saliency estimation model, which directly produces the spatiotemporal saliency inference. A context-aware pyramid feature extraction (CPFE) module is designed for multi-scale high-level feature maps to capture the rich context features. A channel-wise attention (CA) model and a spatial attention (SA) model are respectively

applied to the CPFE feature maps and the low-level feature maps, and then fused to detect salient regions. Finally, an edge preservation loss is proposed to get the accurate boundaries of salient regions.

Tang et al. (2018) [308] used a multiscale spatiotemporal convolutional ConvLSTM network architecture (MSST-ConvLSTM) to combine temporal and spatial information for video saliency detection. This architecture not only retains the original temporal clues but also uses the temporal information in the optical flow map and the structure of LSTM. This part of the study separately learns the information in the time domain and the space domain through neural networks. Generally, to model the information in the time domain, some preprocessing methods, such as the optical flow method, are used. Additionally, the fusion of features extracted in the time and space domains also greatly affect the performance of the network. These works show a better performance and demonstrate the potential advantages in applying neural networks to video saliency problem.

By employing models that explicitly capture both time and space information, the network is able to learn and integrate these dimensions concurrently, thereby ensuring consistency between them. For instance, in reference [309], the author first used a pyramid dilated convolution module to extract multiscale spatial features and further extracted spatio-temporal information through a bidirectional convective ConvLSTM structure. Ingeniously, the author used the forward output of the ConvLSTM units as input and directly fed it into the backward ConvLSTM units, which increases the capabilities to extract deeper spatiotemporal features.

In reference [310], unlike previous video saliency detection with pixel-level datasets, the author collected a densely annotated dataset that covers different scenes, object categories and motion modes. In Li et al. (2018), the authors proposed a flow-guided recurrent neural encoder (FGRNE) architecture, which uses optical flow networks to estimate motion information per frame in the video and sequential feature evolution encoding in terms of LSTM network units to enhance the temporal coherence modeling of the per-frame feature representation [311].

Chaabouni et al. (2016) [312] employed transfer learning to adapt a previously trained deep network for saliency prediction in natural videos. They trained a 5-layer CNN on RGB color planes and residual motion for each video frame. However, their model uses only the very short-term temporal relations of two consecutive frames. In Bazzani et al. (2016), a recurrent mixture density network is proposed for saliency prediction [313]. The input clip of 16 frames is fed to a 3D CNN, whose output becomes the input to a LSTM.

Finally, a linear layer projects the LSTM representation to a Gaussian mixture model, which describes the saliency map. In a similar vein, Mnih et al. (2014) applied LSTMs to predict video saliency maps, relying on both short- and long-term memory of attention deployment [314].

In Leifman et al. (2017), RGB color planes, dense optical flow map, depth map and the previous saliency map are fed to a 7-layered encoder-decoder structure to predict fixations of observers who viewed RGBD videos on a 2D screen [45].

As in their previous work Gorji et al. (2018), here they used a multi-stream ConvLSTM to augment state-of-the-art static saliency models with dynamic attentional push (shared attention) [315]. Their network contains a saliency pathway and three push pathways including gaze following, rapid scene changes, and attentional bounce. The multi-pathway structure is followed by a CNN that learns to combine the complementary and time-varying outputs of the CNN-LSTMs by minimizing the relative entropy between the augmented saliency and viewers fixations on videos.

In Wang et al. (2018), the attentive CNN-LSTM Network which augments a CNN-LSTM with a supervised attention mechanism to enable fast end-to-end saliency learning is introduced [19]. The attention mechanism explicitly encode static saliency information allowing LSTM to focus on learning a more flexible temporal saliency representation across successive frames. Such a design fully leverages existing large-scale static fixation datasets, avoids overfitting, and significantly improves training efficiency.

Sun et al. (2018) proposed a robust deep model that utilizes memory and motion information to capture salient points across successive frames [316]. The memory information was exploited to enhance the model generalization by considering the fact that changes between two adjacent frames are limited within a certain range, and hence the corresponding fixations should remain correlated.

There are some more salient object detection models [317, 318, 319, 320, 321, 322, 284] that attempt to uniformly highlight salient object regions in images or videos. Those models are often task-driven and focus on inferring the main object, in stead of investigating the behavior of the HVS during scene free viewing.

Video Saliency Dataset The recent advancements in the field of human attention and dynamic fixation prediction have been primarily driven by the availability of improved and extensive saliency datasets [323, 324, 325, 326]. These datasets have significantly enhanced the understanding of human visual attention and have greatly improved the

performance of computational models.

One notable dataset is the DHF1K [19], which provides human fixations on a diverse range of representative dynamic nature scenes observed during free-viewing. It consists of 1K video sequences annotated by 17 observers using eye-tracking devices. Each video in DHF1K has been manually labeled with a category, further classified into seven main categories : daily activity, sport, social activity, artistic performance, animal, artifact, and scenery.

Another important dataset is the Hollywood-2 [325], which offers a collection of 3669 video clips with 12 classes of human actions and 10 classes of scenes, totaling approximately 20.1 hours of video. This dataset serves as a comprehensive benchmark for human action recognition in realistic and challenging settings. Analysis conducted by [327] reveals that 84.5

The UCF Sports dataset [325] comprises a range of sports actions typically featured on broadcast television channels like the BBC and ESPN. The video sequences were obtained from various stock footage websites, including BBC Motion gallery and GettyImages. It consists of 150 videos derived from the UCF sports action dataset [328]. According to [327], 82.3

While there exist other datasets such as [326, 324, 323, 329, 289], they are either limited in terms of stimulus variety and scale or collected for specific purposes (e.g., salient objects in videos [320]). Importantly, none of the aforementioned datasets includes a dedicated test set to prevent potential data overfitting, which has posed significant challenges to the research progress.

Moving Object Detection Models

In the early days of moving object detection, researchers formulated a well established background subtraction techniques for stationary camera setting. These techniques have been extended for many years and are able to successfully detect moving objects as long as the camera is stationary [330, 331, 330, 332, 333, 334, 335]. However, a relatively long initialization time to model the background and residual image alignment error on non-stationary camera setting are the main setbacks of this technique [336]. These problems are incontestably addressed by optical flow based methods [337, 338, 304]. However, optical flow is highly dependent on optical flow vectors. The quality of the these vectors is crucial for the motion segmentation performance. Besides, optical flow is highly complex and due to high sensibility of noise corruption, it cannot meet the need of real time object

detection. Optical flow exceptionally works well on large moving objects and fails to detect smaller objects due to blurry edges and low resolution [339].

A variety of frame differencing techniques address the aforementioned problems. For instance, Inter-frame differencing technique generate the difference between two consecutive frames over a period of time for identifying background and foreground pixels. A research in Liang et al. (2010) use inter-frame differencing algorithm to detect moving target in aviation video [340]. Their experiment indicate that the algorithm has few computations and high accuracy to extract moving-target in aviation videos. In Nakashima et al. (2018), interframe differencing and dynamic binarization using discriminant analysis is applied [341]. The positions of the moving object in the image are determined by observing the histograms of each frame.

A slightly different method with comparable result with that of inter-frame differencing is three-frame differencing. This method put three adjacent frames as a group, subtracts both adjacent frames and lets two differential results do the logical AND operation. This has been the most widely used and traditional three-frame differencing method. In Yin et al. (2016), traditional three-frame differencing technique and W4 algorithms are used to detect foreground objects in the infrared video datasets [342]. Another research, Sengar et al. (2016), propose a moving object detection method under static background [343]. The algorithm use a non-overlapping blocks of the difference frames and calculate the intensity sum and mean of each block.

The inter-frame differencing and three-frame differencing techniques suffer from the foreground aperture and ghosting problems due to slow-moving as well as fast-moving foreground objects. Besides, these methods are known for partial or splitted detection of objects [344]. As a result, frame differencing techniques are prone to false positives and sometimes false negatives.

To this end, combined moving object detection methods improved the performance of frame differencing techniques. Relatively robust moving object detection methods combine background subtraction with frame differencing [345, 346, 347, 348] or optical flow [349, 336, 350].

A static background based on three-frame differencing method in combination with background subtraction method is proposed in Weng et al. (2010) [345] and Cheng et al. (2014) [347]. A combination of optical flow and three-frame differencing based moving object detection method is employed in Halidou et al. (2014) [349]. It uses region of interest (ROI) and multi-block local binary pattern descriptors. Another frame differencing and

optical flow based moving object detection technique is proposed in Fernandez et al. (2010) [350]. Here, a thermal infrared camera mounted on autonomous mobile robot is used as a feed to the detection module.

A method for detecting moving people in the indoor environment is proposed with the help of frame differencing and neural network based classification techniques [351]. This method reduces the false alarm and provides a robust classification with the help of a finite state automation. Similarly, a new approach based on fuzzy adaptive resonance theory, neural network with forgetting method for foreground detection and background establishment in natural scenes is proposed in Dou et al. (2014) [352]. On the other hand, the frame differencing and the non-pyramidal Lucas-Kanade approaches [353] are used to detect human candidates based on thermal signatures when the robot stops and moves.

In Xu et al. (2017), an efficient foreground detection method is proposed by combining three-frame differencing and Gaussian mixture model [354]. Another research work, Lee et al. (2013), presents a moving object detection method by combining background subtraction, separable morphological edge detector, and optical flow [355].

Recently, a more sophisticated and efficient moving object detection methods have been proposed by intriguing improved frame differencing techniques with deep neural network technologies. For instance, [21] propose a deep learning based moving object detection method. It uses a Deep Convolutional Neural Network (DCNN) for multi-modal motion segmentation. Improved three-frame differencing and current RGB frame is used to capture temporal information and appearance of the current scene respectively. These inputs are later fused in the DCNN component for effective, efficient and robust motion segmentation. This model improved the performance of three-frame differencing techniques in detecting tiny moving objects.

A research work in Yang et al. (2017) applied a frame differencing technique with Faster Region-Convolutional Neural Network (R-CNN) for highly precise detection and tracking characteristics [356]. Similarly, Mohtavipouret al. (2022) propose a multi-stream CNN and frame differencing based moving object detection method for deep violence detection Mohtavipour et al. (2022) [357]. It uses a handcrafted features related to appearance, speed of movement, and representative image and fed to a convolutional neural network (CNN) as spatial, temporal, and spatiotemporal streams.

Furthermore, Siam et al. (2018) [358] and Wang et al. (2018) [359] show promising results using CNN for moving object detection. They use a two-stream convolutional network to jointly model motion and appearance cues in a single convolutional network.

In Wang et al. (2018) [359] a new framework named moving-object proposals generation and prediction framework (MPGP) is proposed to reduce the searching space and generate some accurate proposals which can reduce computational cost. In addition,they explored the relation of moving regions in feature map of different layers. This method utilize spatial-temporal information to strengthen the detection score and further adjust the location of the bounding boxes.

1.3 Conclusion

The growth in machine learning technologies and enhanced physical dexterity of the new generation of robotic platforms has paved the way towards advancements in the area of human-robot interaction. Machine learning techniques such as interactive machine learning, and deep learning can enable robots to learn from their interactions with humans and adapt their behavior accordingly. This can enhance the robot's ability to understand and respond to human behavior, and to learn from a diverse set of users with different backgrounds and preferences. Specifically, machine learning models that focus on understanding human attention are also critical for designing effective HRI systems. By tracking the user's gaze, body posture, and other environmental cues, attention models can enable robots to understand how humans perceive and process information, and how they prioritize different stimuli. This can enhance the robot's ability to engage with the user, respond to their needs and preferences, and improve the quality of the interaction. Therefore, any research work intending to advance human-robot interaction should emphasize on the synergy between machine learning techniques in general and interactive machine learning models in particular, and neuroscience of attention.

Consequently, in the first part of this chapter, we provided an extensive state-of-the-art in human–robot interaction. Although the hardware components of such systems are among the most crucial for a successful market entry, we focused on the intermediate interfaces and perception mechanisms for improved human and robot perception. We discussed the state-of-the-art in interactive machine learning models in second section of this chapter. We started by discussing general interactive models by dissecting them under the commonly used categories namely : unsupervised learning, reinforcement learning, and supervised learning. Since interactive models built upon the principles of interactive machine learning take a significant portion of machine learning models for HRI, we have conducted an extensive review of the state-of-the-art in IML. Finally, to align the state-

of-the-art with the scope of our dissertation, we undertook a review of state-of-the-art in anthropomorphic attention models. Specifically, we discussed an extensive review of the state-of-the-art of saliency prediction models and moving object detection and segmentation models which has a direct alignment with what we have been working on in the course of this thesis.

Throughout our state-of-the-art analysis, we have come to realize that human-robot interaction, especially within social settings, still face numerous challenges. One of the challenges of human-robot interaction is understanding how humans allocate their attention when interacting with robots. Human attention is complex and dynamic, and robots need to be designed to understand and respond appropriately to human attentional cues. Developing machine learning models that can predict where humans are likely to look next based on contextual information such as the task being performed or the environment is proofed to enhance the intuitiveness and efficiency of HRI. Human attention models also improve robot autonomy, allowing robots to make decisions about where to focus their attention without explicit human input.

In line with these findings, we exerted enormous effort and time researching human visual attention models in human-robot interaction setting. Our primary contribution is the development of video saliency prediction model using the stacked convLSTM approach. Here, we introduce an encoder-decoder based architecture with a prior layer undertaking XY-shift frame differencing, a residual layer fusing spatially processed (VGG-16 based) features with XY-shift frame differenced frames, and a stacked-ConvLSTM component. Since motion out-weights other low-level saliency features in attracting human attention and defining region of interests, we focused on enhancing moving object detection and segmentation techniques on our second research contribution. Here, we proposed a novel frame differencing technique along with a simple three-stream encoder-decoder architecture to effectively and efficiently detect and segment moving objects in a sequence of frames. Our frame differencing component incorporates a novel self-differencing technique, which we call XY-shift frame differencing, and an improved three-frame differencing technique. We fuse the feature maps from the raw frame and the two outputs of our frame differencing component, and fed them to our transfer-learning based convolutional base, VGG-16.

Finally, we contributed a human attention based anthropomorphic human-robot interaction framework for intuitive and efficient human-robot interaction. Specifically, we propose a video saliency and a moving object detection based anthropomorphic human-robot

interaction framework. We employ a state-of-the-art interactive video saliency prediction and moving object detection and segmentation models to provide robots with humanly perceptive and action intelligence. We employ a Robot Operating System (ROS) framework and its third-party modules to implement our framework on a widely known humanoid robot, Pepper.

In the next consecutive chapters, we will be presenting our research contributions in detail. We start by discussing our contribution on interactive video saliency prediction. Then we discuss about our research in the domain of moving object detection and segmentation. Discussion on our anthropomorphic human-robot social interaction framework follows our extensive presentation of the two attention models. Finally, we close our thesis by presenting results and research opportunities that we believe are inspiring for future work in the intersection of human-robot interaction, interactive machine learning, and human attention modeling.

Chapitre 2

Interactive Video Saliency Prediction : The Stacked-convLSTM Approach

---

Cognitive and neuroscience of attention researches suggest the use of spatio-temporal features for an efficient video saliency prediction. This is due to the representative nature of spatio-temporal features for data collected across space and time, such as videos. Video saliency prediction aims to find visually salient regions in a stream of images. Many video saliency prediction models are proposed in the past couple of years. Due to the unique nature of videos from that of static images, the earliest efforts to employ static image saliency prediction models for video saliency prediction task yield reduced performance. Consequently, dynamic video saliency prediction models that use spatio-temporal features were introduced. These models, especially deep learning based video saliency prediction models, transformed the state-of-the-art of video saliency prediction to a better level. However, video saliency prediction still remains a considerable challenge. This has been mainly due to the complex nature of video saliency prediction and scarcity of representative saliency benchmarks. Given the importance of saliency identification for various computer vision tasks, revising and enhancing the performance of video saliency prediction models is crucial. To this end, we propose a novel interactive video saliency prediction model that employs stacked-ConvLSTM based architecture along with a novel XY-shift frame differencing custom layer. Specifically, we introduce an encoder-decoder based architecture with a prior layer undertaking XY-shift frame differencing, a residual layer fusing spatially processed (VGG-16 based) features with XY-shift frame differenced frames, and a stacked-ConvLSTM component. Extensive experimental results over the largest video saliency dataset, DHF1K, show the competitive performance (accuracy) of our model against the state-of-the-art models.

2.1 Introduction

It is crucial that robotic systems employ robust computational models that irreproachably mimic human's perceptive and action intelligence, in real-time. Saliency prediction is among the most significant capabilities of human visual system. The human visual system is able to quickly distinguish important scenes in its visual field. The ability to computationally model this feature of human enables efficient and realistic human-robot interaction in social standard robotic environment [360, 12, 195]. Specifically, it plays a vital role in enabling intuitive and natural human-robot interaction by letting the robot to continuously pay attention to salient regions in its visual field [12, 361]. Besides, these computational models can be used as a source of efficiency in various computer vision tasks [362].

Saliency prediction systems have been applied to various problem domains, such as video segmentation [363, 362], video captioning [364, 365], video compression [295], image captioning [366] autonomous driving [367, 368], human-robotic interaction [12, 105], robot navigation [369, 370], surveillance [371, 372], and other areas [373, 374].

Visual saliency has been studied from the spatial [375, 376] and spatio-temporal perspectives [377] . Spatial information of individual images or frames has been used to build the earliest static image saliency prediction computational models. Several experiments also show that, computational models, especially those inspired by deep neural networks (DNN), suffice the problem of static saliency prediction [276, 277, 289, 292, 291]. However, because of the spatio-temporal or dynamic nature of videos, almost all static image saliency prediction models show hampered performance when employed on video stimulus.

To this end, recent video saliency prediction models are considering spatio-temporal aspects of video saliency dataset. This is mainly due to the recent cognitive and neuroscience of attention research findings, asserting to the importance of spatio-temporal features for data collected across space and time [378, 379]. Besides, advances in deep neural networks and their ability to efficiently handle spatio-temporal data contributed a lot to the growth of DNN inspired dynamic saliency prediction models.

A number of video saliency computational models have been produced in recent years. However, most models use datasets that lack generic, representative, and diverse instances in unconstrained task-independent scenarios. This has been exposing them for over-fitting [380] and incapability to work on real and diverse environment.

Very few computational models have been using diverse and representative datasets,

like DHF1K [19]. The use of large and representative video saliency dataset along with advanced deep neural networks show significant performance improvement [305, 19]. However, video saliency prediction problem in a complex and dynamic environment remains a challenge to this date. To this end, we propose a novel interactive stacked-ConvLSTM based video saliency model. Our architecture incorporates a novel XY-Shift frame differencing custom layer to enhance temporal features within the spatial domain. Additionally, we introduce an innovative approach to fuse temporally magnified spatio-temporal features with features generated by spatial feature extractors such as VGG-16 [381], ensuring a cohesive integration of both temporal and spatial elements. We use stacked-ConvLSTM component [381] for sequential fixation prediction over successive frames. A successive experiments we conducted on the largest video saliency dataset, DHF1K [19], show that our model achieves a competitive result against the state-of-the-art methods.

The rest of this chapter is organized as follows. The second part briefly introduces related research works, the third part introduces the proposed saliency prediction model in detail, the fourth part shows experimental details, and finally, a summary of our research work on video saliency prediction is presented.

2.2 Related Works

Recent researches on visual saliency have been consecutively redefining the state-of-the-art in the area. Most of the earliest saliency models are constructed from still images. These computational models assume that conspicuous visual features "pop-out" and involuntarily capture attention [13]. However, the performance of these models is significantly hampered as it belittles the impact of temporal features. To this end, recent advances on visual saliency prediction consider dynamic features for visual saliency prediction. The growth in this field of saliency is due to the growth in the area of deep learning and the availability of larger video saliency datasets. In this section, existing visual saliency prediction models that define the state-of-the-art in the area are briefly reviewed.

2.2.1 Saliency Models

Researches on human gaze fixation prediction or video saliency prediction is dating back to [275, 276]. The earliest saliency prediction methods are based on various low-level manual features of still image, such as color contrast, edge, center prior and orientation to

produce a "saliency map" [277, 278, 279, 280, 281, 282, 283]. A saliency map is an image that highlights the region on which human gaze could focus on a various probabilistic level.

Low-level feature based saliency models can work robustly on the simplest detection tasks. However, these models fail to perform well on a more complex image structures. To this end, various deep learning based static saliency researches are published Hou et al. [284], Lee et al. [275] and Li and Yu [276] Wang et al. [285] and Zhang et al. [286] [287, 288, 289, 290, 291, 292]. These models have achieved a remarkable result using the powerful learning ability of neural networks and growth in the size and quality of visual saliency datasets [289].

Static image saliency research is almost mature. However, subsequent trials to employ these models on video show a reduced performance [293]. These is mainly due to the frequent change in salient-goal over time in a sequence of frames. Furthermore, convolutional neural networks (CNN) have no memory function, so it is difficult to model video frames that are constantly changing in the time domain with CNN.

To this end, dynamic saliency models leverage both static and temporal features to predict human gaze fixation on videos [294, 295, 293, 296, 297, 298, 299, 300, 301]. Some of these studies [294, 293, 297] can be viewed as extensions of existing static saliency models with additional motion features. Conventionally, video saliency models pair bottom-up feature extraction with an ad-hoc motion estimation that can be performed either by means of optical flow or feature tracking. Frame-differencing [302], background subtraction [303], optical flow [304] and other methods are used to model spatial and motion information. However, these techniques are known for poor performance, especially in complex scene videos.

In contrast, deep video saliency models learn the whole process end-to-end. Some of these saliency models treat spatial and temporal features separately and fuse these features in the last few layers of the DNN architecture in certain way. Other researches simultaneously model the time and space information, directly letting the network simultaneously learn the time and space information and ensure the time and space consistency.

Research works that treat spatial and temporal information separately base on two-stream network architectures [305, 306] that accounts for color images and motion fields separately, or two-layer LSTM with object information [307, 308].

As one of the first attempts, [305] study the use of deep learning for dynamic saliency prediction and propose the so-called spatio-temporal saliency networks. They applied a

two-stream (5 layer each) CNN architecture for video saliency prediction. RGB frames and motion maps were fed to the two streams. They have investigated two different fusion strategies, namely element-wise and convolutional fusion strategies, to integrate spatial and temporal information.

Jiang et al. (2017) [307] concluded that human attention is mainly drawn to objects and their movement. Hence, they propose object-to-motion convolutional neural network (OM-CNN) to learn spatio-temporal features for predicting the intra-frame saliency via exploring the information of both objectness and object motion. Inter-frame saliency is computed by means of a structure-sensitive ConvLSTM architecture.

Zaho et al. (2019) [306] proposes two modules to extract temporal saliency information and spatial information. Moreover, the saliency dynamic information in time is combined with the spatial static saliency estimation model, which directly produces the spatiotemporal saliency inference. A context-aware pyramid feature extraction (CPFE) module is designed for multi-scale high-level feature maps to capture the rich context features. A channel-wise attention (CA) model and a spatial attention (SA) model are respectively applied to the CPFE feature maps and the low-level feature maps, and then fused to detect salient regions. Finally, an edge preservation loss is proposed to get the accurate boundaries of salient regions.

Tang et al. (2018) [308] used a multiscale spatiotemporal convolutional ConvLSTM network architecture (MSST-ConvLSTM) to combine temporal and spatial information for video saliency detection. This architecture not only retains the original temporal clues but also uses the temporal information in the optical flow map and the structure of LSTM. This part of the study separately learns the information in the time domain and the space domain through neural networks. Generally, to model the information in the time domain, some preprocessing methods, such as the optical flow method, are used. Additionally, the fusion of features extracted in the time and space domains also greatly affects the performance of the network. These works show a better performance and demonstrate the potential advantages in applying neural networks to video saliency problem.

Models that simultaneously model the time and space information directly let the network to concurrently learn the time and space information and ensure the time and space consistency. For instance, in reference [309], the author first used a pyramid dilated convolution module to extract multiscale spatial features and further extracted spatio-temporal information through a bidirectional convective ConvLSTM structure. Ingeniously, the author used the forward output of the ConvLSTM units as input and directly fed it into the

backward ConvLSTM units, which increases the capabilities to extract deeper spatiotemporal features.

In reference [310], unlike previous video saliency detection with pixel-level datasets, the author collected a densely annotated dataset that covers different scenes, object categories and motion modes. In Li et al. (2018), the author proposed a flow-guided recurrent neural encoder (FGRNE) architecture, which uses optical flow networks to estimate motion information per frame in the video and sequential feature evolution encoding in terms of LSTM network units to enhance the temporal coherence modeling of the per-frame feature representation [311].

Authors in Chaabouni et al. (2016) employed transfer learning to adapt a previously trained deep network for saliency prediction in natural videos [312]. They trained a 5-layer CNN on RGB color planes and residual motion for each video frame. However, their model uses only the very short-term temporal relations of two consecutive frames. In Bazzani et al. (2016), a recurrent mixture density network is proposed for saliency prediction [313]. The input clip of 16 frames is fed to a 3D CNN, whose output becomes the input to a LSTM. Finally, a linear layer projects the LSTM representation to a Gaussian mixture model, which describes the saliency map. In a similar vein, Mnih et al. (2014) applied LSTMs to predict video saliency maps, relying on both short- and long-term memory of attention deployment [314].

In Leifman et al. (2017), RGB color planes, dense optical flow map, depth map and the previous saliency map are fed to a 7-layered encoder-decoder structure to predict fixations of observers who viewed RGBD videos on a 2D screen [45].

As in their previous work of Gorji et al. (2018), here they used a multi-stream ConvLSTM to augment state-of-the-art static saliency models with dynamic attentional push (shared attention) [315]. Their network contains a saliency pathway and three push pathways including gaze following, rapid scene changes, and attentional bounce. The multi-pathway structure is followed by a CNN that learns to combine the complementary and time-varying outputs of the CNN-LSTMs by minimizing the relative entropy between the augmented saliency and viewers fixations on videos.

Wang et al. (2018) [19], proposed the Attentive CNN-LSTM Network which augments a CNN-LSTM with a supervised attention mechanism to enable fast end-to-end saliency learning. The attention mechanism explicitly encode static saliency information allowing LSTM to focus on learning a more flexible temporal saliency representation across successive frames. Such a design fully leverages existing large-scale static fixation datasets,

avoids overfitting, and significantly improves training efficiency.

Sun et al. (2018) [316] proposed a robust deep model that utilizes memory and motion information to capture salient points across successive frames. The memory information was exploited to enhance the model generalization by considering the fact that changes between two adjacent frames are limited within a certain range, and hence the corresponding fixations should remain correlated.

There are some more salient object detection models [317, 318, 319, 320, 321, 322, 284] that attempt to uniformly highlight salient object regions in images or videos. Those models are often task-driven and focus on inferring the main object, in stead of investigating the behavior of the HVS during scene free viewing.

2.2.2 Video Saliency Dataset

Recent advances in the area of human attention and dynamic fixation prediction are primarily triggered by the release of improved and large saliency dataset [323, 324, 325, 326]. These dataset improved the understanding of human visual attention and boosted the performance of computational models.

The DHF1K [19] dataset provide human fixations on a more diverse and representative dynamic nature scenes while free-viewing. DHF1K includes 1K video sequences annotated by 17 observers with an eye-tracker device.In DHF1K, each video was manually annotated with a category label, which was further classified into 7 main categories : daily activity, sport, social activity, artistic performance, animal artifact and scenery.

The Hollywood-2 [325] provide a dataset with 12 classes of human actions and 10 classes of scenes distributed over 3669 video clips and approximately 20.1 hours of video in total. The dataset intends to provide a comprehensive benchmark for human action recognition in realistic and challenging settings. According to analysis conducted by [327], 84.5 fixations Hollywood-2 dataset are located around the faces.

The UCF Sports dataset [325] consists of a set of actions collected from various sports which are typically featured on broadcast television channels such as the BBC and ESPN. The video sequences were obtained from a wide range of stock footage websites including BBC Motion gallery and GettyImages. It contains 150 videos taken from the UCF sports action dataset [328]. According to [327], 82.3 fixations of UCF sports saliency dataset fall inside the human body area.

Other datasets are either limited in terms of variety and scale of video stimuli [326, 324, 323, 329, 289], or collected for a special purpose (e.g., salient objects in videos [320]).

2.3 Our Approach

2.3.1 Overview

We propose a novel stacked-ConvLSTM based video saliency prediction model. Fig. 2.1 depicts the architecture of our video saliency prediction model. It is a stacked-ConvLSTM architecture that uses both convolutional and recurrent networks. Input to our stacked-ConvLSTM are preprocessed using a novel XY-shift frame differencing layer. This layer takes an absolute difference of an image and its shifted copy and return a high-pass filtered map. Furthermore, a three-frame differencing method takes this data and provides a temporal information aware spatial data map. Three-frame differencing help to magnify the effect of temporal features on the spatial domain and boost the capacity of the stacked-ConvLSTM component on spatio-temporal saliency prediction. Thus, our model produces accurate video saliency prediction with improved generalization. In this section, we introduce our proposed model architecture, and its three important components, namely the stacked-ConvLSTM module, the VGG-16 [382], and the XY-shift frame differencing module in detail.

2.3.2 The stacked-ConvLSTM Model

Fig 2.1 shows our proposed framework, consisting of three parts : the static convolutional component based on VGG-16 and with the weights of ImageNet [383], XY-shift frame differencing and the stacked-ConvLSTM component.

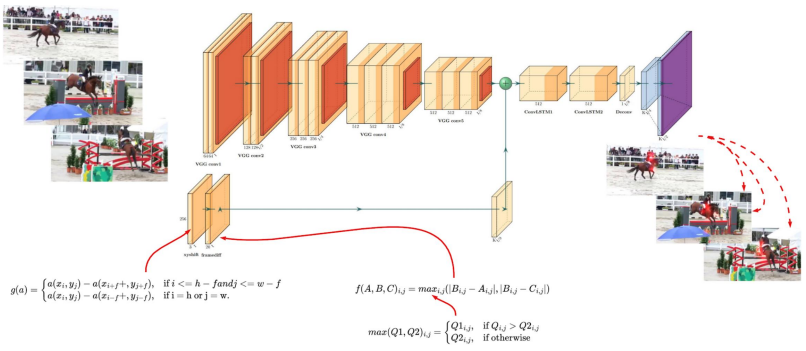


FIGURE 2.1 – Interactive Video Saliency Identification With Attentive ConvLSTM Architecture

2.3.3 Implementation Details

The implementation details are as follows. First, two-stream of data are passed to the VGG-16 and frame differencing components. The VGG-16 [382] extract spatial features from the raw image frames. In order to preserve more spatial details, Pool 4 and Pool 5 layers are removed, resulting in x8 instead of ×32 downsampling. At time step t, the input RGB image X_t size is (224×224×3). The output characteristic size of this component is [32, 40, 512]. Concurrently, we apply a batch level XY-shift frame differencing and three-frame differencing on each members of a batch to magnify temporal features on spatial domain. The XY-shift frame differencing differs a frame from its shifted replica. The effect of this operation is equivalent to the result of a high-pass filter method, but with significantly smaller computational resource. We have mainly used this method to reduce the visibility of irrelevant background objects and expose foreground objects. The mathematical formalization of XY-shift frame differencing is depicted as follows in equation 2.1. Let a be the first channel of image A with a shape of (h,w,3). Then, the XY-shift frame differencing of a is calculated as :

$$g(a) = \begin{cases} a(x_i, y_j) - a(x_{i+f}+, y_{j+f}), & \text{if} \\ i <= h - f and j <= w - f & \\ a(x_i, y_j) - a(x_{i-f}+, y_{j-f}), & \text{if i = h or j = w.} \end{cases} \tag{2.1}$$

where h and w stands for the height and width of the channel and f is a shift factor.

What follows the XY-shift frame differencing is an improved three-frame differencing technique. This technique use the output of XY-shift differencing. It takes three consecutive frames, compute the difference between the current frame and the previous frame, the current frame and the next frame separately, and extract a pixel-wise max between these two resulting frames. This technique is adapted and enhanced to improve the extraction of temporal features from datasets in spatio-temporal domain. The improved three-frame differencing method is formalized as follows in equation 2.2. Consider three consecutive XY-shift frame differenced frames denoted as A, B, and C, where A, B, and C represent the first channel of these frames, and they have a shape of (h, w). Let B be the first channel of the current frame. Then the improved three-frame differencing, f(A,B,C), is calculated as :

$$f(A, B, C)_{i,j} = max_{i,j}(|B_{i,j} - A_{i,j}|, |B_{i,j} - C_{i,j}|) \tag{2.2}$$

where for i,j >= 0 and i<= h and j <= w.

Furthermore, the pixel-wise maximum of two images is computed as shown in 2.3. Let Q1 be the absolute difference of the current frame B and its predecessor frame A. Let Q2 be the absolute difference of the current frame B and its successor frame C. Let both differenced images have a size of (h,w). Then, the pixel-wise maximum, P_{max}, of these two frames is calculated as :

$$max(Q1, Q2)_{i,j} = \begin{cases} Q1_{i,j}, & \text{if } Q_{i,j} > Q2_{i,j} \\ Q2_{i,j}, & \text{if otherwise} \end{cases} \tag{2.3}$$

where for i,j >= 0 and i<= h and j <= w.

A residual layer fusing the VGG-16 extracted spatial features and frame differencing output frames is applied succeeding the aforementioned components. Finally, the output of both VGG-16 and frame differencing mixed layer is deep fused into a single feature space. A [30x40x512] output of the residual layer is further fed to our stacked-ConvLSTM network. The main reason for stacking ConvLSTM is to allow for greater model complexity. Even though there are large-scale datasets like DHF1K that have 1K videos, the amount of training data is still insufficient, considering the high correlation among frames within same video [307]. Hence, increasing the complexity of the model help to extract more complex features in return providing robust video saliency prediction model. The size of the feature map after the stacked-ConvLSTM is 32x40x256. By passing this output through a convolutional layer, with kernel size 1x1, and upsampling the resulting feature map, we get 128x160x1 and 64x80x1 saliency map corresponding to the different loss functions we employed in this research work.

2.3.4 Loss Functions

To better generate robust saliency maps, we use three loss functions as used in Jiang et al. (2018) [384] and Wang et al. (2018) [19]. Linear Correlation Coefficient(CC) [385], the Kullback-Leibler divergence (KLD) [386] and Normalized Scanpath Saliency (NSS) [387]. The essence of using multiple loss functions is to increase the degree of learning and generalization of the model.

We denote the predicted saliency map as $Y \in [0,1]^{28x28}$, the map of fixation locations as $P \in \{0,1\}^{28x28}$ and the continuous saliency map (distribution) as $Q \in [0,1]^{28x28}$. Here

the fixation map P is discrete, that records whether a pixel receives human fixation. The continuous saliency map is obtained via blurring each fixation location with a small Gaussian kernel. Our loss functions is defined as follows :

$$L(Y, P, Q) = L_{KL}(Y, Q) + \alpha_1 L_{CC}(Y, Q) + \alpha_2 L_{NSS}(Y, P) \tag{2.4}$$

where $L_{KL}, L)CCandL_{NSS}$ are the Kullback-Leibler (KL) divergence, the Linear Correlation Coefficient (CC), and the Normalized Scanpath Saliency (NSS), respectively, which are derived from commonly used metrics to evaluate saliency prediction models. αs are balance parameters and are empirically set to $\alpha_1 = \alpha_2 = 0.1$.

Kullback–Leibler divergence (KLD) measures the divergence between the distribution S and $\hat{S}$:

$$L_{KL}(S, \hat{S}) = \sum_{i=1}^{NXM} \hat{S}_i \log \frac{\hat{S}_i}{Si} \tag{2.5}$$

Normalized Scanpath Saliency metric was introduced in Peter et al. (2005), to evaluate the degree of congruency between human eye fixations and a predicted saliency map [387]. Instead of relying on a saliency map as ground truth, the predictions are evaluated against the true fixations map. The value of the saliency map at each fixation point is normalized with the whole saliency map variance :

$$L_{NSS}(S^{fix}, \hat{S}) = \frac{1}{NXM} \sum_{i=1}^{NXM} [\frac{\hat{S}_i - \mu(\hat{S}_i)}{\alpha(\hat{S}_i)}] S_i^{fix} \tag{2.6}$$

Pearson's Correlation Coefficient (CC) measures the linear correlation between the ground truth saliency map and the predicted saliency map :

$$L_{CC}(S, \hat{S}) = \frac{\alpha(S, \hat{S})}{\alpha(S)\alpha(\hat{S})} \tag{2.7}$$

2.3.5 Training Protocol

Our model is iteratively trained with sequential fixation and image data. In training, a video training batch is cascaded with an image training batch. More specifically, in a video training batch, we apply a loss defined over the final dynamic saliency prediction from LSTM. For each video training batch, 20 consecutive frames from the same video are used. Both the video and the start frames are randomly selected. For each image training batch, we set the batch size as 20, and the images are randomly sampled from existing

static fixation dataset.

2.4 Experiments

2.4.1 Datasets and Evaluation Mertrics

Datasets

We use the DHF1K [19] dataset for training and evaluation. We use only the first 70% of the DHF1K dataset and used 60%/10%/30% training/validation/testing ratio to split data for the experiment. Hence, our model is trained and validated on 420 and 70 randomly selected videos. Moreover, the evaluation of our proposed model is undertaken on 210 test video sequences.

Evaluation Metrics

We use five performance evaluation metrics, namely Normalized Scanpath Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), AUC-Judd (AUC-J), and shuffled AUC (s-AUC).

Competitors

To prove the effectiveness of our proposed model, we compare our model with sixteen saliency models. Among them, [19], PQFT [295], Seo et al. [297], Rudoy et al. [296], Hou et al. [298], Fang et al. [299], OBDL [300], AWS-D [301], OM-CNN [307], and Two-stream [305] are dynamic saliency models. Furthermore, ITTI [276], GBVS [277], SALICON [289], DVA [292], Shallow-Net [291], and Deep-Net [291] are state-of-the-art static attention models. OM-CNN, Two-stream, SALICON, DVA, Shallow-Net, and Deep-Net are deep learning models, and others are classic saliency models. We choose these models due to publicly available implementations and their representability of the state-of-the-art.

Computational load

The whole model is trained in an end-to-end manner. The entire training procedure takes about 60 hours with a single NVIDIA Quadro RTX 3000 Max-Q GPU. Our model takes about 0.84s to process a frame image of size 224×224.

2.4.2 Performance Comparison

Performance on DHF1K

Table 2.2 presents the comparative performance of our model against the competitor models. It is observed that our model significantly outperformed all static saliency models and the majority of dynamic models, across all performance metrics. Our model show competitive result with the one reported in Wang et al. (2018) [19]. This is directly attributed to the novel XY-shift frame differencing technique and stacked-ConvLSTM network incorporated in our architecture.

	Models/Datasets	DHF1K				
		AUC-J	SIM	s-AUC	CC	NSS
Dynamic models	[295]	0.699	0.139	0.562	0.137	0.749
	[297]	0.635	0.142	0.499	0.070	0.334
	[296]	0.769	0.214	0.501	0.285	1.498
	[298]	0.726	0.167	0.545	0.150	0.847
	[299]	0.819	0.198	0.537	0.273	1.539
	[300]	0.638	0.171	0.500	0.117	0.495
	[301]	0.703	0.157	0.513	0.174	0.940
	[307]	0.856	0.256	0.583	0.344	1.911
	[305]	0.834	0.197	0.581	0.325	1.632
	[19]	**0.885**	**0.311**	0.553	**0.415**	**2.259**
Static models	[276]	0.774	0.162	0.553	0.233	1.207
	[277]	0.828	0.186	0.554	0.283	1.474
	[289]	0.857	0.232	0.590	0.327	1.901
	[291] Shallow-Net	0.833	0.182	0.529	0.295	1.509
	[291] Deep-Net	0.855	0.201	0.592	0.331	1.775
	[292]	0.860	0.262	0.595	0.358	2.013
Training Setting I	Our model	**0.878**	**0.304**	**0.665**	**0.405**	**2.239**

FIGURE 2.2 – Quantitative results on DHF1K : Training setting I is trained and evaluated using only DHF1K dataset

2.4.3 Analysis

In the course of our research, we have conducted extensive experiments. Here, we analyse our model and competitive models thoroughly with the intention of giving deeper insight to the state-of-the-art models and suggest opportunities that we believe are

inspiring for future work in dynamic video prediction.

We conduct our analysis first by contrasting the effect of employing deep learning methods for static and dynamic saliency prediction. According to our finding, deep learning methods outperform classic methods both in static DVA [292], Deep-Net [291] and dynamic OM-CNN [307], Two-stream [305], ACL [19] saliency prediction problems, and in almost all saliency prediction metrics. On the other hand, classic methods show relatively reduced performance in static saliency predication ITTI [276],GBVS [277]. A significant performance degradation is observed when static saliency prediction algorithms are employed for dynamic saliency prediction problem sets PQFT [295], [297], [296], [298], [299]. This demonstrates the strong learning ability of deep neural network and the promise of developing deep learning network based models in this challenging area. Moreover, the analyses show the inherent incapability of classic machine learning methods for complex problem sets such as, saliency prediction.

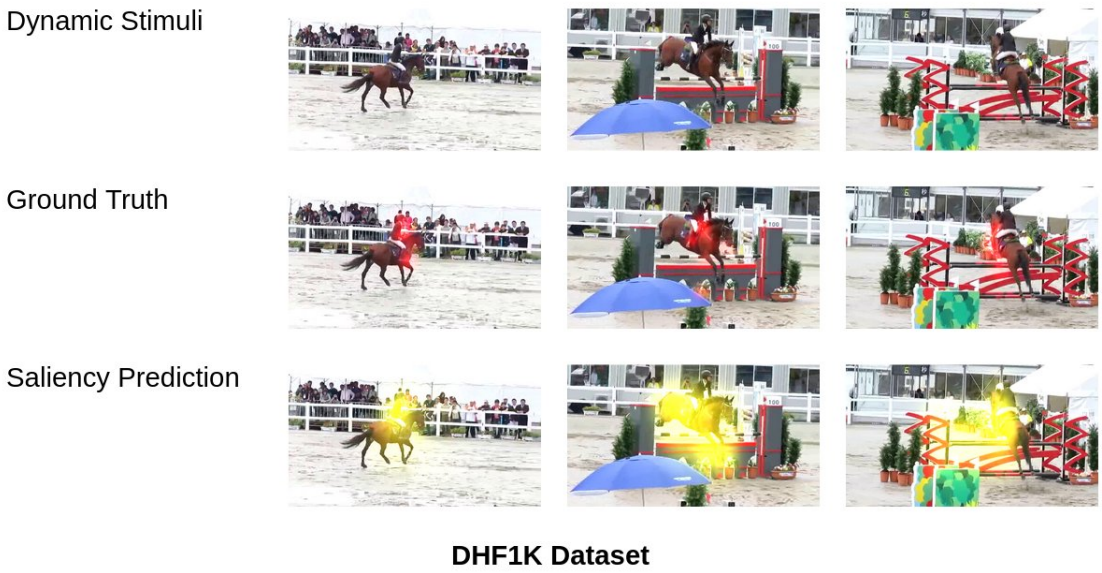


FIGURE 2.3 – Qualitative results of our video saliency model on DHF1K Dataset

2.4.4 Ablation Study

In this section, we discuss component wise contribution of our model. We verify the effectiveness of various components and their order of composition in our model.

The effectiveness of the XY-shift frame differencing technique is analyzed by eliminating its effect from the general architecture. A stacked-ConvLSTM architecture without

our novel frame differencing layer show reduced performance in capturing saliency in highly dynamic scenes. Quantitatively speaking, we noticed 20 to 25 percent performance reduction in all evaluation metrics we employed. Performance gains due to the novel XY-shift frame differencing is attributed to the magnified temporal features in the spatial domain. Magnifying temporal features in the spatial domain help the stacked-ConvLSTM component to easily extract spatio-temporal saliency features.

Besides, due to the complex nature of dynamic saliency prediction, the use of stacked-ConvLSTM component right after a spatial feature extractor component improve our model's performance on complex feature extraction. Consequently, the use of stacked-ConvLSTM rather than a single ConvLSTM architecture show slight performance improvement.

Another interesting finding in the course of our research is the effect of residual layer positioning. The variation in the position of residual layers show significant performance variation. We placed residual layers residual layers in different positions, such as at the end of the primary convolutional base, between the ConvLSTM layer, and finally, at the end of our overall encoder, processing every input in a separate stream. Placing residual layer at the beginning of the stacked-ConvLSTM component yield better saliency prediction performance and relatively better resource utilization.

Similarly, we undertook a through qualitative analysis by randomly selecting sequence of frames from our testing set. On the other hand, the interactivity [274] of our model is evaluated by deploying it in a resource constrained robot called Pepper. The results show the effectiveness of our video saliency prediction model relative to the state-of-the-art video saliency prediction models.

2.5 Conclusion

In this research, we proposed a novel deep learning based dynamic saliency prediction model, which employ the benefits of a novel XY-shift frame differencing technique and stacked-ConvLSTM network. An extensive experimentation on the largest video saliency dataset, DHF1K [19] is undertaken. We compared our results with similar deep learning based dynamic saliency models. Our experimental results show the effectiveness and superiority of our model against 15 state-of-the-art models and its competitiveness against the outperforming dynamic saliency prediction model [19].

Chapitre 3

A New Approach to Moving Object Detection and Segmentation : The XY-shift Frame Differencing

Motion out-weights other low-level saliency features in attracting human attention and defining region of interests. The ability to effectively identify moving objects in a sequence of frames help to solve important computer vision problems, such as moving object detection and segmentation. In this chapter, we propose a novel frame differencing technique along with a simple three-stream encoder-decoder architecture to effectively and efficiently detect and segment moving objects in a sequence of frames. Our frame differencing component incorporates a novel self-differencing technique, which we call XY-shift frame differencing, and an improved three-frame differencing technique. We fuse the feature maps from the raw frame and the two outputs of our frame differencing component, and fed them to our transfer-learning based convolutional base, VGG-16. The result from this sub-component is further deconvolved and the desired segmentation map is produced. The effectiveness of our model is evaluated using the re-labeled multi-spectral CDNet-2014 dataset for motion segmentation. The qualitative and quantitative results show that our technique achieves effective and efficient moving object detection and segmentation results relative to the state-of-the-art methods.

3.1 Introduction

Motion is a key social stimulus that engages visual attention and induces autonomic arousal in the viewer [388, 389]. Motion detection is extensively used in computer vision applications to facilitate the analysis of real world video scenes. Video surveillance [390] being a significant one, monitoring traffic and pedestrian [391], robot control [392], target detection and counting [393], and detection of human activity [394] are some of its

applications in computer vision.

The process of relating moving region of interests (ROI) with object/s is called moving object detection. Moving object detection concerns how to take out moving objects from video frames and remove the background region and noise. Robust moving object detection enable computationally optimal foreground object detection and saliency prediction [335].

The optical flow method use sequence of ordered images to estimate motion of objects as either instantaneous image velocities or discrete image displacements [395]. It is based on the properties of flow vector of the object over time to detect moving object regions. Optical flow method is highly complex and susceptible to noise corruption, fake motion and illumination variation.

On the other hand, background subtraction comprises of two steps : (i) background modeling and (ii) computation of difference between the current background model and the current video frame. The performance of background subtraction method is highly dependent on the accuracy of the background model. This method achieves outstanding performance when the background model is accurate. However, if the background model is inaccurate, it may lead to incorrect detection of moving objects [396]. The performance of background subtraction technique degrades when it face videos with smaller frame rate, camera jitter, and significant illumination change.

The other widely used moving object detection technique is frame differencing. It detects moving objects by taking pixel-by-pixel difference of consecutive frames in a video sequence. Frame differencing is the most common and computationally less complex method for moving object detection in scenarios where the scene is dynamic due to camera movement and mobility of objects in a video sequence. However, this method fails to detect whole relevant pixels of some types of moving objects (foreground aperture problem) [354]. It also wrongly detects a trailing regions as moving object (known as ghost region) when there is an object that is moving fast in the frames [348]. Most significantly, frame differencing fails to detect objects that preserve uniform regions.

To this end, we propose a novel XY-shift frame differencing technique along with three-stream encoder-decoder architecture to address the setbacks of frame differencing based moving object detection and segmentation techniques. The effectiveness of our frame diffrencing technique is analyzed both as a standalone high-pass filter algorithm and as an input for our improved three-frame differencing and encoder-decoder network. Furthermore, we implement three-stream encoder-decoder architecture to build simple but robust moving object detection model.

The rest of this chapter is organized as follows. The second part briefly introduce related research works, the third part introduce the proposed method in detail, the fourth part of this chapter show experimental analysis of the research work, and finally, a summary of this chapter is presented.

3.2 Related Works

In the early days of moving object detection, researchers formulated a well established background subtraction techniques for stationary camera setting. These techniques have been extended for many years and are able to successfully detect moving objects as long as the camera is stationary [330, 331, 330, 332, 333, 334, 335]. However, a relatively long initialization time to model the background and residual image alignment error on non-stationary camera setting are the main setbacks of this technique [336]. These problems are incontestably addressed by optical flow based methods [337, 338, 304]. However, optical flow is highly dependent on optical flow vectors. The quality of the these vectors is crucial for the motion segmentation performance. Besides, optical flow is highly complex and due to high sensibility of noise corruption, it cannot meet the need of real time object detection. Optical flow exceptionally works well on large moving objects and fails to detect smaller objects due to blurry edges and low resolution [339].

A variety of frame differencing techniques address the aforementioned problems. For instance, inter-frame differencing technique generates the difference between two consecutive frames over a period of time for identifying background and foreground pixels. A research in Liang et al. (2010) use inter-frame differencing algorithm to detect moving target in aviation video [340]. Their experiment indicate that the algorithm has few computations and high accuracy to extract moving-target in aviation videos. In Nakashima et al. (2018), interframe differencing and dynamic binarization using discriminant analysis is applied [341]. The positions of the moving object in the image are determined by observing the histograms of each frame.

A slightly different method with comparable result with that of inter-frame differencing is three-frame differencing. This method put three adjacent frames as a group, subtracts both adjacent frames and lets two differential results do the logical AND operation. This has been the most widely used and traditional three-frame differencing method. In Yin et al. (2016), traditional three-frame differencing technique and W4 algorithms are used to detect foreground objects in the infrared video datasets [342]. Another research, Sengar

et al., propose a moving object detection method under static background [343]. The algorithm use a non-overlapping blocks of the difference frames and calculate the intensity sum and mean of each block.

The inter-frame differencing and three-frame differencing techniques suffer from the foreground aperture and ghosting problems due to slow-moving as well as fast-moving foreground objects. Besides, these methods are known for partial or splitted detection of objects [344]. As a result, frame differencing techniques are prone to false positives and sometimes false negatives.

To this end, combined moving object detection methods improved the performance of frame differencing techniques. Relatively robust moving object detection methods combine background subtraction with frame differencing [345, 346, 347, 348] or optical flow [349, 336, 350].

A static background based on three-frame differencing method in combination with background subtraction method is proposed in Weng et al. (2010) [345] and Cheng et al. (2014) [347]. A combination of optical flow and three-frame differencing based moving object detection method is employed in Halidou et al. (2014) [349]. It use region of interest (ROI) and multi-block local binary pattern descriptors. Another frame differencing and optical flow based moving object detection technique is proposed in Fernandez et al. (2010) [350]. Here, a thermal infrared camera mounted on autonomous mobile robot is used as a feed to the detection module.

A method for detecting moving people in the indoor environment is proposed with the help of frame differencing and neural network based classification techniques [351]. This method reduces the false negatives and provides a robust classification with the help of a finite state automata. Similarly, a new approach based on fuzzy adaptive resonance theory, neural network with forgetting method for foreground detection and background establishment in natural scenes is proposed in Dou et al. (2014) [352]. On the other hand, the frame differencing and the non-pyramidal Lucas-Kanade approaches [353] are used to detect human candidates based on thermal signatures when the robot stops and moves.

In Xu et al. (2017), an efficient foreground detection method is proposed by combining three-frame differencing and Gaussian mixture model [354]. Another research work, Lee et al. (2013), present a moving object detection method by combining background subtraction, separable morphological edge detector, and optical flow [355].

Recently, a more sophisticated and efficient moving object detection methods have been proposed by intriguing improved frame differencing techniques with deep neural

network technologies. For instance, Ellenfeld et al. (2021) propose a deep learning based moving object detection method [21]. It use a Deep Convolutional Neural Network (DCNN) for multi-modal motion segmentation. Improved three-frame differencing and current RGB frame is used to capture temporal information and appearance of the current scene respectively. These inputs are later fused in the DCNN component for effective, efficient and robust motion segmentation. This model improved the performance of three-frame differencing techniques in detecting tiny moving objects.

A research work in Yang et al. (2017) applied a frame differencing technique with Faster Region-Convolutional Neural Network (R-CNN) for highly precise detection and tracking characteristics [356]. Similarly, Mohtavipour et al. (2022) propose a multi-stream CNN and frame differencing based moving object detection method for deep violence detection [357]. It use a handcrafted features related to appearance, speed of movement, and representative image and fed to a convolutional neural network (CNN) as spatial, temporal, and spatiotemporal streams.

Furthermore, Siam et al. (2018) [358] and Wang et al. (2018) [359] show promising results using CNN for moving object detection. They use a two-stream convolutional network to jointly model motion and appearance cues in a single convolutional network. In Wang et al. (2018) [359] a new framework named moving-object proposals generation and prediction framework (MPGP) is proposed to reduce the searching space and generate some accurate proposals which can reduce computational cost. In addition, they explored the relation of moving regions in feature map of different layers. This method utilizes spatial-temporal information to strengthen the detection score and further adjust the location of the bounding boxes.

3.3 Our Approach

We propose a novel moving object detection and segmentation method using XY-shift and improved three-frame differencing. Furthermore, we have extended our method by feeding it to a three-stream encoder-decoder network. In this section, we discuss details of our proposed technique.

3.3.1 The Proposed Framework

Fig 3.2 show our proposed framework, consisting of two major components namely : frame differencing and three-stream encoder-decoder network. The frame differencing component consists of two frame differencing methods. The first method is a novel XY-shift frame differencing technique and the second one is an improved three-frame differencing technique. The XY-shift frame differencing differs a frame from its shifted replica. The effect of this operation is equivalent to the result of a high-pass filter method, but with significantly smaller computational resource. We have mainly used this method to reduce the visibility of irrelevant background objects and expose foreground object even if they are in a temporarily static position. The mathematical formalization of XY-shift frame differencing is depicted as follows in equation 3.1. Let a be the first channel of image A with a shape of (h,w,3). Then, the XY-shift frame differencing of a is calculated as :

$$g(a) = \begin{cases} a(x_i, y_j) - a(x_{i+f}+, y_{j+f}), & \text{if } i <= h - f \\ & j <= w - f \\ a(x_i, y_j) - a(x_{i-f}+, y_{j-f}), & \text{if i = h or j = w} \end{cases} \tag{3.1}$$

where h and w stands for the height and width of the channel and f is a shift-factor.

Minuned

00	01	02	03	04	05
10	11	12	13	14	15
20	21	22	23	24	25
30	31	32	33	34	35
40	41	42	43	44	45
50	51	52	53	54	55

Minuned ↦ Subtrahend

11	12	13	14	15	15
21	22	23	24	25	25
31	32	33	34	35	35
41	42	43	44	45	45
51	52	53	54	55	55
51	52	53	54	55	55

FIGURE 3.1 – Tabular representation of XY-shift operands

Figure 3.1 depicts a notational representation of XY-shift frame differencing. Assuming a 6X6 pixel raw image as a minuend, the subtrahend of XY-shift frame differencing technique with 1 shift-factor is constructed starting from the second row and column of the minuend.

What follows the XY-shift frame differencing is an improved three-frame differencing technique. This technique uses the output of XY-shift differencing. It takes three consecutive frames, computes the difference between the current frame and the previous frame, the current frame and the next frame separately, and extracts a pixel-wise max between these two resulting frames. This technique is adapted and enhanced to improve the extraction of temporal features from datasets in spatio-temporal domain. The improved

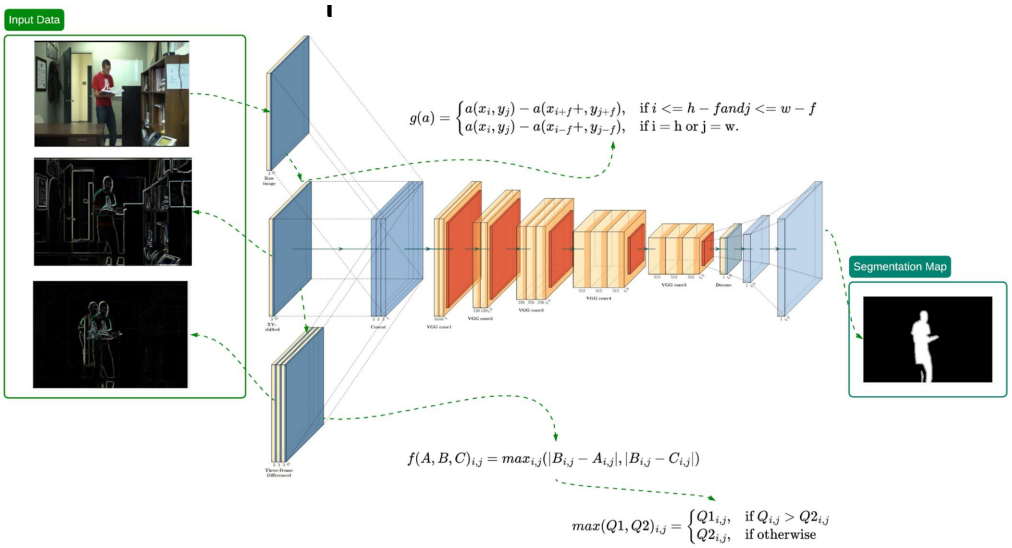


FIGURE 3.2 – Moving object detection and segmentation architecture

three-frame differencing method is formalized as follows in equation 2.2. Let A,B, and C be the first channel of three consecutive XY-shift frame differenced frames with a shape of (h,w). Let B be the first channel of the current frame. Then the improved three-frame differencing, f(A,B,C), is calculated as :

$$f(A, B, C)_{i,j} = max_{i,j}(|B_{i,j} - A_{i,j}|, |B_{i,j} - C_{i,j}|) \tag{3.2}$$

where for i,j >= 0 and i<= h and j <= w. Furthermore, the pixel-wise maximum of two images is computed as shown in equation 3.3. Let Q1 be the absolute difference of the current frame B and its predecessor frame A. Let Q2 be the absolute difference of the current frame B and its successor frame C. Let both differenced images have a size of (h,w). Then, the pixel-wise maximum, P_max, of these two frames is calculated as :

$$max(Q1, Q2)_{i,j} = \begin{cases} Q1_{i,j}, & \text{if } Q_{i,j} > Q2_{i,j} \\ Q2_{i,j}, & \text{if otherwise} \end{cases} \tag{3.3}$$

where for i,j >= 0 and i<= h and j <= w.

The second component of our model constitutes a VGG-16 based encoder and a decoder network. The top five convolutional layers of VGG-16 along with the weights of ImageNet [382] are used as an encoder with the intention of avoiding excessive sparsity of hidden units.

The implementation detail is as follows. Three consecutive raw image frames are passed to the frame differencing component. The XY-shift frame differencing component takes each frame separately and performs XY-shift frame differencing. The result of XY-shift frame differencing on random images is presented in figure 3.3 Concurrently, the improved three-frame differencing takes all XY-Shift frame differenced frames and perform three-frame differencing for each consecutive frames using a pixel-wise maximum function, shown in equation 3.3.

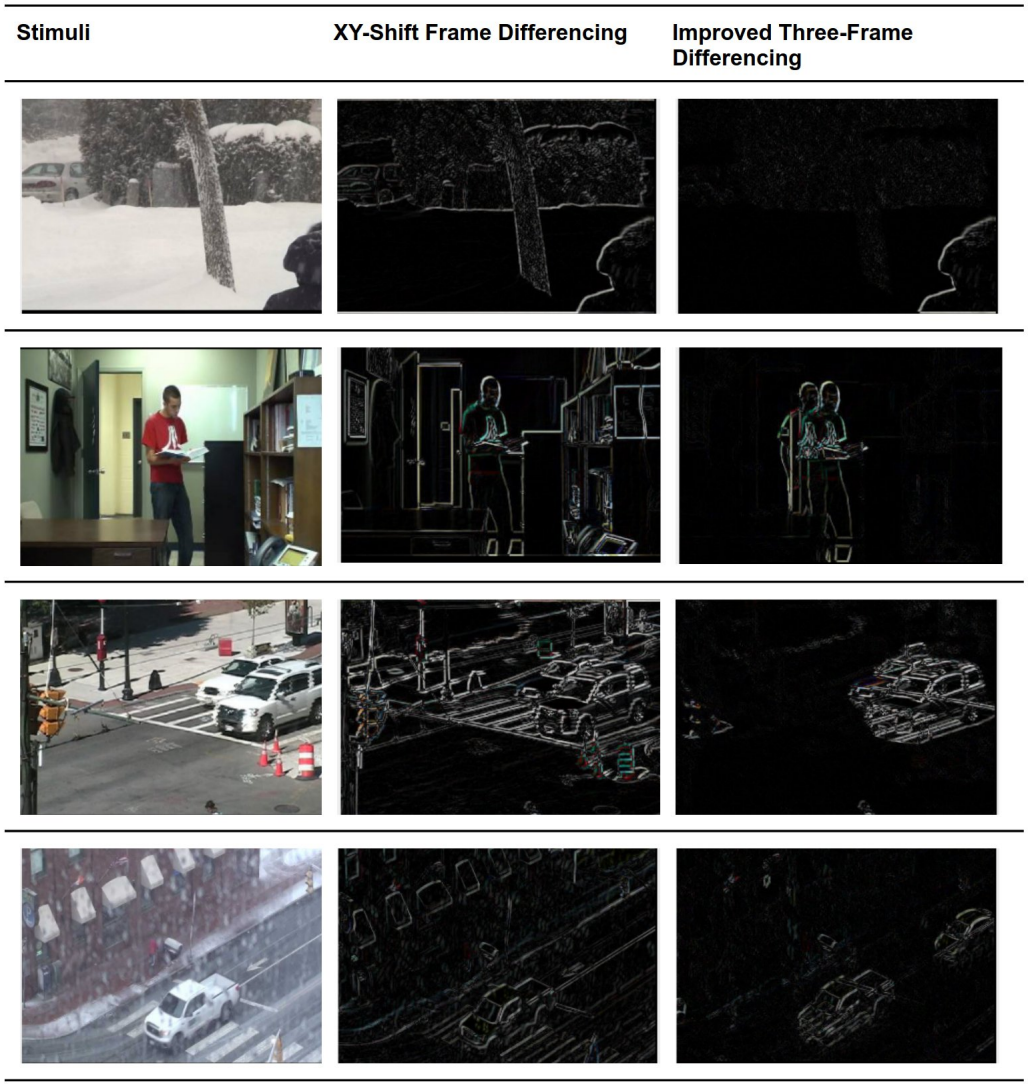


FIGURE 3.3 – A random presentation of XY-shift frame differenced and Three-Frame differenced frames : frames not thresholded

The purpose of the XY-shift frame differencing as depicted in column two of Fig. 3.3 is to clear irrelevant background objects. This contributed a lot in reducing textures that

affects the three-frame differencing negatively. The third column of Fig. 3.3 clearly shows the contribution of the XY-shift frame differencing in enhancing three-frame differencing techniques.

Furthermore, the result of the improved three-frame differencing, the XY-shift frame differencing, and the raw RGB image frames is fused into one future space using a residual layer and fed to the last encoder-decoder network. The VGG-16 based encoder extracts further features and the decoder segment produces the desired segmentation map as shown in Fig. 3.2.

3.3.2 Loss Function

Moving object detection and segmentation task is a binary pixel-wise classification task. Consequently, binary cross-entropy is chosen as the loss function to evaluate the model performance during training.

The binary crossentropy loss function calculates the loss as shown in equation 3.4 :

$$L_{BCI} = -\frac{1}{S}\sum_{i=1}^{S} y_i . \log \hat{y}_i + (1 - y_i).(\log(1 - y_i) \tag{3.4}$$

where $\hat{y}_i$ is the i^{th} scalar value in the model output, y_i is the corresponding target value, and S, the output size, is the number of scalar values in the model output.

3.3.3 Training Protocol

We use the [397]'s 4-fold cross validation strategy to split between training and test data of CDNet-2014. All CDNet-2014 sequences are divided equally into four disjoint splits. The model is trained on three of the splits. The remaining split is used to evaluate the model performance. We used a built-in python randomize function to select sequence of frames in each iteration from a varity of videos. A randomizer function is set to pick n number of frames in each iteration where n is the batch size. In this way we enhanced the representatives and complexity of the data, overcoming the possibility of overfitting at the same time. The designated training data is further divided into a training split (90%) and validation split (10%). Moreover, early stopping is used to prevent overfitting : if the validation loss does not improve in two consecutive epochs, the training process is terminated.

We used Adaptive Moment Estimation (Adam) [398] to optimize the network during

the training process. During training, the learning rate was set to 0.0001 and was decreased by a factor of 10 every 2 epochs. The network was trained for 10 epochs. The whole model is trained in an end-to-end manner. The entire training procedure takes about 8 hours using a single NVIDIA Quadro RTX 3000 Max-Q GPU.

3.4 Experiments

3.4.1 Datasets

We use the relabeled version of CDNet-2014, [20] dataset, for training and evaluation. CDNet-2014 includes over 160,000 pixel-wise annotated frames in 53 video sequences subdivided in 11 categories and two spectra : VIS and thermal IR. The 53 sequences contain a large variety of different scenes with varying image quality and resolution ranging from (320 × 240) to (720 × 480) pixels. Most scenes show an urban environment with persons or cars. The dataset contains both indoor and outdoor scenes and covers many different real world challenges such as shadows, dynamic backgrounds, and camera motion. The ground truth for each image is a gray-scale image that describes the 4 motion classes : static, hard shadow, unknown motion, and motion. An additional class is used to mark areas that are outside the region of interest (non-ROI). Pixels annotated as non-ROI are discarded during evaluation.

3.4.2 Evaluation Metrics

We evaluate our model on the testing sets of CDNet-2014, in total of 11 video sequences with nearly 39,820 frames. We emphasized on the Recall (Re), Precision (Pr), and F1-score from the the standard evaluation measures [20]. This is mainly due to the sufficiency of the selected metrics for the problem at hand.

Table 3.1 – Quantitative comparison of results

Approach	**Precision**	**Recall**	**F1-Score**
[399]	0.626	0.673	0.553
[346]	0.462	0.513	0.42
STBGS [334]	0.406	0.549	0.401
[343]	0.375	0.58	0.389
[21]	0.774	0.751	0.745
Ours	**0.801**	**0.795**	**0.772**

3.4.3 Frame Differencing Experiments

We undertook an ablation analysis on the different components of our architecture. The significance of the XY-shift frame differencing and improved three-frame differencing is thoroughly analysed by eliminating and replacing each of them with inter-frame differencing [340, 341] and traditional frame differencing [342, 343] techniques. The substitution of both of these frame differencing methods exhibits a reduced model performance. Especially, elimination of our XY-shift frame differencing technique and the use of inter-frame differencing as a source of input for our model caused a major performance degradation. The elimination of our frame differencing techniques significantly reduced our model performance on most of its salient features, such as robustness against false-motion, temporarily at-rest object detection, and tiny moving object detection.

The use of our XY-shift frame differencing technique along with the improved three-frame differencing technique exhibit an outstanding performance. This is mainly due to the power of our XY-shift frame differencing technique in eliminating dynamic and noisy backgrounds. Moreover, the use of XY-shift differenced frames for three-frame differencing component further improved the performance of our model, especially in dynamic background scene videos. From the qualitative analysis point of view, the absence of XY-shift frame differencing affected our model performance on tiny moving objects and temporarily at-rest foreground objects.

3.4.4 Optical Flow Experiments

Our second phase of ablation analysis concerns with the popular optical flow technique. We converted our model into a two-stream architecture and assessed the impact of optical flow technique. Here, we used optical flow output and raw image as input source of a two-stream encoder-decoder architecture. The ability of our model to detect false-negatives was slightly compromised in this setup. We extended our model to a three-stream network by replacing the improved three-frame differencing segment by the optical flow output. This setting slightly improved its performance but with major deficiency to yield a competitive result.

3.4.5 Comparisons With The State-of-the-art

To compare models with the state-of-the-art, it is necessary to have readily available repositories of research works that form the code base for the state-of-the-art. It has been

difficult to find code bases of moving object detection and segmentation research works. However, we were able to compare our proposed model with five other state-of-the-art methods for motion detection and segmentation namely, [399], [346], [343], and [21]. Table 3.1 shows the quantitative results regarding Precision, Recall, and F1-score. Our approach outperforms most of these methods by a large margin and scores an outstanding result with the recent research work that combines three-frame differencing with DCNN [21].

The qualitative result is visualized in figure 3.4. What is depicted in the first row is the ground truth set by the CDNet-2014 dataset. The second raw show the raw appearance images for the corresponding ground truth. We used columns to showcase the performance of the state-of-the-art papers and our model over these images. Images are selected from different scene videos to show the generalization capacity of models in different environment. The qualitative analysis show the poor object detection and segmentation performance of background subtraction and frame differencing based algorithms in complex scene environment like, images with water in the background and dynamic scenes. As it can be seen in figure 3.4, these algorithms are highly prone to false positives and negatives. Compared to these algorithms, our model was able to detect both moving objects and temporarily at-rest objects effectively and efficiently.

The most robust model that we analysed in this section is [21]. As it is discussed in the previous sections, this model combine improved three-frame differencing technique with appearance frame in a two stream encoder-decoder architecture. Compared to other deep learning based algorithms, these model show relatively better moving object detection and segmentation performance. Finally, our qualitative and quantitative results show the efficiency and effectiveness of our model. It has also scored a competitive result compared to the state-of-the-art methods. However, given the big error gap shown in Table 3.1, there is still a need to further enhance methods. Exploiting robust digital image processing and deep neural network technologies should be the focus of our next phase of research.

3.5 Conclusion

We proposed a novel moving object detection and segmentation technique. Our contribution in terms of architectural component is two fold : we propose a novel frame differencing technique, XY-shift frame differencing, and enhanced the traditional three-frame differencing technique. The XY-shift frame differencing is mainly used to sharpen the image and transform it to a high-pass filtered frame in relatively smaller computatio-

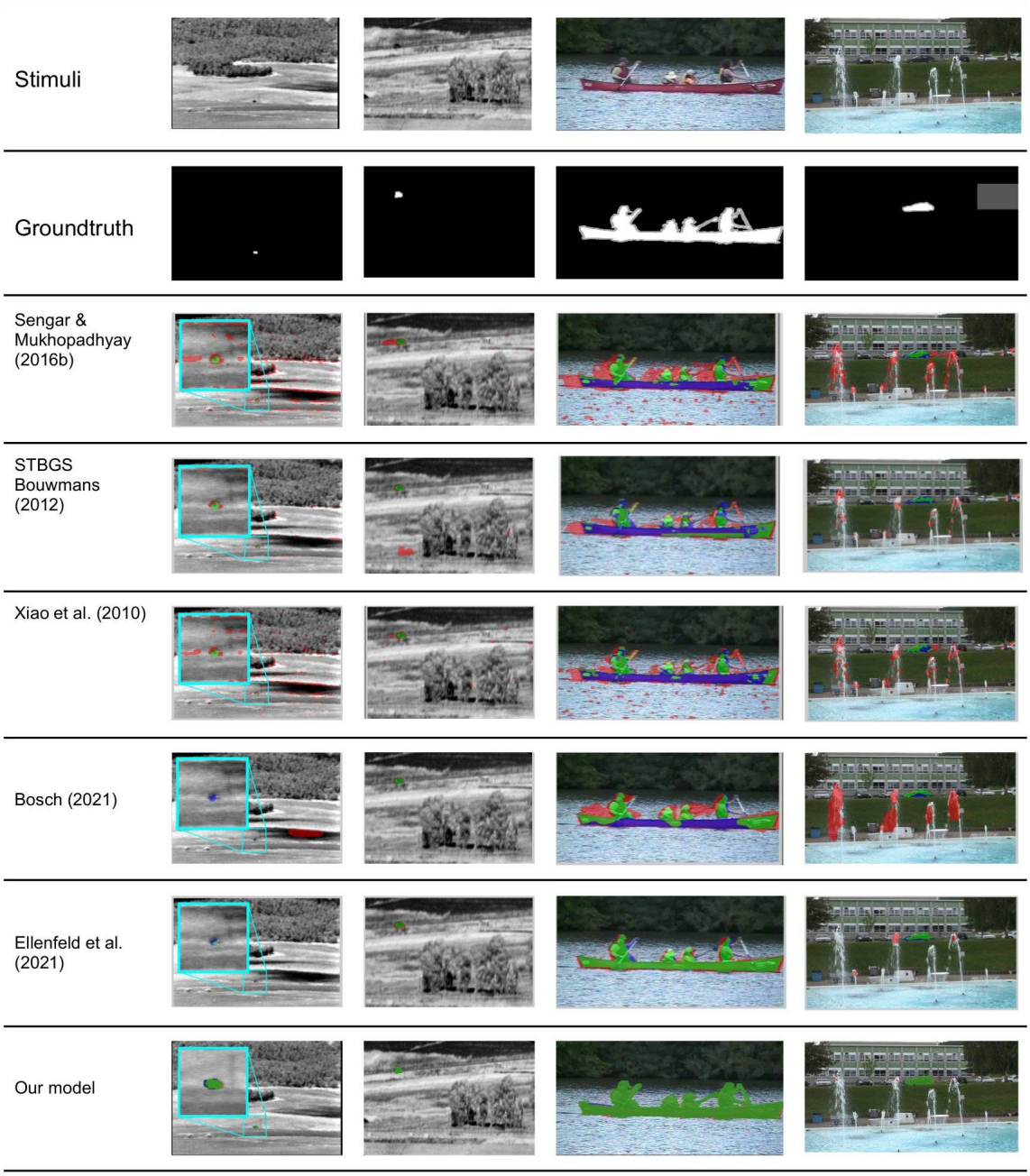


FIGURE 3.4 – Qualitative evaluation of different models against our model : The green, blue, and red shades indicate the correct, false negative, and false positive classifications

nal resource. Image passed through this component eliminate most types of noises and irrelevant background objects. Feeding the output of the XY-shift frame differencing to the traditional three-frame differencing technique happen to overcome the most common defects of three-frame differencing. The performance of the the three-frame differencing

technique is improved by feeding XY-shift frame differenced frames instead of raw image and computing the pixel wise maximum of differences.

We have also introduced an efficient way of fusing raw appearance images with frame differenced images resulting in a significant improvement on moving object detection and segmentation models. We used a three-stream encoder-decoder deep neural network architecture. The raw appearance image, XY-shift frame differenced image, and three-frame differenced images are fed to their corresponding stream and later a feature space which is a deep fusion of the three stream data is produced. This intermediate feature space is fed to the encoder, which the top 5 layers of VGG-16, and the resulting feature has been deconvolved to get the desired segmentation map. The CDNet-2014 change detection dataset were used to build and analyse our model.

Our experimental analysis covered multiple perspectives of moving object detection and segmentation. We undertook an exhaustive ablation analysis by replacing our proposed frame differencing component with background subtraction, three-frame differencing, and optical flow based moving object detection and segmentation techniques. For the evaluation, we used precision, recall and F1-score metrics. Both qualitative and quantitative results show that the proposed approach outperform the state-of-the-art moving object detection and segmentation methods.

Chapitre 4

Anthropomorphic Human-Robot Interaction Framework : Attention Based Approach

In recent years, robots have become increasingly integrated into our daily lives, from automated assistants in our homes to manufacturing robots in factories. However, for robots to effectively interact with humans, they must be able to identify environmental cues like humans do. This is where the field of human-robot interaction comes in, which aims to create intuitive and natural interactions between robots and humans.

One important aspect of HRI is the ability for robots to process visual information in a way that is similar to humans. Human attention models have been developed to simulate the way humans process visual information, making them useful for identifying important regions in images and videos. In this chapter, we explore the use of human attention models in developing intuitive and anthropomorphic HRI. Our approach involves combining a saliency model and a moving object detection model to identify regions of interest in images and videos. The framework is implemented using the ROS framework on Pepper, a humanoid robot. To evaluate the effectiveness of our system, we conducted both subjective and objective measures. The subjective measures included rating measures to evaluate intuitiveness, trust, engagement, and user satisfaction, while the objective measures evaluated the performance of their human attention subsystem against state-of-the-art models.

The results of their experiments demonstrated the significant impact of our framework in enabling intuitive and anthropomorphic human-robot interaction. The subjective measures showed that participants found the interactions with the robot to be more natural and engaging, while the objective measures demonstrated that their human attention subsystem outperformed state-of-the-art models in identifying important regions in images and videos.

Overall, this chapter highlights the importance of incorporating human attention models into the development of HRI systems, and demonstrates the potential of such systems to create more intuitive and natural interactions between humans and robots.

4.1 Introduction

AI agents are frequently described as having the capacity to reason and behave in a manner that resembles human and rational thinking [400]. HRI is improved by agents that can emulate human thinking and behavior because HRI is primarily defined by the observable actions and results of human thought in behaviorism [401].

Human attention models enable robots to respond to visual stimuli like humans [17]. By integrating these models into their behavior, robots can identify relevant visual information, make informed decisions, and interact more naturally with humans. This human-like allocation of visual attention helps robots perform tasks effectively and efficiently, improving the user experience. Specifically, saliency prediction [2] and moving object detection models [6] simulate human attention in computer vision. Saliency prediction models identify attention-grabbing regions in an image using low and high-level features while moving object detection models identify motion in a scene. Humans attend to moving objects more than static ones. Combining these models can simulate different aspects of human attention, creating more complex attentional simulations that resemble how humans attend to the visual world in complex anthropomorphic behavior.

Human-robot interaction is an essential aspect of the RoboCup@Home competition, where robots are tasked with performing various domestic tasks. Anthropomorphic behaviors, which mimic human characteristics and mannerisms, are highly valued in this competition. These behaviors may include social cues such as maintaining eye contact, appropriate gesturing, and active listening skills. In particular, two of the most challenging tasks in the competition are the Receptionist and General Purpose Service Robot tasks [402]. In these tasks, the robot must be able to autonomously identify and engage with human beings, determining which individual wants to interact and focusing its attention on the speaker for the duration of the conversation. To succeed in these tasks, the robot must possess advanced sensory and cognitive capabilities, including the ability to recognize and interpret saliency and other verbal and nonverbal cues.

In this chapter, we propose a framework for anthropomorphic human-robot interaction based on human attention. Our framework comprises a human attention sub-system that

uses state-of-the-art video saliency prediction [2] and moving object detection and segmentation [6] models, as well as the Robot Operating System framework [18]. Finally, the behavior manager sub-system that controls the behavior of the Pepper robot [27] based on inputs from the human attention sub-system is implemented.

This chapter is structured in the following manner. The second section provides a concise overview of the state-of-the-art, followed by the third section which delves into the research methodology. The fourth section presents the experimental findings, while the concluding section offers a summary of the research work and highlights potential areas for future investigation.

4.2 State of the art

Currently, there is a vast and varied research and design effort focused on human-robot interaction [403]. HRI research can be broadly categorized into three areas : human-supervised HRI [404], autonomous HRI [405], and human-robot social interaction [406], where social robots engage with humans to accomplish specific objectives.

Human-supervised HRI involves human guidance to control robots during interactions. It is used in industrial settings for repetitive and dangerous tasks. The goal is to develop robots that can work alongside humans safely and efficiently [404]. Autonomous HRI, on the other hand, focuses on independent robot operation [405]. Both types of research focus on efficiency [403], while HRSI research aims for anthropomorphic and intuitive robot behavior [407]. This involves designing robots with human-like characteristics for social interaction and communication [1].

Human attention models are crucial for effective human-robot social interaction. The Attentional Intensity Model (AIM) [408] propose that attentional focus intensity determines perception efficiency and accuracy, while the integration of perceptual and action processes is essential for generating intuitive and adaptive behavior [409].

In the field of HRI, research has focused on improving social interaction through emotional expressions and intuitive recognition of human gestures. Emotional expressions can enhance social intelligence and naturalness of robot behavior [406], while recognizing and responding to human gestures is important for building a friendly and natural relationship [37, 38, 39]. Although these models have been studied to enhance HRI, they may not be sufficient to address all intuitiveness challenges associated with it. Emotions and gestures can be ambiguous and not always reflect a person's intentions or needs, creating

communication barriers.

Research has shown that human attention-based systems are a reliable and effective approach to improving HRI [274, 6]. These systems track a person's gaze, attention, and movement to better understand their needs and intentions, leading to more intuitive interactions. They are also more universal since they are less reliant on cultural or individual differences in emotional or gestural expression.

To this end, human attention has been utilized to establish joint attention between humans and robots. One method is to utilize eye-tracking technology to monitor a person's gaze direction, which offers valuable insight into where a person is directing their attention. Another approach is to use attention cues like head movements [410], gestures [37, 38, 39], or vocalizations [411], which can provide additional information about a person's attentional state and objectives.

Several studies have investigated the use of human attention cues in achieving joint attention between humans and robots. For example, [101] proposed a method of using pointing gestures and a saliency map to establish a joint focus of attention. In Saran et al. (2018), a deep learning approach that tracks a person's gaze in real time was proposed, using gaze heat map statistics to predict whether a person is facing the robot's camera [103]. Similarly, Shi et al. (2019) proposed an approach that uses Earth Mover's Distance to measure gaze similarity and predict which object a person is looking at [104]. Other studies have investigated the role of gaze cues in predicting a person's intentions, such as Huang et al. (2015) [412]. Additionally, Schillaci et al. (2013) proposed a saliency-based attentional model combined with attention manipulation skills to enable the robot to engage in an interactive game with objects as the first step towards joint attention [12]. Several studies have also explored the use of saliency models for joint attention, such as Schauerte et al. (2014) [105] and Frintrop et al. (2010) [413].

HRI researchers have found that joint attention is essential for human-robot interaction, but simply following a person's gaze is not enough. To achieve more intuitive and efficient interactions, robots need to model human attention in real-time, considering factors such as saliency and object permanence. Many current HRI systems lack this ability, leading to inefficient interactions. To address this, researchers suggest combining saliency prediction and moving object detection to better understand where humans are looking and which objects they are attending to in real-time, allowing robots to adjust their behavior accordingly [414].

Consequently, we propose a novel framework for anthropomorphic HRI based on hu-

man attention, specifically saliency prediction [2] and moving object detection [6]. Our approach aims to enhance the robot's understanding of potential user's focus of attention, which can facilitate more intuitive and anthropomorphic communication. By leveraging the latest advances in computer vision and machine learning, our framework can enable the robot to anticipate and respond to the user's needs, leading to a more seamless and personalized interaction. In the following section, we describe the key components of our framework and demonstrate its effectiveness through experiments and evaluations.

4.3 Proposal

4.3.1 Overview

Our anthropomorphic HRI framework is a sophisticated system that enables robots to interact with humans in a more natural and intuitive way. As it is shown in Figure 4.1, the framework consists of several key components that work together to create a seamless interaction between the robot and the human user.

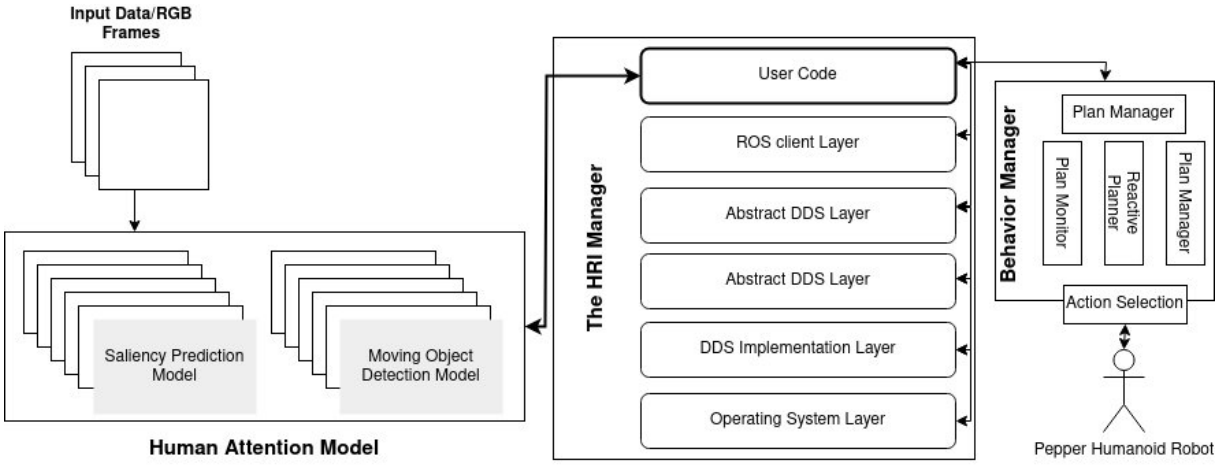


FIGURE 4.1 – Anthropomorphic human-robot interaction framework

At the heart of the system is a human attention model, that detects and highlights important regions and moving objects in the video feed, helping the robot understand the user's focus. State-of-the-art interactive video saliency prediction [2] and moving object detection and segmentation [6] models were adapted for this purpose.

The human-robot interaction manager is responsible for managing the interaction between the robot and the human. It uses a ROS framework to enable communication bet-

ween modules, robots, and humans [18]. ROS provides various tools for building complex robotic systems, allowing the robot to receive commands and respond naturally.

The behavior manager manages the robot's behavior based on the human attention model. It uses this information to determine the most appropriate response to the user's actions and ensure that the robot behaves consistently with the user's expectations.

Overall, the anthropomorphic human-robot interaction framework is a highly advanced system that enables robots to interact with humans efficiently and in a way that is natural and intuitive. The system is made up of several key components, including the human attention model, the HRI manager, and the behavior manager, which work together to create a seamless interaction between the robot and the human user.

4.3.2 Human Attention Model

Interactive Video Saliency Prediction Model

This is a novel video saliency prediction model [2] that utilizes stacked-ConvLSTM networks and convolutional networks. The model is specifically designed for video saliency prediction and is evaluated on the DHF1K dataset, which consists of 1,000 videos with human eye fixation annotations [19]. The architecture in Figure 4.1 includes an XY-shift frame differencing layer that generates a high-pass filtered map and a three-frame differencing method to enhance temporal features. The model fuses the VGG16 spatial features with the frame differencing output and passes them through a residual layer to create a single feature space. The stacked-ConvLSTM network extracts complex features and improves the robustness of the saliency prediction. The model is trained using sequential fixation and image data and three loss functions to enhance learning and generalization. A detailed desciption of this model and its architecture can be found in Chapter 2 of this document.

A Moving Object Detection and Segmentation Model

This is a novel XY-shift frame differencing technique and a three-stream encoder-decoder architecture for moving object detection and segmentation [6]. The XY-shift frame differencing reduces irrelevant background objects and exposes foreground objects, while the improved three-frame differencing extracts temporal features in the spatio-temporal domain. The three-stream encoder-decoder network consists of a VGG16-based encoder and decoder and uses binary cross-entropy as the loss function for training. The model is

trained using the CDNet-2014 dataset [20] with 4-fold cross-validation and early stopping to prevent overfitting. Adaptive Moment Estimation (Adam) is used to optimize the network during training. A detailed desciption of this model and its architecture can be found in Chapter 3 of this document.

4.3.3 Simulation Environment

The proposed framework is evaluated in two different settings : the Gazebo environment with ROS and the real Pepper robot. The experiment was conducted using a 3D model of the Pepper robot in the Gazebo environment, providing a realistic simulation to test the interactive video saliency prediction and moving object detection and segmentation models in a controlled environment. The process flow includes setting up the Gazebo environment, implementing the computer vision models as ROS nodes, developing a behavior manager module to control the virtual robot's camera movements, and evaluating the system's performance in different scenarios. This approach enables the testing and refinement of human attention models in a simulated environment before deploying them on a real Pepper robot.

4.3.4 Real Environment

The integration of the Pepper robot with ROS was made possible through the use of the ROSBridge protocol, which enables seamless communication between the robot and external systems. This integration was facilitated by the Naoqi driver, which enabled the Pepper robot to interact with ROS and utilize its vast array of tools and functionalities. The integration is accomplished through the ROS message protocol, enabling data exchange in a standardized format that is native to ROS. This approach offers several benefits including a more streamlined and efficient communication process and greater compatibility with ROS-based tools and functionalities.

We developed our human attention model using ROS nodes, dedicating nodes to computer vision, navigation, and the model itself. The human attention model utilizes input from the robot's vision sensors to determine where the robot's attention should be directed. This information is then communicated to the robot's actuator, specifically the wheel motors, to move the robot's body accordingly. To enable communication with the Pepper robot, we designed a ROS package for both ROS1 Noetic and ROS2 Humble versions, utilizing the Naoqi_driver. The package acquires image information in RGB format through

the `/image_raw` topic and sends linear and angular velocity commands to the `/cmd_vel` controller. This approach enables seamless integration of the human attention model with the Pepper robot, allowing for more intuitive and adaptive interactions.

4.4 Results

4.4.1 Models

Video Saliency Prediction Model

Our model was trained and evaluated on the DHF1K dataset, with the first 70% of the dataset used for training and evaluation. We used a 60/10/30 split ratio for training, validation, and testing, respectively, and randomly selected 420 and 70 videos for training and validation, respectively. To measure the model's performance, we employed five evaluation metrics, including Normalized Scanpath Saliency, Similarity Metric, Linear Correlation Coefficient, AUC-Judd, and shuffled AUC. Our video saliency prediction model exhibited an outstanding performance in all evaluation metrics when compared against 6 static saliency models adapted for video saliency prediction and 10 dynamic saliency models, demonstrating its effectiveness. For more detailed results on our proposed video saliency prediction model, we invite readers to refer to [2] and Chapter two 2 of this thesis.

Moving Object Detection and Segmentation

We used a dataset called CDNet-2014 [20] for training and evaluating our model. The dataset contains over 160,000 annotated frames in 53 video sequences, divided into 11 categories and two spectra : visible and thermal infrared. The scenes in the dataset include urban environments with people and cars, both indoor and outdoor, and with real-world challenges such as shadows, dynamic backgrounds, and camera motion. The ground truth for each image is a gray-scale image that describes four motion classes : static, hard shadow, unknown motion, and motion. An additional class is used for areas outside the region of interest. We evaluated our model on the testing sets of CDNet-2014, which contains 11 video sequences with almost 39,820 frames. We focused on the Recall, Precision, and F1-score metrics as they were deemed sufficient for the problem at hand. We evaluated our model against five state-of-the-art moving object detection and segmentation models. Consequently, our model outperformed all models in all evaluation

metrics. For more detailed results on our proposed video saliency prediction model, we invite readers to refer to [6] and Chapter three 3 of this thesis.

4.4.2 Simulation Environment

In order to assess the performance of the Pepper robot in various human-robot interaction scenarios, we designed and conducted an evaluation study using Gazebo platform. The study consisted of four distinct scenarios that aimed to simulate different real-world situations. The scenarios were as follows :

Moving Objects in a Closed Room In this scenario, we exposed the Pepper robot to a closed room environment where it encountered moving objects.

Interaction with Multiple Humans/Figures Pepper interacted with a combination of multiple humans and figures, some of which were moving and some interacted directly with the robot.

Operating in an Open Space with Visual Variations Pepper operated in an open space with various visual variations over time, simulating a dynamic environment.

Dynamic Environment with Humans and Moving Objects Pepper was left in a dynamic environment where it encountered both humans and other moving objects.

To ensure a diverse participant pool, we selected seven individuals from Addis Ababa University, considering their relevance to the discipline while also preserving gender variety. Each participant had the opportunity to observe the responses of the Pepper robot when it was placed in an environment relevant to each scenario.

To evaluate the performance of the Pepper robot, we employed the following metrics : Intuitiveness, Trust, Engagement, and User Satisfaction. To put the metrics into context,

Intuitiveness Intuitiveness refers to the ease with which users can understand and interact with the Pepper robot. It assesses how quickly and naturally users can comprehend the robot's behavior and functionalities. This metric is important because a robot that is intuitive to interact with can enhance user acceptance and engagement.

Trust Trust evaluates the level of confidence and reliance that users place in the robot. It reflects users' belief in the robot's competence, reliability, and adherence to safety measures. Trust is crucial in human-robot interaction as it influences user willingness to engage and collaborate with the robot.

Engagement Engagement measures the extent to which users are actively involved and interested in the interaction with the Pepper robot. It assesses the ability of the robot to captivate users' attention, generate interest, and promote a sense of involvement. High engagement levels indicate a successful interaction that keeps users engaged and attentive.

User Satisfaction User Satisfaction measures users' overall contentment and fulfillment with the interaction experience. It reflects users' subjective evaluation of the robot's performance, including aspects such as usability, usefulness, and meeting their expectations. User Satisfaction is a crucial metric as it provides insight into the overall acceptability and desirability of the human-robot interaction.

These metrics were chosen because they collectively provide a comprehensive evaluation of the human-robot interaction experience. By assessing Intuitiveness, Trust, Engagement, and User Satisfaction, we can gather valuable insights into different dimensions of the user experience, including ease of use, perceived reliability, user involvement, and overall satisfaction. By using these metrics, we aim to understand how well the Pepper robot performs in the evaluated scenarios, identify areas for improvement, and guide future enhancements in the design and implementation of human-robot interaction systems. The chosen metrics allow us to capture both objective and subjective aspects of the interaction, providing a holistic evaluation of the robot's performance from the user's perspective.

Data collection from the participants was conducted using a single form. We used HTML5, Bootstrap, and CSS3 for the front-end and Django framework for the backend. Participants were requested to assess the performance of the Pepper robot in each scenario using a rating scale ranging from 1 to 10. The scale was designed such that 1 represented poor performance, while 10 indicated the best performance achievable. Participants based their ratings on the predefined metrics established for the evaluation.This approach enabled us to gather valuable feedback from the participants, allowing for a quantitative assessment of the Pepper robot's performance across the different scenarios. Each perspective were managed to include exactly 10 objective questions. Later, the result from user's feedback is aggregated and averaged for quantitative representation. Details of the

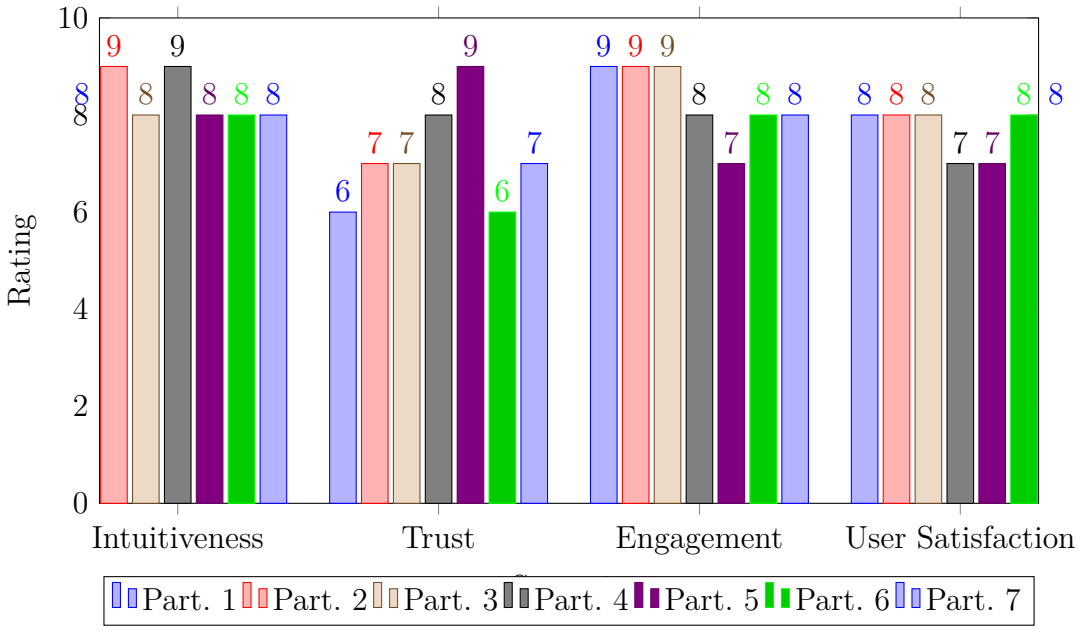


FIGURE 4.2 – Performance Ratings of Participants in Scenario 1

questionarie and results is put in the Chapter4.5 of this thesis document.

Accordingly, participants rated the performance of the robot in each scenario and teh aggregated result is presented in the following manner.

The bar chart in Figure 4.2 illustrates the performance ratings of different participants in scenario I. It was evaluated based on four metrics : Intuitiveness, Trust, Engagement, and User Satisfaction. The chart clearly depicts the variation in ratings across participants for each metric and scenario. Notably, Participant 3 consistently rated the Pepper robot highly in terms of Intuitiveness, Trust, Engagement, and User Satisfaction, while Participant 4 consistently provided lower ratings across all metrics. These findings highlight the individual differences in perception and experience among participants, emphasizing the importance of considering multiple perspectives when evaluating human-robot interaction scenarios.

The second scenarion in Figure 4.3 represents the performance ratings of participants while letting Pepper interact with multiple figures/humans and across four metrics : Intuitiveness, Trust, Engagement, and User Satisfaction. Each participant (Participant I to Participant VII) evaluated the performance of the robot according to the four metrics.

In terms of Intuitiveness, all participants rated the scenario highly, with ratings ranging from 8 to 10. Participant IV provided the highest rating of 10, indicating a strong perception of the scenario's intuitiveness.

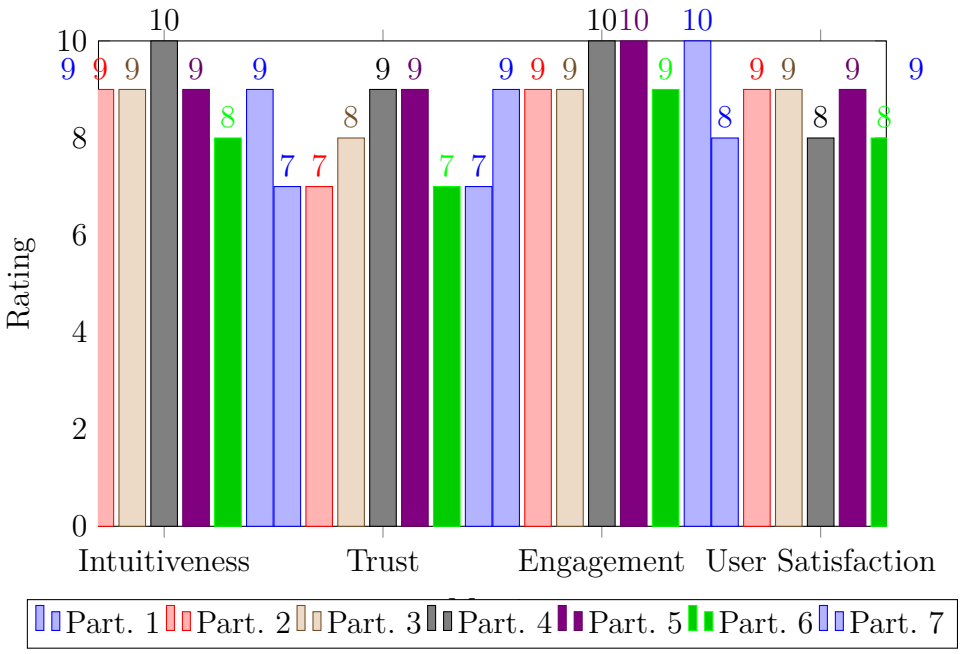


FIGURE 4.3 – Performance Ratings of Participants in Scenario II

Regarding Trust, the ratings were relatively consistent among participants, ranging from 7 to 9. Participants III, IV, and V provided higher ratings, suggesting a greater level of trust in the scenario compared to other participants.

For Engagement, the ratings were consistently high across all participants, ranging from 9 to 10. Participants IV, V, and VII rated the scenario with the highest engagement levels, reflecting a strong sense of involvement and interaction with the scenario.

In terms of User Satisfaction, the ratings ranged from 8 to 9, indicating a generally positive evaluation among the participants. Participants II and V provided the highest ratings for user satisfaction, highlighting their overall contentment with the scenario.

In Figure 4.4, the performance ratings of participants across four metrics : Intuitiveness, Trust, Engagement, and User Satisfaction is evaluated while spawning Pepper to operate in an open space with visual variations. Each participant (Participant I to Participant VII) evaluated based on the four metrics set.

In terms of Intuitiveness, participants provided consistently high ratings, with ratings ranging from 8 to 10. Participant I and Participant III rated the scenario with the highest level of intuitiveness, indicating a strong perception of the scenario's ease of understanding and use.

Regarding Trust, most participants rated the scenario with high levels of trust, with

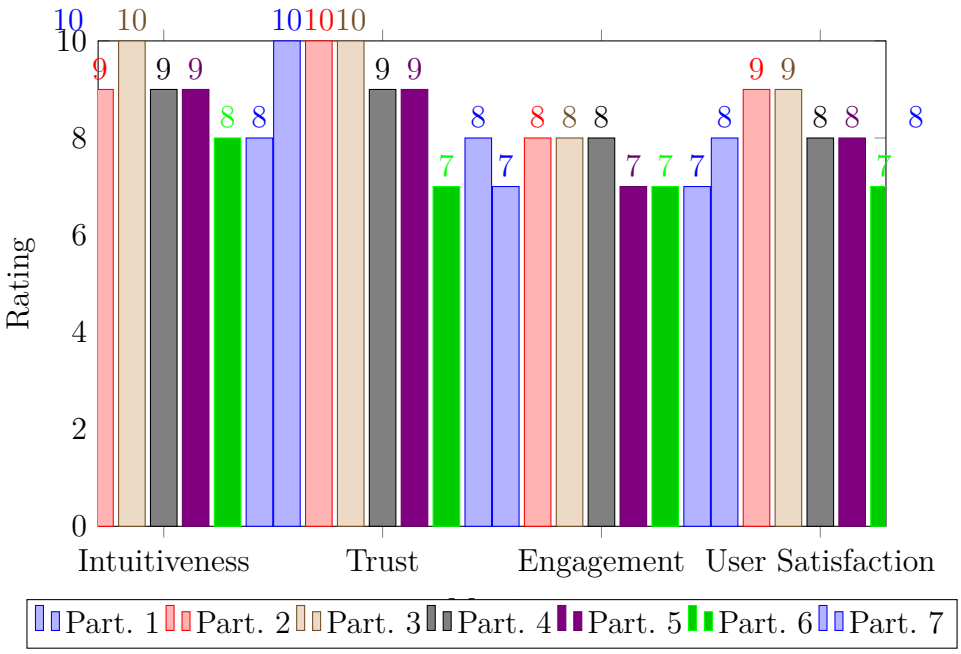


FIGURE 4.4 – Performance Ratings of Participants in Scenario III

ratings ranging from 9 to 10. Participants I, II, and III provided the highest ratings, indicating a high degree of confidence and reliance on the scenario.

For Engagement, the ratings varied among participants, ranging from 7 to 8. Participants I and II rated the scenario with the highest engagement levels, indicating a strong sense of involvement and interaction with the scenario, while other participants provided slightly lower ratings in terms of engagement.

In terms of User Satisfaction, the ratings ranged from 7 to 9. Participants II and III provided the highest ratings for user satisfaction, indicating their overall contentment with the scenario. Other participants also rated the scenario positively, reflecting a generally satisfactory experience.

The bar chart in 4.5 represents the performance ratings of participants in Scenario IV across four metrics : Intuitiveness, Trust, Engagement, and User Satisfaction. Each participant (Participant I to Participant VII) is evaluated based on their ratings for each metric.

In terms of Intuitiveness, the ratings varied among participants, ranging from 7 to 8. Participants II and V provided the highest ratings for intuitiveness, indicating a relatively high level of ease of understanding and use in the scenario.

Regarding Trust, most participants rated the scenario with moderate to high levels of

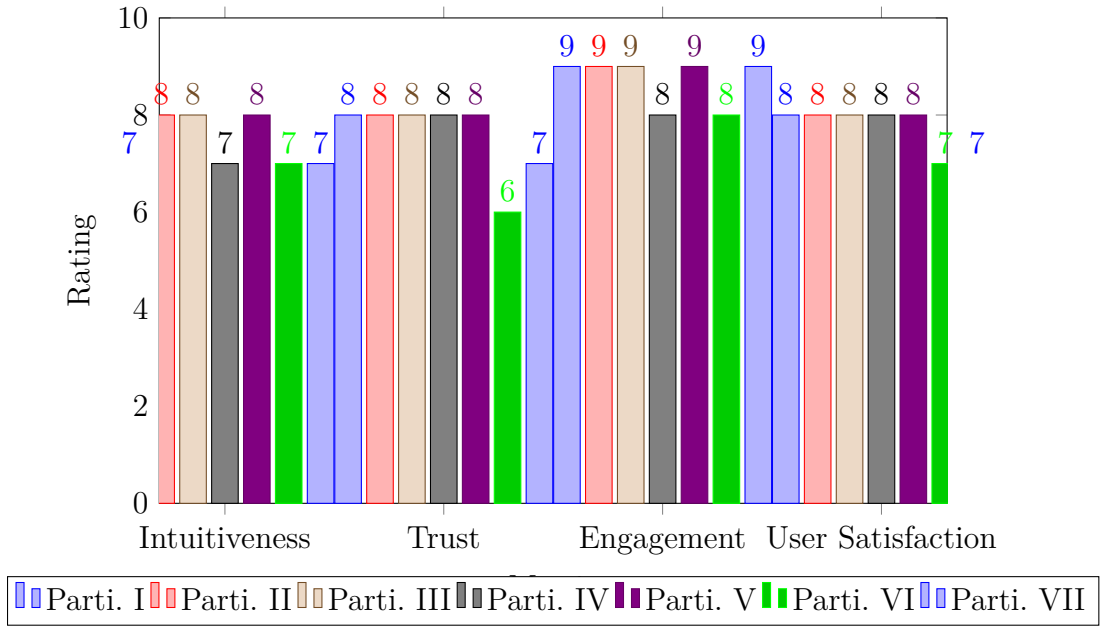


FIGURE 4.5 – Performance Ratings of Participants in Scenario IV

trust, with ratings ranging from 6 to 8. Participants I, II, III, IV, and V provided similar ratings, reflecting a reasonable level of confidence and reliance on the scenario.

For Engagement, the ratings were generally high among participants, ranging from 8 to 9. Participants I, II, III, V, and VII rated the scenario with the highest engagement levels, indicating a strong sense of involvement and interaction with the scenario.

In terms of User Satisfaction, the ratings ranged from 7 to 8. Participants II and III provided the highest ratings for user satisfaction, indicating their overall contentment with the scenario. Other participants also rated the scenario positively, reflecting a generally satisfactory experience.

As it is clearly shown in Figure 4.6, across all four scenarios, participants' evaluations were conducted based on four metrics : Intuitiveness, Trust, Engagement, and User Satisfaction. The ratings provided by participants varied for each metric across the different scenarios.

In terms of Intuitiveness, participants generally rated the scenarios with high scores, indicating that the scenarios were perceived as easy to understand and use. There was consistency in the high ratings across all scenarios, suggesting that the scenarios were well-designed in terms of intuitiveness.

For Trust, participants demonstrated a moderate to high level of confidence and reliance on the scenarios. The ratings for trust were generally positive, with some variations

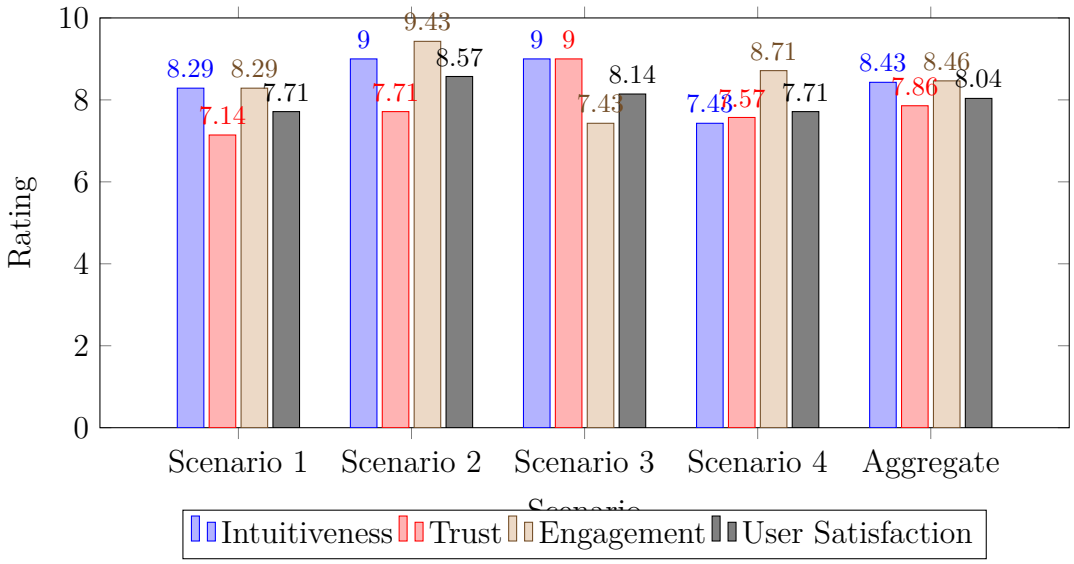


FIGURE 4.6 – Evaluation Ratings for Scenarios

among participants and scenarios. This indicates that the scenarios were able to instill a reasonable level of trust in the participants.

In terms of Engagement, participants were actively involved and interacted with the scenarios. The ratings for engagement were consistently high across all scenarios, indicating that the scenarios successfully captivated and engaged the participants.

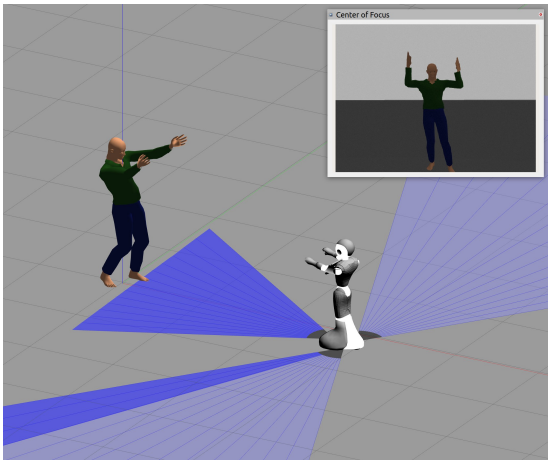
Regarding User Satisfaction, participants expressed overall satisfaction with their experiences in the scenarios. The ratings for user satisfaction were positive, suggesting that the scenarios met or exceeded participants' expectations and provided a satisfactory user experience.

To provide a visual representation of these scenarios, we included screenshots in Figure 4.7a [1], Figure 4.7b [2], Figure 4.7c [3], and Figure 4.7d [4].

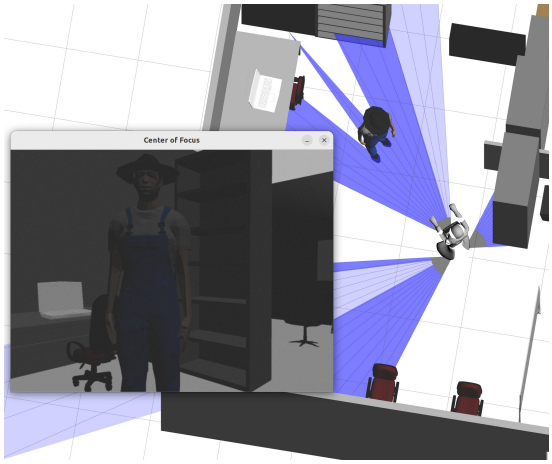
Hence, according to the results, our human attention based anthropomorphic human-robot interaction framework enabled intuitive, trustworthy, and fairly satisfying interaction capabilities with the virtual environment.

1. Pepper paying attention to moving bodies
2. Pepper interacting with temporarily salient body
3. Pepper acting humanly in a still environment
4. Pepper attending a very dynamic environment and on the move

(a) Moving Object Detection Scenario



(b) User Interaction Scenario



(c) Anthropomorphic Robot Behaviour Scenario

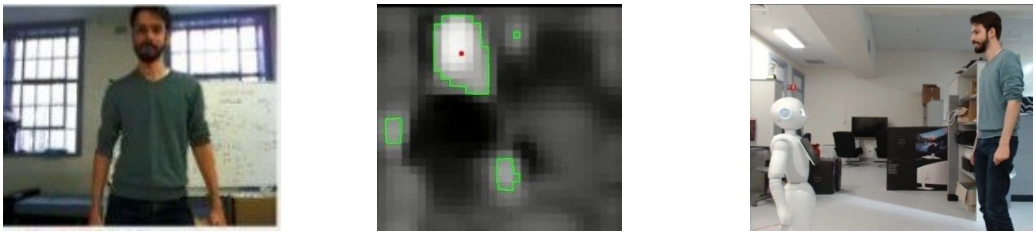
(d) Dynamic space Interaction Scenario

FIGURE 4.7 – Pepper Humanoid Robot Operating in Simulated Environment

4.4.3 Real-Time Embedded Strategy

Although we conducted an extensive evaluation in a simulated environment, we also performed a qualitative analysis of our HRI framework using a Pepper humanoid robot. The tests we conducted involved moving object detection using a moving object detection model, user interaction saliency model, and anthropomorphic robot action using both saliency and moving object detection models.

To demonstrate the efficiency of our approach in the real-time and low-resources constraints of the RoboCup@Home, we employed ourselves to optimize and deployed the architecture onboard the Pepper robot. To do so, models weights are first converted



(a) Pepper view (b) Real-time saliency (c) Global view

FIGURE 4.8 – Pepper Humanoid Robot Operating in the Real World : Human Interaction Scenario.

to Tensorflow Lite [5] format for lightweight inference on the robot CPU [6]. Then, a synchronizer is added to ensure low latency communication between the attention module, moving object module, and behavior module.

Figure 4.8 displays a screenshot of the Pepper humanoid robot operating with our anthropomorphic HRI framework.

4.5 Conclusion and Future work

In this pivotal chapter, we have delved into the realms of Anthropomorphic Human-Robot Interaction (HRI) by introducing a novel framework that harnesses the power of the human attention model. By successfully integrating and evaluating this framework across four distinct scenarios, we have obtained valuable insights into the intuitive, trustworthy, engaging, and satisfying nature of human-robot interactions.

Through extensive testing within simulated environments and rigorous functionality tests on the renowned humanoid robot Pepper, our research has laid a solid foundation for advancing the field of HRI and human attention models. By providing a comprehensive framework that enables robots to respond based on human attention cues, we have bridged the gap between humans and robots, unlocking a new realm of efficient and intuitive HRI systems.

The results obtained from the subjective ratings of seven carefully selected participants highlight the efficacy and potential of our framework. Each scenario served as a crucial stepping stone towards understanding and improving the intricate dynamics of human-robot interactions. By examining the participants' assessments of intuitiveness, trust,

5. `https://www.tensorflow.org/lite/guide?hl=en`
6. Intel Atom™ E3845 @ 1.91GHz x 4

engagement, and user satisfaction, we gained valuable insights into how our framework enhances the overall HRI experience.

Moreover, this research serves as the very heart of our thesis, representing the culmination of our efforts to fuse the futures of multiple attention models and elevate the state of the art in human-robot interaction. By pushing the boundaries of HRI research, we have not only created a tangible framework but also paved the way for future investigations and advancements in this exciting field.

As we conclude this chapter, it is essential to emphasize the significance of our work and its broader implications. Our research contributes to the research community by providing a robust framework that harmonizes human attention models and empowers robots to act in alignment with human cognitive patterns. This transformative capability opens doors to a myriad of potential applications and paves the way for a future where humans and robots seamlessly interact, collaborate, and coexist.

Looking ahead, we envision a range of promising research avenues that can build upon our foundation. These include exploring advanced cognitive models, refining the framework's adaptability to different environments, and investigating novel interaction techniques that further enhance the efficiency and intuitiveness of HRI systems.

In summary, this chapter represents a significant milestone in our thesis, where we have amalgamated cutting-edge attention models and propelled the field of HRI towards greater horizons. By delivering a comprehensive and innovative framework, we have established a strong foothold in the realm of intuitive and efficient human-robot interactions. Our research serves as a catalyst for future advancements in HRI, ushering in a new era of seamless collaboration between humans and robots.

In conclusion, this chapter presented an Anthropomorphic HRI framework that is based on the human attention model. The framework was tested and evaluated using subjective ratings from 7 participants in four different scenarios, assessing intuitiveness, trust, engagement, and user satisfaction. The framework was extensively tested in a simulated environment and underwent functionality tests on the real humanoid robot Pepper. This work contributes to the field of HRI and human attention models by providing a framework that can be used to guide robots on how to act based on human attention. Future research problems that might interest the research community in the area of HRI and human attention models were also highlighted. Overall, this work provides a significant step towards developing more efficient and intuitive HRI systems.

Conclusion

As we come to the conclusion of this thesis, we reflect on the journey that has led us here. Over the past several chapters, we have explored the field of Human-Robot Interaction and the role of human attention models in improving the performance of HRI systems. We have discussed the development and evaluation of attention models for both intuitive and efficient HRI, and the integration of these models into a framework for the Pepper Humanoid Robot.

This thesis has aimed to address the need for research on the application of attention models to HRI systems. Our objectives were to develop and test attention models as heuristic functions in HRI applications and to design an integrated framework to apply these models to HRI scenarios. We have made significant contributions in the areas of interactive video saliency prediction [2], moving object detection and segmentation [6], anthropomorphic [415] and efficient [416] HRI framework development.

In this concluding chapter, we will summarize our findings and contributions, discuss the limitations and future directions of this work, and provide some concluding remarks on the potential impact of our research in the field of HRI.

Summary of research problem and objectives

The research problem addressed in this thesis is the need for attention-based models in human-robot interaction to develop intuitive, anthropomorphic, and efficient HRI. The overarching objective of this thesis is to develop and evaluate attention-based models to improve the HRI performance of Pepper, a humanoid robot, in various scenarios.

To achieve this objective, the thesis proposed four specific research objectives. First, to develop an attention-based model for video saliency prediction using stacked ConvLSTM networks. Second, to develop a novel approach for moving object detection and segmentation using the XY shift frame differencing algorithm. Third, to develop an anthropomorphic HRI framework using a human attention approach, which involves integrating the attention-based models with other computer vision models and control algorithms to improve the overall performance of Pepper in various HRI scenarios.

The thesis also hypothesized that the developed attention-based models would significantly improve the performance of social standard robots like Pepper in various HRI scenarios, leading to more intuitive, anthropomorphic, and efficient HRI. To evaluate these hypotheses, the thesis conducted extensive experiments using large dynamic gaze fixation datasets for video saliency prediction and customised activity recognition datasets for moving object detection and segmentation. The results of these experiments provided strong evidence to support the hypotheses, as the attention-based models significantly improved the performance of Pepper in various HRI scenarios.

This thesis contribute to the field of HRI by addressing the critical need for attention-based models to improve the HRI performance of humanoid robots. The developed attention-based models have practical applications in various fields which are significantly represented by requirements set in RoboCup@Home Social Standard Robotics, and have the potential to revolutionize the way humans interact with robots.

Summary of research methodology

The research methodology used in this thesis involved a systematic approach to developing and evaluating attention-based models for HRI. The aim was to improve the performance of HRI systems by incorporating attention models as a source of intuitiveness and heuristics. The methodology included several stages, beginning with a review of the state-of-the-art to identify the gaps and challenges in the field of HRI.

The first stage involved the development of an attention-based video saliency prediction model. This model was developed using a stacked convLSTM approach and was trained on the DHF1K dataset. The performance of the model was evaluated using various metrics such as sAUC, AUC-Judd, CC, SIM, and NSS. The results showed that the attention-based model outperformed deep learning based video saliency prediction models in terms of accuracy and robustness.

The second stage focused on the development of a moving object detection and segmentation model using the XY shift frame differencing approach. The model was trained on the CDNet2014 dataset, and its performance was evaluated using various metrics such as precision, recall, and F1 score. The results showed that our moving object detection and segmentation model outperformed motion information based deep learning models for moving object detection and segmentation.

In the third stage, an anthropomorphic HRI framework was developed using a human

attention approach. The framework was based on the Robot Operating System framework and was designed to work with the Pepper Humanoid Robot. The attention model was integrated with other control algorithms to improve the anthropomorphic, intuitive nature of the robot in various HRI scenarios.

In the fourth stage, an efficient HRI framework was developed using video saliency-based heuristics for heavy machine learning models. We used various heavy computer vision models such as for face recognition, object detection and activity recognition. The framework was designed to improve the efficiency of HRI systems by reducing the computational load of heavy machine learning models. The attention-based heuristics were integrated with various computer vision models and control algorithms to improve the overall performance of the system in various HRI scenarios.

The research methodology used in this thesis also involved the use of large datasets, such as DHF1K [19] and CDNet2014 [20], for training and evaluating the attention-based models. The methodology included a rigorous evaluation of the performance of the models using various metrics. The results showed that the attention-based models outperformed traditional models in terms of accuracy and robustness.

In conclusion, the research methodology used in this thesis involved a systematic approach to developing and evaluating attention-based models for HRI. The methodology included several stages, beginning with a review of the literature and ending with the development of efficient HRI frameworks using attention-based heuristics. The use of large datasets and rigorous evaluation of the performance of the models using various metrics ensured the validity and reliability of the results. The contributions made by this thesis include the development of attention-based models for video saliency prediction, moving object detection and segmentation, as well as the development of anthropomorphic and efficient HRI frameworks using human attention and attention based heuristics. These contributions have the potential to revolutionize the field of HRI and have practical applications in various domains, such as RoboCup@Home Social Standard Robotics.

Summary of Findings and contributions to the field

Throughout this thesis, we have made significant contributions to the field of HRI and computer vision. Our research has explored the development and application of human attention based models to HRI scenarios, with a particular focus on video saliency prediction and moving object detection and segmentation.

In the state-of-the-art section, we began our review by discussing the significance of interactive machine learning in the field of human-robot interaction. We highlighted the growing interest in using machine learning techniques for interactive systems, which allows the robot to learn from user feedback and improve its performance over time. We then delved into the specific applications of machine learning in the context of HRI, including saliency prediction and moving object detection. These models are critical for enabling the robot to understand its environment, identify important objects or regions, and react accordingly. Finally, we discussed the impact of these models on the development of HRI systems. We explored the challenges associated with developing anthropomorphic HRI systems and how these models can help overcome these challenges. We highlighted the importance of using human attention models as a heuristic function for other computer vision models to improve the overall efficiency and accuracy of HRI systems.

As the second contribution of this thesis, we developed a stacked ConvLSTM approach for interactive video saliency prediction. Our proposed approach outperformed existing state-of-the-art methods in terms of accuracy and efficiency. Moreover, we demonstrated the potential of our model in improving the performance of various HRI scenarios, such as object grasping and following.

In the third part of our research, we proposed a novel XY shift frame differencing approach for moving object detection and segmentation. Our proposed method not only achieved high accuracy in detecting and segmenting moving objects but also showed promising results in terms of computational efficiency. Furthermore, we have shown that the proposed approach can be used in various HRI scenarios, such as tracking and navigation.

In the fourth part of our research, we developed an anthropomorphic HRI framework based on human attention models. We integrated the attention models with other computer vision models and control algorithms to improve the overall performance of the robot in various HRI scenarios. Our approach showed significant improvements in terms of intuitive and natural HRI, as well as the efficiency of the HRI system.

Lastly, we proposed a novel attention-based heuristic function for improving the efficiency of heavy machine learning models in HRI scenarios. Our approach showed promising results in reducing the computational cost of heavy models while maintaining their accuracy in various HRI tasks.

Put succinctly, our research has made significant contributions to the field of HRI and computer vision. We have shown the potential of attention-based models in improving the performance of HRI scenarios, particularly in terms of intuitiveness, anthropomorphism,

and efficiency. Our proposed methods have shown promising results in various HRI tasks, such as moving object detection and segmentation, object tracking, navigation, and following. Our contributions can address various real-world applications, especially those set by RoboCup@Home Social Standard Robotics.

Limitations and future research directions

While our research has made significant contributions to the field of HRI, there are also some limitations that must be acknowledged. One such limitation is the fact that our research solely focuses on video saliency prediction and moving object detection and segmentation problems based on spatio-temporal aspects of all datasets used. While this approach is effective in predicting areas of visual attention, it does not take into account the auditory aspects of human attention. A more comprehensive human-attention model should incorporate both visual and auditory cues to create a more accurate prediction of human attention.

Another limitation of our research is that we only used existing large dynamic gaze fixation datasets like DHF1K [19] and activity recognition datasets like CDNet2014 [20] for our experiments. Although these datasets have been widely used in the literature, they may not fully capture the complexities of real-world HRI scenarios. Future research should consider collecting more comprehensive and diverse datasets to ensure that HRI systems are accurately trained and tested.

Additionally, while our anthropomorphic HRI framework using human attention approach has shown promising results, there are still limitations in terms of the expressiveness and naturalness of the robot's behavior. Future research could focus on developing more sophisticated algorithms that can generate more natural and human-like behavior, while also considering cultural and individual differences in human behavior.

Lastly, while our efficient HRI approach using human attention models as heuristic function for other heavy computer vision models has shown significant improvements in performance, it is still limited by the processing power of current hardware. Future research could focus on developing more efficient algorithms that can be run on smaller and more portable devices, such as mobile phones or tablets.

In summary, while our research has made significant contributions to the field of HRI, there are still limitations that must be addressed in future research. By considering these limitations and continuing to develop more sophisticated algorithms and comprehensive

datasets, we can further advance the field of HRI and create more effective and natural human-robot interactions.

Conclusion and final remarks

In this thesis, we presented a comprehensive study of human attention models and their application to HRI scenarios. Our research aimed to develop an integrated framework for intuitive, anthropomorphic, and efficient HRI by leveraging the human attention models as heuristic functions.

We started our study with an extensive literature review in the state-of-the-art chapter, which provided a clear understanding of the current trends and challenges in the field of human attention modeling and its impact on HRI. Our research found that the integration of human attention models can significantly improve the performance of HRI systems, leading to more natural and efficient interactions between humans and robots.

To achieve our research objectives, we proposed two foundational research works. The first experimental chapter proposed a stacked convLSTM approach for interactive video saliency prediction. The proposed approach achieved state-of-the-art results on the DHF1K dataset, demonstrating its effectiveness in predicting the visual attention of humans in videos. The second experimental chapter proposed an XY shift frame differencing approach for moving object detection and segmentation, which also achieved state-of-the-art results on the CDNet2014 dataset.

We then developed an anthropomorphic HRI framework using the human attention approach, which integrated the attention models with other computer vision models and control algorithms to improve the overall performance of the robot in various HRI scenarios. The developed framework was based on the ROS and was designed to work with the Pepper Humanoid Robot and for RoboCup@Home Social Standard Robot setting.

Finally, we proposed an efficient human-robot social interaction through video saliency-based heuristics for heavy machine learning models. The proposed approach used the human attention models as heuristic functions to reduce the computational burden of heavy machine learning models, leading to faster and more efficient HRI systems.

However, our research has some limitations, such as the lack of consideration for the auditory aspects of human attention. Future research should explore the integration of auditory attention models into HRI systems to improve their performance further.

In conclusion, our research contributes to the field of HRI by providing a comprehensive

study of human attention models and their integration into HRI systems. Our proposed approaches achieved state-of-the-art results in video saliency prediction and moving object detection and segmentation. The developed anthropomorphic HRI framework and efficient human-robot social interaction through video saliency-based heuristics approach can significantly improve the performance of HRI systems, leading to more natural and efficient interactions between humans and robots.

PUBLICATIONS

The purpose of this chapter is to provide a comprehensive overview of the publications related to this thesis. The chapter will specifically focus on publications that have contributed to the research conducted in this thesis through their investigation of saliency prediction, moving object detection and segmentation, interactive machine learning, anthropomorphic HRI, and efficent HRI.

In our research, we have investigated the application of attention-based models to improve the human-robot interaction experience. Two of our experimental researches have specifically focused on saliency prediction and moving object detection and segmentation, which are crucial components in understanding human attention and behavior. These publications provide key insights into the development of attention-based models for robot perception, which is an important aspect of human-robot interaction.

In addition to the experimental research papers, a pre-print paper on the state of the art in interactive machine learning has been published. This paper also has a high correlation with the thesis due to its usability in the HRI environment.

Finally, our recently submitted results on anthropomorphic and efficient HRI using human attention models at the 2023's Robocup symposium represent the latest findings in our research. These results demonstrate the practical application of our attention-based approach for improving human-robot interactions.

List of Publications

Interactive Video Saliency Prediction : The Stacked-convLSTM Approach

— Title : Interactive Video Saliency Prediction : The Stacked-convLSTM Approach
— Authors : Natnael Wondimu, Ubbo Visser, and Cédric Buche
— Book Title : Proceedings of the 15th International Conference on Agents and Artificial Intelligence
— Volume : 2

— Publication date : 2023
— pages : 157-168
— Publisher : SciTePress
— organization : INSTICC
— DOI : 10.5220/0011664600003393
— isbn : 978-989-758-623-1
— issn : 2184-433X

A New Approach to Moving Object Detection and Segmentation : The XY-shift Frame Differencing

— Title : A New Approach to Moving Object Detection and Segmentation : The XY-shift Frame Differencing
— Authors : Natnael Wondimu, Ubbo Visser, and Cédric Buche
— Book Title : Proceedings of the 15th International Conference on Agents and Artificial Intelligence
— Volume : 3
— Publication date : 2023
— pages=309-318
— Publisher : SciTePress
— organization=INSTICC
— DOI : 10.5220/0011664500003393
— isbn=978-989-758-623-1
— issn=2184-433X

Saliency Prediction : Deep Learning Based Approach

— Title : Saliency Prediction : Deep Learning Based Approach
— Authors : Natnael Wondimu, Ubbo Visser, and Cédric Buche
— Book Title : Proceedings of the Black in AI workshop at the Conference on Neural Information Processing Systems (NeurIPS)
— Publication date : 2022
— Publisher : Conference Publishing Services
— organization= Neural Information Processing Systems Foundation,
— URL : https ://nips.cc/virtual/2022/63611

Interactive Machine Learning : A State of the Art Review

— Title : Interactive Machine Learning : A State of the Art Review
— Authors : Wondimu, N and Visser, U and Buche, Cédric
— Journal : arXiv preprint arXiv :2207.06196
— Publication date : 2022
— Publisher : Cornell Tech

Accepted Articles

Anthropomorphic Human-Robot Interaction Framework : Attention Based Approach

— Title : Anthropomorphic Human-Robot Interaction Framework : Attention Based Approach
— Authors : Natnael Wondimu, Maelic Neau, Antoine Dizet, Ubbo Visser, and Cédric Buche
— Book Title : Proceedings of the 2023 RoboCup Symposium
— Publication date : 2023
— Publisher : Springer
— Organization : RoboCup Federation

Enhanced Human-Robot Interaction through Spatio-Temporal Saliency Prediction

— Title : Enhanced Human-Robot Interaction through Spatio-Temporal Saliency Prediction
— Authors : Wondimu, N and Visser, U and Buche, Cédric
— Journal : Springer Nature of Computer Science
— Publication date : 2023
— Publisher : Springer
— Organization : Springer Nature

Position Papers

Efficient Human-Robot Social Interaction through Video Saliency-Based Heuristics for Heavy Machine Learning Models

— Title : Efficient Human-Robot Social Interaction through Video Saliency-Based Heuristics for Heavy Machine Learning Models
— Authors : Wondimu, N and Visser, U and Buche, Cédric

Enhancing Human-Robot Interaction in Social Standard Robotics through Multi-Modal Saliency Prediction : Audio-Visual Based Approach

— Title : Enhancing Human-Robot Interaction in Social Standard Robotics through Multi-Modal Saliency Prediction : Audio-Visual Based Approach
— Authors : Wondimu, N and Visser, U and Buche, Cédric

Affiliations

The following table lists the institutions or organizations that the authors are affiliated with.

Conclusion

This chapter has presented a comprehensive list of publications related to the thesis topic of Application of Interactive Machine Learning Models For Human-Robot Interaction : Attention-based Approach. The publications listed, which include experimental research papers, a pre-print paper, and two position papers, are significant contributions to the field of human-robot interaction and demonstrate the importance of incorporating human attention modeling in developing effective and efficient human-robot interaction systems.

The experimental research papers on saliency prediction and moving object detection and segmentation provide valuable insights into how humans attend to visual stimuli, and how this knowledge can be applied to the development of robotic systems. The pre-print

Photo	Name	Affiliations
	Natnael Argaw Wondimu	National Engineering School of Brest, Bretagne, France ; Addis Ababa University, Addis Ababa, Ethiopia
	Dr. Ubbo Visser	University of Miami, Florida, United States
	Prof. Cédric Buche	National Engineering School of Brest, Bretagne, France ; CROSSING, CNRS IRL 2010, SA, Australia ; Flinders University, SA, Australia
	Maelic Neau	National Engineering School of Brest, Bretagne, France ; CROSSING, CNRS IRL 2010, Australia ; Lab-STICC, CNRS UMR 6285, France ; Flinders University, Australia
	Antoine Dizet	Lab-STICC, CNRS UMR 6285, France ; Flinders University, Australia

paper on the state of the art in interactive machine learning contextualizes the research and highlights the relevance of this field to the thesis topic.

The position paper on efficient human-robot social interaction through video saliency-based heuristics for heavy machine learning models presents an innovative approach to improving the efficiency of human-robot interaction systems by incorporating video saliency-based heuristics. The other position paper on multi-modal saliency prediction with attentive convLSTM architecture presents a novel approach to saliency prediction, which includes both audio and visual cues and demonstrates the potential for multi-modal approaches to improve the performance of human-robot interaction systems.

In conclusion, the publications presented in this chapter provide a strong foundation for the proposed attention-based approach to human-robot interaction. They highlight the

importance of incorporating human attention modeling into the design of robotic systems. These publications, together with the proposed interactive machine learning models, have the potential to revolutionize the field of human-robot interaction.

APPENDIX

Human-Robot Interaction Framework Evaluation Questionnaire

Scenario I : Moving Objects in a Closed Room

Intuitiveness

1. How intuitive did you find the robot's responses when encountering a moving human in the closed room scenario ?
2. To what extent did the robot's behavior align with your expectations of how it should interact with a moving human in the environment ?
3. How easily did you understand the robot's intentions and actions when it encountered a moving human ?
4. Did the robot's reactions to a moving human enhance or hinder its overall intuitiveness in the scenario ?
5. How intuitive was the robot's navigation and movement when there was a moving human in the closed room ?
6. Did the robot's interactions with a moving human feel natural and effortless to comprehend ?
7. To what degree did the robot's behavior towards the moving human reflect a human-like understanding of the situation ?
8. How much did the robot's intuitive responses contribute to your willingness to trust and collaborate with it ?
9. Did the robot's intuitiveness play a significant role in making the overall interaction with the moving human more enjoyable and efficient ?
10. Considering the robot's intuitiveness in the closed room environment with a moving human, how likely are you to use or recommend it for similar tasks in the future ?

Trust

1. To what extent did you trust the robot's ability to identify and avoid obstacles in the closed room scenario ?
2. How confident were you in the robot's capability to navigate safely without colliding with any objects ?
3. Did the robot's responses to unexpected obstacles enhance your trust in its competence ?
4. How reliable did you find the robot's reactions when encountering moving objects in the closed room ?
5. To what degree did the robot's behavior align with your expectations of a safe and efficient interaction ?
6. How much trust did you place in the robot's ability to follow safety measures while moving through the closed room ?
7. Did the robot's performance in handling obstacles influence your overall trust in its capabilities ?
8. How much did you rely on the robot to handle its tasks autonomously without human intervention ?
9. To what extent did the robot's behavior foster a sense of confidence in its decision-making process ?
10. Considering the robot's performance in the closed room environment, how likely are you to collaborate with it in similar scenarios in the future ?

User Engagement

1. How engaged were you with the robot's interactions and movements in the closed room scenario ?
2. To what extent did the robot's behavior capture and maintain your attention throughout the interaction ?
3. Did the presence of a moving human in the closed room enhance or hinder your overall engagement with the robot ?
4. How much did the robot's responses to the moving objects contribute to a sense of active participation and involvement in the interaction ?

5. Did the robot's engagement strategies make you feel more immersed and connected with the task or environment ?
6. To what degree did the robot's actions spark your curiosity or interest during the closed room scenario ?
7. How enjoyable was the overall experience of interacting with the robot in the presence of moving objects, including humans ?
8. Did the robot's ability to adapt and respond to the environment's dynamic elements positively impact your engagement levels ?
9. How likely are you to seek out further interactions with the robot in similar closed room environments with moving objects ?
10. Considering your level of engagement with the robot, how inclined are you to recommend it to others for similar tasks or scenarios ?

User Satisfaction

1. How satisfied were you with the overall performance of the robot in the closed room scenario ?
2. To what extent did the robot meet your expectations regarding its interaction with moving objects, including humans ?
3. How well did the robot's behavior align with your desired outcomes and objectives during the interaction ?
4. Did the robot's responses to moving objects contribute significantly to your overall satisfaction with the interaction ?
5. How satisfied were you with the robot's ability to navigate safely and efficiently in the presence of moving objects ?
6. Did the robot's performance in the closed room environment, including its responses to a moving human, enhance your sense of satisfaction ?
7. To what degree did the robot's actions make you feel comfortable and at ease during the interaction ?
8. How satisfied were you with the robot's communication and interaction capabilities in the closed room scenario ?
9. Did the robot's performance in the presence of moving objects meet your criteria for a successful and satisfactory human-robot interaction ?

10. Considering your overall satisfaction with the robot's performance, how likely are you to use it again for similar tasks or recommend it to others ?

Scenario II : Interaction with Multiple Humans/Figures

Intuitiveness

1. How intuitive did you find the robot's responses when interacting with multiple moving humans and figures ?
2. To what extent did the robot's behavior align with your expectations of how it should interact with various moving and static entities ?
3. How easily did you understand the robot's intentions and actions during the complex interactions with multiple humans and figures ?
4. Did the robot's reactions to different human and figure interactions enhance or hinder its overall intuitiveness ?
5. How intuitive was the robot's navigation and movement while interacting with multiple dynamic elements ?
6. Did the robot's interactions with both moving and static entities feel natural and coherent to you ?
7. To what degree did the robot's behavior towards multiple humans and figures reflect a human-like understanding of the situation ?
8. How much did the robot's intuitive responses contribute to your willingness to trust and engage with it during the interactions ?
9. Did the robot's intuitiveness play a significant role in making the overall interaction with multiple humans and figures more enjoyable and efficient ?
10. Considering the robot's intuitiveness in handling complex interactions, how likely are you to use or recommend it for similar tasks in the future ?

Trust

1. To what extent did you trust the robot's ability to handle complex interactions with multiple moving humans and figures ?
2. How confident were you in the robot's competence and reliability during the interactions ?

3. Did the robot's responses to different human and figure interactions enhance your trust in its capabilities ?
4. How reliable did you find the robot's reactions when dealing with the variety of dynamic elements present during the interactions ?
5. To what degree did the robot's behavior foster a sense of confidence in its decision-making process during the interactions ?
6. How much trust did you place in the robot's ability to ensure safety while engaging with multiple humans and figures ?
7. Did the robot's performance in handling complex interactions influence your overall trust in its capabilities ?
8. How much did the robot's trustworthiness affect your willingness to collaborate and engage with it ?
9. To what extent did the robot's adherence to safety measures impact your trust during the interactions ?
10. Considering the robot's performance in complex interactions, how likely are you to trust and rely on it for similar tasks in the future ?

User Engagement

1. How engaged were you with the robot during the interactions with multiple moving humans and figures ?
2. To what extent did the robot's behavior capture and maintain your attention throughout the complex interactions ?
3. Did the presence of multiple dynamic elements enhance or hinder your overall engagement with the robot ?
4. How much did the robot's responses to different human and figure interactions contribute to a sense of active participation and involvement ?
5. Did the robot's engagement strategies make you feel more immersed and connected with the tasks or environment ?
6. To what degree did the robot's actions spark your curiosity or interest during the interactions ?
7. How enjoyable was the overall experience of interacting with the robot in the presence of multiple moving humans and figures ?

8. Did the robot's ability to adapt and respond to the dynamic elements positively impact your engagement levels ?
9. How likely are you to seek out further interactions with the robot in similar scenarios involving multiple humans and figures ?
10. Considering your level of engagement with the robot, how inclined are you to recommend it to others for similar tasks or scenarios ?

User Satisfaction

1. How satisfied were you with the overall performance of the robot during the interactions with multiple moving humans and figures ?
2. To what extent did the robot meet your expectations regarding its interaction with the complex mix of moving and static entities ?
3. How well did the robot's behavior align with your desired outcomes and objectives during the interactions ?
4. Did the robot's responses to different human and figure interactions significantly contribute to your overall satisfaction with the experience ?
5. How satisfied were you with the robot's ability to navigate and operate effectively while engaging with multiple dynamic elements ?
6. Did the robot's performance in the presence of multiple moving humans and figures meet your criteria for a successful and satisfactory interaction ?
7. To what degree did the robot's actions make you feel comfortable and satisfied during the interactions ?
8. How satisfied were you with the robot's communication and interaction capabilities in the context of complex interactions ?
9. Did the robot's performance in handling complex scenarios contribute significantly to your overall satisfaction ?
10. Considering your overall satisfaction with the robot's performance, how likely are you to use it again for similar tasks or recommend it to others ?

Scenario III : Operating in an Open Space with Visual Variations

Intuitiveness

1. How intuitive did you find the robot's responses when operating in the open space with various visual variations ?
2. To what extent did the robot's behavior align with your expectations of how it should interact in an open and visually diverse setting ?
3. How easily did you understand the robot's intentions and actions during the interactions with visual variations in the environment ?
4. Did the robot's reactions to different visual variations enhance or hinder its overall intuitiveness ?
5. How intuitive was the robot's navigation and movement when dealing with a variety of visual elements and changes in the environment ?
6. Did the robot's responses to visual variations feel natural and coherent to you ?
7. To what degree did the robot's behavior in the open space with visual variations reflect a human-like understanding of the situation ?
8. How much did the robot's intuitive responses contribute to your willingness to trust and engage with it during the interactions ?
9. Did the robot's intuitiveness play a significant role in making the overall interaction in the visually diverse environment more enjoyable and efficient ?
10. Considering the robot's intuitiveness in handling complex interactions with visual variations, how likely are you to use or recommend it for similar tasks in the future ?

Trust

1. To what extent did you trust the robot's ability to handle complex interactions in the open space with various visual variations ?
2. How confident were you in the robot's competence and reliability during the interactions with visual changes in the environment ?
3. Did the robot's responses to different visual variations enhance your trust in its capabilities ?
4. How reliable did you find the robot's reactions when dealing with the diverse visual elements present during the interactions ?

5. To what degree did the robot's behavior foster a sense of confidence in its decision-making process in the visually diverse environment ?
6. How much trust did you place in the robot's ability to ensure safety while operating in the open space with visual variations ?
7. Did the robot's performance in handling complex interactions with visual changes influence your overall trust in its capabilities ?
8. How much did the robot's trustworthiness affect your willingness to collaborate and engage with it ?
9. To what extent did the robot's adherence to safety measures impact your trust during the interactions in the visually diverse environment ?
10. Considering the robot's performance in complex interactions with visual variations, how likely are you to trust and rely on it for similar tasks in the future ?

User Engagement

1. How engaged were you with the robot during the interactions in the open space with various visual variations ?
2. To what extent did the robot's behavior capture and maintain your attention throughout the interactions with visual changes in the environment ?
3. Did the presence of various visual elements and changes enhance or hinder your overall engagement with the robot ?
4. How much did the robot's responses to different visual variations contribute to a sense of active participation and involvement ?
5. Did the robot's engagement strategies make you feel more immersed and connected with the tasks or environment in the visually diverse setting ?
6. To what degree did the robot's actions spark your curiosity or interest during the interactions with visual variations ?
7. How enjoyable was the overall experience of interacting with the robot in the open space with visual variations ?
8. Did the robot's ability to adapt and respond to the visual elements and changes positively impact your engagement levels ?
9. How likely are you to seek out further interactions with the robot in similar visually diverse scenarios ?

10. Considering your level of engagement with the robot, how inclined are you to recommend it to others for similar tasks or scenarios ?

User Satisfaction

1. How satisfied were you with the overall performance of the robot during the interactions in the open space with visual variations ?
2. To what extent did the robot meet your expectations regarding its interaction with the diverse visual elements and changes ?
3. How well did the robot's behavior align with your desired outcomes and objectives during the interactions in the visually diverse environment ?
4. Did the robot's responses to different visual variations significantly contribute to your overall satisfaction with the experience ?
5. How satisfied were you with the robot's ability to navigate and operate effectively in the visually diverse environment ?
6. Did the robot's performance in the open space with visual variations meet your criteria for a successful and satisfactory interaction ?
7. To what degree did the robot's actions make you feel comfortable and satisfied during the interactions ?
8. How satisfied were you with the robot's communication and interaction capabilities in the context of complex interactions with visual changes ?
9. Did the robot's performance in handling complex scenarios with visual variations contribute significantly to your overall satisfaction ?
10. Considering your overall satisfaction with the robot's performance, how likely are you to use it again for similar tasks or recommend it to others ?

Scenario IV : Dynamic Environment with Humans and Moving Objects

Intuitiveness

1. How intuitive did you find the robot's responses when interacting with both moving objects and humans in the dynamic environment ?

2. To what extent did the robot's behavior align with your expectations of how it should interact in a dynamic and unpredictable setting ?
3. How easily did you understand the robot's intentions and actions during the interactions with both humans and moving objects ?
4. Did the robot's reactions to different human and moving object interactions enhance or hinder its overall intuitiveness ?
5. How intuitive was the robot's navigation and movement when dealing with a diverse set of dynamic elements ?
6. Did the robot's interactions with both moving objects and humans feel natural and coherent to you ?
7. To what degree did the robot's behavior towards humans and moving objects reflect a human-like understanding of the situation ?
8. How much did the robot's intuitive responses contribute to your willingness to trust and engage with it during the interactions ?
9. Did the robot's intuitiveness play a significant role in making the overall interaction in the dynamic environment more enjoyable and efficient ?
10. Considering the robot's intuitiveness in handling complex interactions, how likely are you to use or recommend it for similar tasks in the future ?

Trust

1. To what extent did you trust the robot's ability to handle complex interactions with both humans and moving objects in the dynamic environment ?
2. How confident were you in the robot's competence and reliability during the interactions ?
3. Did the robot's responses to different human and moving object interactions enhance your trust in its capabilities ?
4. How reliable did you find the robot's reactions when dealing with the variety of dynamic elements present during the interactions ?
5. To what degree did the robot's behavior foster a sense of confidence in its decision-making process during the interactions ?
6. How much trust did you place in the robot's ability to ensure safety while engaging with both humans and moving objects ?

7. Did the robot's performance in handling complex interactions influence your overall trust in its capabilities ?
8. How much did the robot's trustworthiness affect your willingness to collaborate and engage with it ?
9. To what extent did the robot's adherence to safety measures impact your trust during the interactions ?
10. Considering the robot's performance in complex interactions, how likely are you to trust and rely on it for similar tasks in the future ?

User Engagement

1. How engaged were you with the robot during the interactions with both humans and moving objects in the dynamic environment ?
2. To what extent did the robot's behavior capture and maintain your attention throughout the complex interactions ?
3. Did the presence of both moving objects and humans enhance or hinder your overall engagement with the robot ?
4. How much did the robot's responses to different human and moving object interactions contribute to a sense of active participation and involvement ?
5. Did the robot's engagement strategies make you feel more immersed and connected with the tasks or environment ?
6. To what degree did the robot's actions spark your curiosity or interest during the interactions ?
7. How enjoyable was the overall experience of interacting with the robot in the dynamic environment with humans and moving objects ?
8. Did the robot's ability to adapt and respond to the dynamic elements positively impact your engagement levels ?
9. How likely are you to seek out further interactions with the robot in similar scenarios involving both humans and moving objects ?
10. Considering your level of engagement with the robot, how inclined are you to recommend it to others for similar tasks or scenarios ?

User Satisfaction

1. How satisfied were you with the overall performance of the robot during the interactions in the dynamic environment with humans and moving objects ?
2. To what extent did the robot meet your expectations regarding its interaction with the complex mix of humans and moving objects ?
3. How well did the robot's behavior align with your desired outcomes and objectives during the interactions ?
4. Did the robot's responses to different human and moving object interactions significantly contribute to your overall satisfaction with the experience ?
5. How satisfied were you with the robot's ability to navigate and operate effectively while engaging with a variety of dynamic elements ?
6. Did the robot's performance in the dynamic environment with humans and moving objects meet your criteria for a successful and satisfactory interaction ?
7. To what degree did the robot's actions make you feel comfortable and satisfied during the interactions ?
8. How satisfied were you with the robot's communication and interaction capabilities in the context of complex interactions ?
9. Did the robot's performance in handling complex scenarios contribute significantly to your overall satisfaction ?
10. Considering your overall satisfaction with the robot's performance, how likely are you to use it again for similar tasks or recommend it to others ?

Sources primaires

[2] N Wondimu, U Visser et Cédric Buche, « Interactive Video Saliency Prediction : The Stacked-convLSTM Approach », in : *15th International Conference on Agents and Artificial Intelligence*, SCITEPRESS-Science et Technology Publications, 2023, p. 157-168.

[3] Michael I Posner, « Attention in cognitive neuroscience : An overview », in : *The cognitive neurosciences*, The MIT Press, 1995, p. 615-624.

[6] N Wondimu, U Visser et Cédric Buche, « A New Approach to Moving Object Detection and Segmentation : The XY-shift Frame Differencing », in : *15th International Conference on Agents and Artificial Intelligence*, SCITEPRESS-Science et Technology Publications ; SCITEPRESS-Science and . . ., 2023, p. 309-318.

[7] Robert Desimone, « Visual attention mediated by biased competition in extrastriate visual cortex », in : *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences* 353.*1373* (1998), p. 1245-1255.

[8] Marvin M Chun et Yuhong Jiang, « Contextual cueing : Implicit learning and memory of visual context guides spatial attention », in : *Cognitive psychology* 36.*1* (1998), p. 28-71.

[11] Yu Du, Clarence W de Silva, Ming Cong, Dong Liu et Wenlong Qin, « An integrated model of visual attention for homecare robot with self-awareness », in : *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, 2015, p. 1752-1757.

[12] Guido Schillaci, Sasa Bodiroza et Verena Vanessa Hafner, « Evaluating the effect of saliency detection and attention manipulation in human-robot interaction », in : *International Journal of Social Robotics* 5 (2013), p. 139-152.

[13] Ali Borji et Laurent Itti, « State-of-the-art in visual attention modeling », in : *IEEE transactions on pattern analysis and machine intelligence* 35.*1* (2012), p. 185-207.

[14] Liang Wang, Weiming Hu et Tieniu Tan, « Recent developments in human motion analysis », in : *Pattern recognition* 36.*3* (2003), p. 585-601.

[16] Natalie Sebanz, Harold Bekkering et Gunther Knoblich, « Joint action : bodies and minds moving together », in : *Trends in cognitive sciences* 10.*2* (2006), p. 70-76.

[17] Paul Schweizer, « The truly total Turing test », in : *Minds and Machines* 8 (1998), p. 263-272.

[18] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler et Andrew Y Ng, « ROS : an open-source Robot Operating System », in : *ICRA workshop on open source software*, t. 3, Kobe, Japan, 2009, p. 5.

[19] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng et Ali Borji, « Revisiting video saliency : A large-scale benchmark and a new model », in : *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, p. 4894-4903.

[20] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth et Prakash Ishwar, « CDnet 2014 : An expanded change detection benchmark dataset », in : *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, p. 387-394.

[21] Marc Ellenfeld, Sebastian Moosbauer, Ruben Cardenes, Ulrich Klauck et Michael Teutsch, « Deep Fusion of Appearance and Frame Differencing for Motion Segmentation », in : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, p. 4339-4349.

[25] Saleema Amershi, Maya Cakmak, William Bradley Knox et Todd Kulesza, « Power to the people : The role of humans in interactive machine learning », in : *Ai Magazine* 35 (2014), p. 105-120.

[27] Amit Kumar Pandey et Rodolphe Gelin, « A mass-produced sociable humanoid robot : Pepper : The first machine of its kind », in : *IEEE Robotics & Automation Magazine* 25.*3* (2018), p. 40-48.

[28] Ben Goertzel, Julia Mossbridge, Eddie Monroe, David Hanson et Gino Yu, « Humanoid robots as agents of human consciousness expansion », in : *arXiv preprint arXiv :1709.07791* (2017).

[29] Nikolaos G Tsagarakis, Darwin G Caldwell, Francesca Negrello, Wooseok Choi, Lorenzo Baccelliere, Vo-Gia Loc, J Noorden, Luca Muratore, Alessio Margan et Alberto Cardellino, « Walk-man : A high-performance humanoid platform for realistic environments », in : *Journal of Field Robotics* 34.*7* (2017), p. 1225-1259.

[30] Ch Ott, Oliver Eiberger, Werner Friedl, B Bauml, Ulrich Hillenbrand, Ch Borst, Alin Albu-Schaffer, Bernhard Brunner, H Hirschmuller et S Kielhofer, « A humanoid two-arm system for dexterous manipulation », in : *2006 6th IEEE-RAS international conference on humanoid robots*, IEEE, 2006, p. 276-283.

[31] Kenji Kaneko, Kensuke Harada, Fumio Kanehiro, Go Miyamori et Kazuhiko Akachi, « Humanoid robot HRP-3 », in : *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2008, p. 2471-2478.

[32] Nicolaus A Radford, Philip Strawser, Kimberly Hambuchen, Joshua S Mehling, William K Verdeyen, A Stuart Donnan, James Holley, Jairo Sanchez, Vienny Nguyen et Lyndon Bridgwater, « Valkyrie : Nasa's first bipedal humanoid robot », in : *Journal of Field Robotics* 32.*3* (2015), p. 397-419.

[33] Yoshihiro Miyake et Hiroshi Shimizu, « Mutual entrainment based human-robot communication field-paradigm shift from" human interface" to" communication field" », in : *Proceedings of 1994 3rd IEEE International Workshop on Robot and Human Communication*, IEEE, 1994, p. 118-123.

[34] Maurizio Ficocelli, Junichi Terao et Goldie Nejat, « Promoting interactions between humans and robots using robotic emotional behavior », in : *IEEE transactions on cybernetics* 46.*12* (2015), p. 2911-2923.

[35] Hee-Deok Yang, A-Yeon Park et Seong-Whan Lee, « Gesture spotting and recognition for human–robot interaction », in : *IEEE Transactions on robotics* 23.*2* (2007), p. 256-270.

[36] Stefan Waldherr, Roseli Romero et Sebastian Thrun, « A gesture based interface for human-robot interaction », in : *Autonomous Robots* 9 (2000), p. 151-173.

[37] Maha Salem, Katharina Rohlfing, Stefan Kopp et Frank Joublin, « A friendly gesture : Investigating the effect of multimodal robot behavior in human-robot interaction », in : *2011 ro-man*, IEEE, 2011, p. 247-252.

[38] Xing Li, « Human–robot interaction based on gesture and movement recognition », in : *Signal Processing : Image Communication* 81 (2020), p. 115686.

[39] Aaron St Clair, Ross Mead et Maja J Mataric, « Investigating the effects of visual saliency on deictic gesture production by a humanoid robot », in : *2011 RO-MAN*, IEEE, 2011, p. 210-216.

[40] Aaron Chau, Kouhei Sekiguchi, Aditya Arie Nugraha, Kazuyoshi Yoshii et Kotaro Funakoshi, « Audio-Visual SLAM towards Human Tracking and Human-Robot Interaction in Indoor Environments », in : *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019, p. 1-8, doi : 10.1109/RO-MAN46459.2019.8956321.

[41] Julie A Adams, Ruzena Bajcsy, Jana Kosecka, Vijay Kumar, Robert Mandelbaum, Max Mintz, R Paul, Curtis Wang, Yoshio Yamamoto et Xiaoping Yun, « Cooperative material handling by human and robotic agents : Module development and system synthesis », in : *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, t. 1, IEEE, 1995, p. 200-205.

[42] Roland S Johansson, « Sensory input and control of grip », in : *Novartis Foundation Symposium 218-Sensory Guidance of Movement : Sensory Guidance of Movement : Novartis Foundation Symposium 218*, Wiley Online Library, 2007, p. 45-63.

[43] Jaehyun Shim, « Haptic cues in bimanual cooperative transport of large objects », thèse de doct., University of British Columbia, 2018.

[44] Allison M Okamura, *Haptic dimensions of human-robot interaction*, 2018.

[45] George Leifman, Dmitry Rudoy, Tristan Swedish, Eduardo Bayro-Corrochano et Ramesh Raskar, « Learning gaze transitions from depth to improve video saliency estimation », in : *Proceedings of the IEEE International Conference on Computer Vision*, 2017, p. 1698-1707.

[46] Volker Klingspor, John Demiris et Michael Kaiser, « Human-robot communication and machine learning », in : *Applied Artificial Intelligence* 11.7 (1997), p. 719-746.

[48] Kyle B Reed et Michael A Peshkin, « Physical collaboration of human-human and human-robot teams », in : *IEEE transactions on haptics* 1.2 (2008), p. 108-120.

[49] Guillaume Morel, Ezio Malis et Sylvie Boudet, « Impedance based combination of visual and force control », in : *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No. 98CH36146)*, t. 2, IEEE, 1998, p. 1743-1748.

[50] Dennis Perzanowski, Alan C Schultz et William Adams, « Integrating natural language and gesture in a robotics domain », in : *Proceedings of the 1998 IEEE International Symposium on Intelligent Control (ISIC) held jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA) Intell*, IEEE, 1998, p. 247-252.

[51] Raveesh Meena, Kristiina Jokinen et Graham Wilcock, « Integration of gestures and speech in human-robot interaction », in : *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, IEEE, 2012, p. 673-678.

[52] Edgar Seemann, Kai Nickel et Rainer Stiefelhagen, « Head pose estimation using stereo vision for human-robot interaction », in : *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* IEEE, 2004, p. 626-631.

[53] Brian Reily, Fei Han, Lynne E Parker et Hao Zhang, « Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction », in : *Autonomous Robots* 42 (2018), p. 1281-1298.

[54] Kenji Sakita, Koichi Ogawara, Shinji Murakami, Kentaro Kawamura et Katsushi Ikeuchi, « Flexible cooperation between human and robot by interpreting human intention from gaze information », in : *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, t. 1, IEEE, 2004, p. 846-851.

[55] Kelsey P Hawkins, Nam Vo, Shray Bansal et Aaron F Bobick, « Probabilistic human action prediction and wait-sensitive planning for responsive human-robot collaboration », in : *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, IEEE, 2013, p. 499-506.

[56] Umme Zakia et Carlo Menon, « Human–robot collaboration in 3D via force myography based interactive force estimations using cross-domain generalization », in : *IEEE Access* 10 (2022), p. 35835-35845.

[57] Don Joven Agravante, Andrea Cherubini, Antoine Bussy, Pierre Gergondet et Abderrahmane Kheddar, « Collaborative human-humanoid carrying using vision and haptic sensing », in : *2014 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2014, p. 607-612.

[58] Leonel Rozo, Danilo Bruno, Sylvain Calinon et Darwin G Caldwell, « Learning optimal controllers in human-robot cooperative transportation tasks with position and force constraints », in : *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, IEEE, 2015, p. 1024-1030.

[59] Vincent Duchaine et Clement M Gosselin, « General model of human-robot cooperation using a novel velocity based variable impedance control », in : *Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'07)*, IEEE, 2007, p. 446-451.

[60] Ryojun Ikeura et Hikaru Inooka, « Variable impedance control of a robot for cooperation with a human », in : *Proceedings of 1995 IEEE International Conference on Robotics and Automation*, t. 3, IEEE, 1995, p. 3097-3102.

[61] Toru Tsumugiwa, Ryuichi Yokogawa et Kei Hara, « Variable impedance control with virtual stiffness for human-robot cooperative peg-in-hole task », in : *IEEE/RSJ international conference on intelligent robots and systems*, t. 2, IEEE, 2002, p. 1075-1081.

[62] Paul Evrard, Elena Gribovskaya, Sylvain Calinon, Aude Billard et Abderrahmane Kheddar, « Teaching physical collaborative tasks : object-lifting case study with a humanoid », in : *2009 9th IEEE-RAS international conference on humanoid robots*, IEEE, 2009, p. 399-404.

[63] Paul Evrard et Abderrahmane Kheddar, « Homotopy switching model for dyad haptic interaction in physical collaborative tasks », in : *World haptics 2009-third joint EuroHaptics conference and symposium on haptic interfaces for virtual environment and teleoperator systems*, IEEE, 2009, p. 45-50.

[64] Andrej Gams, Bojan Nemec, Jan Ijspeert Auke et Ales Ude, « Coupling movement primitives : Interaction with the environment and bimanual tasks », in : *IEEE Transactions on Robotics* 30.*4* (2014), p. 816-830.

[65] Ivana Palunko, Philine Donner, Martin Buss et Sandra Hirche, « Cooperative suspended object manipulation using reinforcement learning and energy-based control », in : *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2014, p. 885-891.

[66] Philine Donner et Martin Buss, « Cooperative swinging of complex pendulum-like objects : Experimental evaluation », in : *IEEE Transactions on Robotics* 32.*3* (2016), p. 744-753.

[67] Luka Peternel et Jan Babic, « Learning of compliant human–robot interaction using full-body haptic interface », in : *Advanced Robotics* 27.*13* (2013), p. 1003-1012.

[68] Shuhei Ikemoto, Heni Ben Amor, Takashi Minato, Bernhard Jung et Hiroshi Ishiguro, « Physical human-robot interaction : Mutual learning and adaptation », in : *IEEE robotics & automation magazine* 19.*4* (2012), p. 24-35.

[69] Jeremy A Marvel, Shelly Bagchi, Megan Zimmerman et Brian Antonishek, « Towards effective interface designs for collaborative HRI in manufacturing : Metrics and measures », in : *ACM Transactions on Human-Robot Interaction (THRI)* 9.*4* (2020), p. 1-55.

[70] Jorg Kruger, Terje K Lien et Alexander Verl, « Cooperation of human and machines in assembly lines », in : *CIRP annals* 58.*2* (2009), p. 628-646.

[71] Kangwagye Samuel, Kevin Haninger et Sehoon Oh, « High-Performance Admittance Control of An Industrial Robot Via Disturbance Observer », in : *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society*, IEEE, 2022, p. 1-6.

[72] Luka Peternel, Tadej Petric, Erhan Oztop et Jan Babic, « Teaching robots to cooperate with humans in dynamic manipulation tasks based on multi-modal human-in-the-loop approach », in : *Autonomous robots* 36 (2014), p. 123-136.

[73] Yu Zhuang, Shaowei Yao, Chenming Ma et Rong Song, « Admittance control based on EMG-driven musculoskeletal model improves the human–robot synchronization », in : *IEEE Transactions on Industrial Informatics* 15.*2* (2018), p. 1211-1218.

[74] Achim Buerkle, William Eaton, Niels Lohse, Thomas Bamber et Pedro Ferreira, « EEG based arm movement intention recognition towards enhanced safety in symbiotic Human-Robot Collaboration », in : *Robotics and Computer-Integrated Manufacturing* 70 (2021), p. 102137.

[75] Iolanda Leite, Rui Henriques, Carlos Martinho et Ana Paiva, « Sensors in the wild : Exploring electrodermal activity in child-robot interaction », in : *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2013, p. 41-48.

[76] Pramila Rani, Changchun Liu, Nilanjan Sarkar et Eric Vanman, « An empirical study of machine learning techniques for affect recognition in human–robot interaction », in : *Pattern Analysis and Applications* 9 (2006), p. 58-69.

[77] Christian Cipriani, Franco Zaccone, Silvestro Micera et M Chiara Carrozza, « On the shared control of an EMG-controlled prosthetic hand : analysis of user–prosthesis interaction », in : *IEEE Transactions on Robotics* 24.*1* (2008), p. 170-184.

[78] Francis HY Chan, Yong-Sheng Yang, FK Lam, Yuan-Ting Zhang et Philip A Parker, « Fuzzy EMG classification for prosthesis control », in : *IEEE transactions on rehabilitation engineering* 8.*3* (2000), p. 305-311.

[79] Tommaso Lenzi, Stefano Marco Maria De Rossi, Nicola Vitiello et Maria Chiara Carrozza, « Intention-based EMG control for powered exoskeletons », in : *IEEE transactions on biomedical engineering* 59.*8* (2012), p. 2180-2190.

[80] Christian Fleischer, Andreas Wege, Konstantin Kondak et Günter Hommel, *Application of EMG signals for controlling exoskeleton robots*, 2006, doi : doi : 10.1515/BMT.2006.063.

[81] Jorn Vogel, Claudio Castellini et Patrick van der Smagt, « EMG-based teleoperation and manipulation with the DLR LWR-III », in : *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2011, p. 672-678.

[82] Luka Peternel, Nikos Tsagarakis et Arash Ajoudani, « Towards multi-modal intention interfaces for human-robot co-manipulation », in : *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, IEEE, 2016, p. 2663-2669.

[83] Christian J Bell, Pradeep Shenoy, Rawichote Chalodhorn et Rajesh PN Rao, « Control of a humanoid robot by a noninvasive brain–computer interface in humans », in : *Journal of neural engineering* 5.*2* (2008), p. 214.

[84] Daniel Szafir et Bilge Mutlu, « Pay attention ! Designing adaptive agents that monitor and improve user engagement », in : *Proceedings of the SIGCHI conference on human factors in computing systems*, 2012, p. 11-20.

[85] Stefan Ehrlich, Agnieszka Wykowska, Karinne Ramirez-Amaro et Gordon Cheng, « When to engage in interaction—And how ? EEG-based enhancement of robot's ability to sense social signals in HRI », in : *2014 IEEE-RAS International Conference on Humanoid Robots*, IEEE, 2014, p. 1104-1109.

[86] Pramila Rani, Nilanjan Sarkar, Craig A Smith et Leslie D Kirby, « Anxiety detecting robotic system–towards implicit human-robot collaboration », in : *Robotica* 22.*1* (2004), p. 85-95.

[87] Luka Peternel, Nikos Tsagarakis, Darwin Caldwell et Arash Ajoudani, « Adaptation of robot physical behaviour to human fatigue in human-robot comanipulation », in : *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, IEEE, 2016, p. 489-494.

[88] Juan Antonio Corrales, Francisco A Candelas et Fernando Torres, « Hybrid tracking of human operators using IMU/UWB data fusion by a Kalman filter », in : *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 2008, p. 193-200.

[90] Gizem Ates, Martin Fodstad Stolen et Erik Kyrkjebo, « Force and gesture-based motion control of human-robot cooperative lifting using imus », in : *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2022, p. 688-692.

[91] Philipp Mittendorfer, Eiichi Yoshida et Gordon Cheng, « Realizing whole-body tactile interactions with a self-organizing, multi-modal artificial skin on a humanoid robot », in : *Advanced Robotics* 29.*1* (2015), p. 51-67.

[92] Leonel Rozo, Sylvain Calinon, Darwin G. Caldwell, Pablo Jiménez et Carme Torras, « Learning Physical Collaborative Robot Behaviors From Human Demonstrations », in : *IEEE Transactions on Robotics* 32.*3* (2016), p. 513-527, doi : 10.1109/TRO.2016.2540623.

[94] Chenguang Yang, Peidong Liang, Arash Ajoudani, Zhijun Li et Antonio Bicchi, « Development of a robotic teaching interface for human to human skill transfer », in : *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2016, p. 710-716.

[95] Serena Ivaldi, Sebastien Lefort, Jan Peters, Mohamed Chetouani, Joelle Provasi et Elisabetta Zibetti, « Towards engagement models that consider individual factors in HRI : on the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task : experiments with the iCub humanoid », in : *International Journal of Social Robotics* 9 (2017), p. 63-86.

[97] Stephanie Lackey, Daniel Barber, Lauren Reinerman, Norman I Badler et Irwin Hudson, « Defining next-generation multi-modal communication in human robot interaction », in : *Proceedings of the human factors and ergonomics society annual meeting*, SAGE Publications Sage CA : Los Angeles, CA, 2011, p. 461-464.

[98] Amelie Legeleux, Cedric Buche et Dominique Duhaut, « Gaussian Mixture Model with Weighted Data for Learning by Demonstration », in : *The International FLAIRS Conference Proceedings*, t. 35, 2022.

[99] Martin Lawitzky, Jose Ramon Medina, Dongheui Lee et Sandra Hirche, « Feedback motion planning and learning from demonstration in physical robotic assistance : differences and synergies », in : *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, p. 3646-3652.

[100] Michael Tomasello, *Why we cooperate*, MIT press, 2009.

[101] Boris Schauerte et Gernot A Fink, « Focusing computational visual attention in multi-modal human-robot interaction », in : *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, 2010, p. 1-8.

[102] Serena Ivaldi, Salvatore M Anzalone, Woody Rousseau, Olivier Sigaud et Mohamed Chetouani, « Robot initiative in a team learning task increases the rhythm of interaction but not the perceived engagement », in : *Frontiers in neurorobotics* 8 (2014), p. 5.

[103] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz et Scott Niekum, « Human gaze following for human-robot interaction », in : *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, p. 8615-8621.

[104] Lei Shi, Cosmin Copot et Steve Vanlanduit, « What are you looking at ? detecting human intention in gaze based human-robot interaction », in : *arXiv preprint arXiv :1909.07953* (2019).

[105] Boris Schauerte et Rainer Stiefelhagen, « "Look at this !" learning to guide visual saliency in human-robot interaction », in : *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2014, p. 995-1002.

[106] Shuwen Qiu, Hangxin Liu, Zeyu Zhang, Yixin Zhu et Song-Chun Zhu, « Human-robot interaction in a shared augmented reality workspace », in : *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, p. 11413-11418.

[107] Michael Walker, Hooman Hedayati, Jennifer Lee et Daniel Szafir, « Communicating robot motion intent with augmented reality », in : *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, p. 316-324.

[108] Scott A Green, J Geoffrey Chase, XiaoQi Chen et Mark Billinghurst, « Evaluating the augmented reality human-robot collaboration system », in : *International journal of intelligent systems technologies and applications* 8.*1-4* (2010), p. 130-143.

[109] John Glassmire, Marcia O'Malley, William Bluethmann et Robert Ambrose, « Cooperative manipulation between humans and teleoperated agents », in : *12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2004. HAPTICS'04. Proceedings.* IEEE, 2004, p. 114-120.

[110] Florian Leutert, Christian Herrmann et Klaus Schilling, « A spatial augmented reality system for intuitive display of robotic data », in : *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2013, p. 179-180.

[111] Filippo Brizzi, Lorenzo Peppoloni, Alessandro Graziano, Erika Di Stefano, Carlo Alberto Avizzano et Emanuele Ruffaldi, « Effects of augmented reality on the performance of teleoperated industrial assembly tasks in a robotic embodiment », in : *IEEE Transactions on Human-Machine Systems* 48.*2* (2017), p. 197-206.

[112] HC Fang, SK Ong et AYC Nee, « A novel augmented reality-based interface for robot path planning », in : *International Journal on Interactive Design and Manufacturing (IJIDeM)* 8 (2014), p. 33-42.

[114] Taeho Jo, « Machine Learning Foundations », in : *Supervised, Unsupervised, and Advanced Learning. Cham : Springer International Publishing* (2021).

[115] Wenqiang Chi, Jindong Liu, Hedyeh Rafii-Tari, Celia Riga, Colin Bicknell et Guang-Zhong Yang, « Learning-based endovascular navigation through the use of non-rigid registration for collaborative robotic catheterization », in : *International journal of computer assisted radiology and surgery* 13 (2018), p. 855-864.

[116] Yanlong Huang, Joao Silverio, Leonel Rozo et Darwin G Caldwell, « Generalized task-parameterized skill learning », in : *2018 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2018, p. 5667-5474.

[117] Sean R Eddy, « What is a hidden Markov model ? », in : *Nature biotechnology* 22.*10* (2004), p. 1315-1316.

[118] Elena Corina Grigore, Alessandro Roncone, Olivier Mangin et Brian Scassellati, « Preference-based assistance prediction for human-robot collaboration tasks », in : *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, p. 4441-4448.

[119] Leonel Rozo, Joao Silverio, Sylvain Calinon et Darwin G Caldwell, « Learning controllers for reactive and proactive behaviors in human–robot collaboration », in : *Frontiers in Robotics and AI* 3 (2016), p. 30.

[120] David Vogt, Simon Stepputtis, Richard Weinhold, Bernhard Jung et Heni Ben Amor, « Learning human-robot interactions from human-human demonstrations (with applications in lego rocket assembly) », in : *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, IEEE, 2016, p. 142-143.

[121] Shingo Murata, Wataru Masuda, Jiayi Chen, Hiroaki Arie, Tetsuya Ogata et Shigeki Sugano, « Achieving human–robot collaboration with dynamic goal inference by gradient descent », in : *Neural Information Processing : 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part II 26*, Springer, 2019, p. 579-590.

[122] Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu et Julie Shah, « Efficient model learning from joint-action demonstrations for human-robot collaborative tasks », in : *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, p. 189-196.

[123] Ali Ghadirzadeh, Xi Chen, Wenjie Yin, Zhengrong Yi, Maarten Bjorkman et Danica Kragic, « Human-centered collaborative robots with deep reinforcement learning », in : *IEEE Robotics and Automation Letters* 6.*2* (2020), p. 566-571.

[124] Ali Ghadirzadeh, Judith Butepage, Atsuto Maki, Danica Kragic et Maarten Bjorkman, « A sensorimotor reinforcement learning framework for physical human-robot interaction », in : *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2016, p. 2682-2688.

[125] Min Wu, Yanhao He et Steven Liu, « Shared impedance control based on reinforcement learning in a human-robot collaboration task », in : *Advances in Service and Industrial Robotics : Proceedings of the 28th International Conference on Robotics in Alpe-Adria-Danube Region (RAAD 2019) 28*, Springer, 2020, p. 95-103.

[126] Min Wu, Yanhao He et Steven Liu, « Adaptive impedance control based on reinforcement learning in a human-robot collaboration task with human reference estimation », in : *Int J Mech Control* 21.*1* (2020), p. 21-31.

[127] Zhen Deng, Jinpeng Mi, Dong Han, Rui Huang, Xiaofeng Xiong et Jianwei Zhang, « Hierarchical robot learning for physical collaboration between humans and robots », in : *2017 IEEE international conference on robotics and biomimetics (robio)*, IEEE, 2017, p. 750-755.

[128] Weifeng Lu, Zhe Hu et Jia Pan, « Human-robot collaboration using variable admittance control and human intention prediction », in : *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2020, p. 1116-1121.

[129] Weitian Wang, Rui Li, Yi Chen, Z Max Diekel et Yunyi Jia, « Facilitating human–robot collaborative tasks by teaching-learning-collaboration from human demonstrations », in : *IEEE Transactions on Automation Science and Engineering* 16.*2* (2018), p. 640-653.

[130] Loris Roveda, Jeyhoon Maskani, Paolo Franceschi, Arash Abdi, Francesco Braghin, Lorenzo Molinari Tosatti et Nicola Pedrocchi, « Model-based reinforcement learning variable impedance control for human-robot collaboration », in : *Journal of Intelligent & Robotic Systems* 100.*2* (2020), p. 417-433.

[131] Samuele Vinanzi, Angelo Cangelosi et Christian Goerick, « The role of social cues for goal disambiguation in human-robot cooperation », in : *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2020, p. 971-977.

[132] Luka Peternel, Cheng Fang, Nikos Tsagarakis et Arash Ajoudani, « A selective muscle fatigue management approach to ergonomic human-robot co-manipulation », in : *Robotics and Computer-Integrated Manufacturing* 58 (2019), p. 69-79.

[133] Xiongjun Chen, Ning Wang, Hong Cheng et Chenguang Yang, « Neural learning enhanced variable admittance control for human–robot collaboration », in : *Ieee Access* 8 (2020), p. 25727-25737.

[134] Loris Roveda, Shaghayegh Haghshenas, Marco Caimmi, Nicola Pedrocchi et Lorenzo Molinari Tosatti, « Assisting operators in heavy industrial tasks : On the design of an optimized cooperative impedance fuzzy-controller with embedded safety rules », in : *Frontiers in Robotics and AI* 6 (2019), p. 75.

[135] Marta Lorenzini, Wansoo Kim, Elena De Momi et Arash Ajoudani, « A synergistic approach to the real-time estimation of the feet ground reaction forces and centers of pressure in humans with application to human–robot collaboration », in : *IEEE Robotics and Automation Letters* 3.*4* (2018), p. 3654-3661.

[136] Shaolong Kuang, Yucun Tang, Andi Lin, Shumei Yu et Lining Sun, « Intelligent control for human-robot cooperation in orthopedics surgery », in : *Intelligent Orthopaedics : Artificial intelligence and smart image-guided technology for orthopaedics* (2018), p. 245-262.

[137] Jianjing Zhang, Hongyi Liu, Qing Chang, Lihui Wang et Robert X Gao, « Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly », in : *CIRP annals* 69.*1* (2020), p. 9-12.

[138] Shingo Murata, Yuxi Li, Hiroaki Arie, Tetsuya Ogata et Shigeki Sugano, « Learning to achieve different levels of adaptability for human–robot collaboration utilizing a neuro-dynamical system », in : *IEEE Transactions on Cognitive and Developmental Systems* 10.*3* (2018), p. 712-725.

[139] Liang Yan, Xiaoshan Gao, Xiongjie Zhang et Suokui Chang, « Human-robot collaboration by intention recognition using deep LSTM neural network », in : *2019 IEEE 8th International Conference on Fluid Power and Mechatronics (FPM)*, IEEE, 2019, p. 1390-1396.

[140] Mirza Awais Ahmad, Mouloud Ourak, Caspar Gruijthuijsen, Jan Deprest, Tom Vercauteren et Emmanuel Vander Poorten, « Deep learning-based monocular placental pose estimation : towards collaborative robotics in fetoscopy », in : *International Journal of Computer Assisted Radiology and Surgery* 15 (2020), p. 1561-1571.

[141] Xiongjun Chen, Yiming Jiang et Chenguang Yang, « Stiffness estimation and intention detection for human-robot collaboration », in : *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, 2020, p. 1802-1807.

[142] Ana Cunha, Flora Ferreira, Emanuel Sousa, Luis Louro, Paulo Vicente, Sergio Monteiro, Wolfram Erlhagen et Estela Bicho, « Towards Collaborative Robots as Intelligent Co-workers in Human-Robot Joint Tasks : what to do and who does it ? », in : *ISR 2020 ; 52th International Symposium on Robotics*, VDE, 2020, p. 1-8.

[143] Weronika Wojtak, Flora Ferreira, Paulo Vicente, Luís Louro, Estela Bicho et Wolfram Erlhagen, « A neural integrator model for planning and value-based decision making of a robotics assistant », in : *Neural Computing and Applications* 33 (2021), p. 3737-3756, doi : 10.1007/s00521-020-05224-8.

[144] Dustin Arendt, Caner Komurlu et Leslie M Blaha, « CHISSL : A human-machine collaboration space for unsupervised learning », in : *International Conference on Augmented Cognition*, Springer, 2017, p. 429-448.

[145] Jerry Alan Fails et Dan R Olsen Jr, « Interactive machine learning », in : *Proceedings of the 8th international conference on Intelligent user interfaces*, 2003, p. 39-45.

[146] Liu Jiang, Shixia Liu et Changjian Chen, « Recent research advances on interactive machine learning », in : *Journal of Visualization* 22.*2* (2019), p. 401-417.

[147] Andreas Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, Gloria Cerasela Crişan, C. Pintea et V. Palade, « Interactive machine learning : experimental evidence for the human in the algorithmic loop », in : *Applied Intelligence* 49 (2018), p. 2401-2414.

[148] Nadia Boukhelifa, Anastasia Bezerianos et Evelyne Lutton, « Evaluation of interactive machine learning systems », in : *Human and Machine Learning*, Springer, 2018, p. 341-360.

[149] Shixia Liu, Xiting Wang, Mengchen Liu et Jun Zhu, « Towards better analysis of machine learning models : A visual analytics perspective », in : *Visual Informatics* 1.*1* (2017), p. 48-56.

[150] Jules Françoise, *An overview of Interactive Machine Learning*, 2020.

[151] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy et Klaus Mueller, « Explainable active learning (xal) toward ai explanations as interfaces for machine teachers », in : *Proceedings of the ACM on Human-Computer Interaction* 4.*CSCW3* (2021), p. 1-28.

[152] Dana Angluin, « Queries and concept learning », in : *Machine learning* 2.*4* (1988), p. 319-342.

[153] Andreas Holzinger, « Interactive machine learning for health informatics : when do we need the human-in-the-loop ? », in : *Brain Informatics* 3.*2* (2016), p. 119-131.

[154] Mu-Huan Chung, Mark Chignell, Lu Wang, Alexandra Jovicic et Abhay Raman, « Interactive Machine Learning for Data Exfiltration Detection : Active Learning with Human Expertise », in : *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2020, p. 280-287.

[155] James Fogarty, Desney Tan, Ashish Kapoor et Simon Winder, « CueFlik : interactive concept learning in image search », in : *Proceedings of the sigchi conference on human factors in computing systems*, 2008, p. 29-38.

[156] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein et J Doug Tygar, « Adversarial machine learning », in : *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 2011, p. 43-58.

[157] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik et Ananthram Swami, « Practical black-box attacks against machine learning », in : *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, p. 506-519.

[158] Nicholas Carlini et David Wagner, « Towards evaluating the robustness of neural networks », in : *2017 ieee symposium on security and privacy (sp)*, IEEE, 2017, p. 39-57.

[159] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow et Rob Fergus, « Intriguing properties of neural networks », in : *arXiv preprint arXiv :1312.6199* (2013).

[160] Wenlong Sun, Olfa Nasraoui et Patrick Shafto, « Iterated Algorithmic Bias in the Interactive Machine Learning Process of Information Filtering. », in : *KDIR*, 2018, p. 108-116.

[161] Wieland Brendel, Jonas Rauber et Matthias Bethge, « Decision-based adversarial attacks : Reliable attacks against black-box machine learning models », in : *arXiv preprint arXiv :1712.04248* (2017).

[162] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson et Kilian Weinberger, « Simple black-box adversarial attacks », in : *International Conference on Machine Learning*, PMLR, 2019, p. 2484-2493.

[163] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu et Cho-Jui Hsieh, « Sign-opt : A query-efficient hard-label adversarial attack », in : *arXiv preprint arXiv :1909.10773* (2019).

[164] Chuan Guo, Jared S Frank et Kilian Q Weinberger, « Low frequency adversarial perturbation », in : *arXiv preprint arXiv :1809.08758* (2018).

[165] Reid Porter, James Theiler et Don Hush, « Interactive machine learning in data exploitation », in : *Computing in Science & Engineering* 15.*5* (2013), p. 12-20.

[166] Yuxin Ma, Tiankai Xie, Jundong Li et Ross Maciejewski, « Explaining vulnerabilities to adversarial machine learning through visual analytics », in : *IEEE transactions on visualization and computer graphics* 26.*1* (2019), p. 1075-1085.

[167] Nilaksh Das, Haekyu Park, Zijie J Wang, Fred Hohman, Robert Firstman, Emily Rogers et Duen Horng Chau, « Massif : Interactive interpretation of adversarial attacks on deep learning », in : *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, p. 1-7.

[168] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh et Himabindu Lakkaraju, « Fooling lime and shap : Adversarial attacks on post hoc explanation methods », in : *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, p. 180-186.

[169] Ehsan Emamjomeh-Zadeh et David Kempe, « A general framework for robust interactive learning », in : *arXiv preprint arXiv :1710.05422* (2017).

[170] Stefano Teso et Kristian Kersting, « Explanatory interactive machine learning », in : *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, p. 239-245.

[171] R. S. Gutzwiller et J. Reeder, « Human interactive machine learning for trust in teams of autonomous robots », in : 2017, p. 1-3, doi : 10.1109/COGSIMA.2017.7929607.

[172] Josua Krause, Adam Perer et Kenney Ng, « Interacting with predictions : Visual inspection of black-box machine learning models », in : *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, p. 5686-5697.

[173] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin, « "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier », in : *CoRR* abs/1602.04938 (2016), arXiv : 1602.04938, url : http://arxiv.org/abs/1602.04938.

[174] M. Mozina, « Arguments in Interactive Machine Learning », in : *Informatica (Slovenia)* 42 (2018).

[175] Matteo Turchetta, Felix Berkenkamp et Andreas Krause, « Safe exploration for interactive machine learning », in : *arXiv preprint arXiv :1910.13726* (2019).

[176] Felix Berkenkamp, Andreas Krause et Angela P Schoellig, « Bayesian optimization with safety constraints : safe and automatic parameter tuning in robotics », in : *arXiv preprint arXiv :1602.04450* (2016).

[177] Yanan Sui, Joel Burdick et Yisong Yue, « Stagewise safe bayesian optimization with gaussian processes », in : *International Conference on Machine Learning*, PMLR, 2018, p. 4781-4789.

[178] Stef Van Den Elzen et Jarke J Van Wijk, « Baobabview : Interactive construction and analysis of decision trees », in : *2011 IEEE conference on visual analytics science and technology (VAST)*, IEEE, 2011, 151-1inproceedings60.

[179] Shixia Liu, Jiannan Xiao, Junlin Liu, Xiting Wang, Jing Wu et Jun Zhu, « Visual diagnosis of tree boosting methods », in : *IEEE transactions on visualization and computer graphics* 24.*1* (2017), p. 163-173.

[180] Xun Zhao, Yanhong Wu, Dik Lun Lee et Weiwei Cui, « iForest : Interpreting random forests via visual analytics », in : *IEEE transactions on visualization and computer graphics* 25.*1* (2018), p. 407-416.

[181] T. Mühlbacher, H. Piringer, S. Gratzl, M. Sedlmair et M. Streit, « Opening the Black Box : Strategies for Increased User Involvement in Existing Algorithm Implementations », in : *IEEE Transactions on Visualization and Computer Graphics* 20.*12* (2014), p. 1643-1652, doi : 10.1109/TVCG.2014.2346578.

[182] Vamshi Ambati, « Active Learning and Crowdsourcing for Machine Translation in Low Resource Scenarios », AAI3528171, thèse de doct., USA, 2012, isbn : 9781267582157.

[183] Spencer Frazier et Mark Riedl, « Improving deep reinforcement learning in minecraft with action advice », in : *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2019, p. 146-152.

[184] Davy Preuveneers, Ilias Tsingenopoulos et Wouter Joosen, « Resource usage and performance trade-offs for machine learning models in smart environments », in : *Sensors* 20.*4* (2020), p. 1176.

[185] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M Pintea et Vasile Palade, « A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop », in : *arXiv preprint arXiv :1708.01104* (2017).

[186] Saleema Amershi, James Fogarty, Ashish Kapoor et Desney Tan, « Effective end-user interaction with machine learning », in : *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.

[187] Agnes Tegen, Paul Davidsson et Jan A Persson, « Activity recognition through interactive machine learning in a dynamic sensor setting », in : *Personal and Ubiquitous Computing* (2020), p. 1-14.

[188] Saleema Amershi, James Fogarty et Daniel Weld, « Regroup : Interactive machine learning for on-demand group creation in social networks », in : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, p. 21-30.

[189] Vladimir Dzyuba, Matthijs van Leeuwen, Siegfried Nijssen et Luc De Raedt, « Interactive learning of pattern rankings », in : *International Journal on Artificial Intelligence Tools* 23.*06* (2014), p. 1460026.

[190] Huang Li, Shiaofen Fang, Snehasis Mukhopadhyay, Andrew J Saykin et Li Shen, « Interactive machine learning by visualization : A small data solution », in : *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, p. 3513-3521.

[191] Vishal Jain, William Fedus, Hugo Larochelle, Doina Precup et Marc G Bellemare, « Algorithmic improvements for deep reinforcement learning applied to interactive fiction », in : *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, p. 4328-4336.

[192] Rebecca Fiebrink et Perry R Cook, « The Wekinator : a system for real-time, interactive machine learning in music », in : *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, t. 3, 2010.

[193] Nicholas Gillian et Joseph A Paradiso, « The gesture recognition toolkit », in : *The Journal of Machine Learning Research* 15.*1* (2014), p. 3483-3487.

[194] Margaret Schedel, Phoenix Perry et Rebecca Fiebrink, « Wekinating 000000Swan : Using Machine Learning to Create and Control Complex Artistic Systems. », in : *NIME*, Citeseer, 2011, p. 453-456.

[195] Carlos Gonzalez Diaz, Phoenix Perry et Rebecca Fiebrink, « Interactive machine learning for more expressive game interactions », in : *2019 IEEE Conference on Games (CoG)*, IEEE, 2019, p. 1-2.

[196] Dustin Arendt, Emily Grace et Svitlana Volkova, « Interactive machine learning at scale with CHISSL », in : *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[197] Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau et Thomas A Trikalinos, « Deploying an interactive machine learning system in an evidence-based practice center : abstrackr », in : *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, 2012, p. 819-824.

[199] Ilker Kose, Mehmet Gokturk et Kemal Kilic, « An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance », in : *Applied Soft Computing* 36 (2015), p. 283-299.

[200] P Qian, Y Zhou et C Rudin, « Using Gaussian processes to monitor diabetes development », in : *Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2011) P. Qian, Y. Zhou, C. Rudin, eds*, Citeseer, 2011.

[201] Ankita Tyagi, Ritika Mehra et Aditya Saxena, « Interactive thyroid disease prediction system using machine learning technique », in : *Proceedings on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, 2018, p. 689-693.

[202] Andre Dantas de Medeiros, Nayara Pereira Capobiango, Jose Maria da Silva, Laercio Junio da Silva, Clissia Barboza da Silva et Denise Cunha Fernandes dos Santos Dias, « Interactive machine learning for soybean seed and seedling quality classification », in : *Scientific reports* 10.*1* (2020), p. 1-10.

[203] Simon Flutura, Andreas Seiderer, Tobias Huber, Katharina Weitz, Ilhan Aslan, Ruben Schlagowski, Elisabeth Andre et Joachim Rathmann, « Interactive Machine Learning and Explainability in Mobile Classification of Forest-Aesthetics », in : *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, 2020.

[206] Marco Gillies, « Understanding the role of interactive machine learning in movement interaction design », in : *ACM Transactions on Computer-Human Interaction (TOCHI)* 26.*1* (2019), p. 1-34.

[207] Maria Kyrarini, Muhammad Abdul Haseeb, Danijela Ristic-Durrant et Axel Graser, « Robot learning of industrial assembly task via human demonstrations », in : *Autonomous Robots* 43.*1* (2019), p. 239-257.

[208] Andrew Ilyas, Logan Engstrom, Anish Athalye et Jessy Lin, « Black-box adversarial attacks with limited queries and information », in : *International Conference on Machine Learning*, PMLR, 2018, p. 2137-2146.

[209] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su et Jun Zhu, « Improving black-box adversarial attacks with a transfer-based prior », in : *arXiv preprint arXiv :1906.06919* (2019).

[210] Ian H Witten et Eibe Frank, « Data mining : practical machine learning tools and techniques with Java implementations », in : *Acm Sigmod Record* 31.*1* (2002), p. 76-77.

[211] James D Hollan, Edwin L Hutchins et Louis Weitzman, « STEAMER : An interactive inspectable simulation-based training system », in : *AI magazine* 5.*2* (1984), p. 15-15.

[212] Chenliang Zhou, Dominic Kuang, Jingru Liu, Hanbo Yang, Zijia Zhang, Alan Mackworth et David Poole, « AISpace2 : an interactive visualization tool for learning and teaching artificial intelligence », in : *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, p. 13436-13443.

[213] Justin Talbot, Bongshin Lee, Ashish Kapoor et Desney S Tan, « EnsembleMatrix : interactive visualization to support machine learning with multiple classifiers », in : *Proceedings of the SIGCHI conference on human factors in computing systems*, 2009, p. 1283-1292.

[214] Simon Tong et Edward Chang, « Support vector machine active learning for image retrieval », in : *Proceedings of the ninth ACM international conference on Multimedia*, 2001, p. 107-118.

[215] Ching-Chih Tsai, You-Zhu Chen et Ching-Wen Liao, « Interactive emotion recognition using support vector machine for human-robot interaction », in : *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, IEEE, 2009, p. 407-412.

[216] Malihe Javidi, Baharak Shakeri Aski, Hale Homaei et Hamid Reza Pourreza, « A new approach for interactive image retrieval based on fuzzy feedback and support vector machine », in : *Proceedings on Computational Intelligence for Modelling Control & Automation*, IEEE, 2008, p. 1205-1210.

[217] Hee-Su Kim et Sung-Bae Cho, « Application of interactive genetic algorithm to fashion design », in : *Engineering applications of artificial intelligence* 13.*6* (2000), p. 635-644.

[218] Chih-Chin Lai et Ying-Chuan Chen, « A user-oriented image retrieval system based on interactive genetic algorithm », in : *IEEE transactions on instrumentation and measurement* 60.*10* (2011), p. 3318-3325.

[219] Sung-Bae Cho et Joo-Young Lee, « A human-oriented image retrieval system using interactive genetic algorithm », in : *IEEE Transactions on Systems, Man, and Cybernetics-Part A : Systems and Humans* 32.*3* (2002), p. 452-458.

[220] Shang-Fei Wang, Xu-Fa Wang et Jia Xue, « An improved interactive genetic algorithm incorporating relevant feedback », in : *2005 International Conference on Machine Learning and Cybernetics*, t. 5, IEEE, 2005, p. 2996-3001.

[221] E Poirson, J-F Petiot et David Blumenthal, « Interactive Genetic Algorithm to Collect User Perceptions. Application to the Design of Stemmed Glasses », in : *Nature-Inspired Methods for Metaheuristics Optimization*, Springer, 2020, p. 35-51.

[222] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M Pintea et Vasile Palade, « Interactive machine learning : experimental evidence for the human in the algorithmic loop », in : *Applied Intelligence* 49.*7* (2019), p. 2401-2414.

[223] Jingwen Meng, Xiaoming You et Sheng Liu, « Heterogeneous ant colony optimization based on adaptive interactive learning and non-zero-sum game », in : *europe PMC* (2021).

[224] Weikai Yang, Xiting Wang, Jie Lu, Wenwen Dou et Shixia Liu, « Interactive steering of hierarchical clustering », in : *IEEE Transactions on Visualization and Computer Graphics* (2020).

[225] Fabricio Aparecido Breve, « Simple Interactive Image Segmentation using Label Propagation through kNN graphs », in : *arXiv preprint arXiv :2002.05708* (2020).

[226] David Ha et Douglas Eck, « A neural representation of sketch drawings », in : *arXiv preprint arXiv :1704.03477* (2017).

[229] Miguel Angel Meza Martinez, Mario Nadj et Alexander Maedche, « Towards an integrative theoretical framework of interactive machine learning systems », in : *ais* (2019).

[230] Stefano Teso et Oliver Hinz, *Challenges in interactive machine learning*, 2020.

[231] Josua Krause, Adam Perer et Enrico Bertini, « INFUSE : interactive feature selection for predictive modeling of high dimensional data », in : *IEEE transactions on visualization and computer graphics* 20.*12* (2014), p. 1614-1623.

[232] Fan Yang, Mengnan Du et Xia Hu, « Evaluating explanation without ground truth in interpretable machine learning », in : *arXiv preprint arXiv :1907.06831* (2019).

[233] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs et Enrico Bertini, « A workflow for visual diagnostics of binary classifiers using instance-level explanations », in : *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE, 2017, p. 162-172.

[234] Michael Felderer et Rudolf Ramler, « Quality Assurance for AI-Based Systems : Overview and Challenges (Introduction to Interactive Session) », in : *International Conference on Software Quality*, Springer, 2021, p. 33-42.

[235] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow et Rob Fergus, « Intriguing properties of neural networks », in : *arXiv preprint arXiv :1312.6199* (2013).

[236] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh et Shin-Ming Cheng, « Autozoom : Autoencoder-based zeroth order optimization method for attacking black-box neural networks », in : *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, p. 742-749.

[237] Florian Tramer, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh et Patrick McDaniel, « Ensemble adversarial training : Attacks and defenses », in : *arXiv preprint arXiv :1705.07204* (2017).

[238] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha et Michael Wellman, « Towards the science of security and privacy in machine learning », in : *arXiv preprint arXiv :1611.03814* (2016).

[240] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndic, Pavel Laskov, Giorgio Giacinto et Fabio Roli, « Evasion attacks against machine learning at test time », in : *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 2013, p. 387-402.

[241] Hyun Kwon, Yongchul Kim, Ki-Woong Park, Hyunsoo Yoon et Daeseon Choi, « Friend-safe evasion attack : An adversarial example that is correctly recognized by a friendly classifier », in : *computers & security* 78 (2018), p. 380-397.

[242] Hyun Kwon, Yongchul Kim, Ki-Woong Park, Hyunsoo Yoon et Daeseon Choi, « Multi-targeted adversarial example in evasion attack on deep neural network », in : *IEEE Access* 6 (2018), p. 46084-46096.

[243] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu et Dawn Song, « Targeted backdoor attacks on deep learning systems using data poisoning », in : *arXiv preprint arXiv :1712.05526* (2017).

[244] Jiale Zhang, Junjun Chen, Di Wu, Bing Chen et Shui Yu, « Poisoning attack in federated learning using generative adversarial nets », in : *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, IEEE, 2019, p. 374-380.

[245] Jonas Geiping, Liam Fowl, Gowthami Somepalli, Micah Goldblum, Michael Moeller et Tom Goldstein, « What Doesn't Kill You Makes You Robust (er) : Adversarial Training against Poisons and Backdoors », in : *arXiv preprint arXiv :2102.13624* (2021).

[246] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi et Pascal Frossard, « Deepfool : a simple and accurate method to fool deep neural networks », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 2574-2582.

[247] Florian Tramer, Fan Zhang, Ari Juels, Michael K Reiter et Thomas Ristenpart, « Stealing Machine Learning Models via Prediction {APIs} », in : *25th USENIX security symposium (USENIX Security 16)*, 2016, p. 601-618.

[248] Nicolas Papernot, Patrick McDaniel et Ian Goodfellow, « Transferability in machine learning : from phenomena to black-box attacks using adversarial samples », in : *arXiv preprint arXiv :1605.07277* (2016).

[249] Ishai Rosenberg, Asaf Shabtai, Lior Rokach et Yuval Elovici, « Generic black-box end-to-end attack against rnns and other api calls based malware classifiers », in : *arXiv preprint arXiv :1707.05970* 282 (2017).

[250] Jiakai Wang, « Adversarial examples in physical world », in : *Proc. Workshop Track Int. Conf. Learn. Represent.(ICLR)*, 2021.

[251] John X Morris, Eli Lifland, Jin Yong Yoo et Yanjun Qi, « Textattack : A framework for adversarial attacks in natural language processing », in : *Proceedings of the 2020 EMNLP, Arvix* (2020).

[252] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Li Chen, Michael E Kounavis et Duen Horng Chau, « Adagio : Interactive experimentation with adversarial attack and defense for audio », in : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2018, p. 677-681.

[253] Cynthia Rudin, « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », in : *Nature Machine Intelligence* 1.*5* (2019), p. 206-215.

[255] Nick Littlestone, « Learning quickly when irrelevant attributes abound : A new linear-threshold algorithm », in : *Machine learning* 2.*4* (1988), p. 285-318.

[256] Thorsten Joachims, « Optimizing search engines using clickthrough data », in : *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, p. 133-142.

[257] Filip Radlinski et Thorsten Joachims, « Query chains : learning to rank from implicit feedback », in : *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, p. 239-248.

[258] Pranjal Awasthi, Maria Balcan et Konstantin Voevodski, « Local algorithms for interactive clustering », in : *International Conference on Machine Learning*, PMLR, 2014, p. 550-558.

[259] Ehsan Emamjomeh-Zadeh et David Kempe, « Adaptive hierarchical clustering using ordinal queries », in : *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 2018, p. 415-429.

[260] Maria-Florina Balcan et Avrim Blum, « Clustering with interactive feedback », in : *International Conference on Algorithmic Learning Theory*, Springer, 2008, p. 316-328.

[261] Nilaksh Das, Siwei Li, Chanil Jeon, Jinho Jung, Shang-Tse Chen, Carter Yagemann, Evan Downing, Haekyu Park, Evan Yang et Li Chen, « MLsploit : A Framework for Interactive Experimentation with Adversarial Machine Learning Research », in : *Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ACM, 2019.

[262] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya et Aad Van Moorsel, « The relationship between trust in AI and trustworthy machine learning technologies », in : *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, p. 272-283.

[263] Ming Yin, Jennifer Wortman Vaughan et Hanna Wallach, « Understanding the effect of accuracy on trust in machine learning models », in : *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, p. 1-12.

[264] Donald Honeycutt, Mahsan Nourani et Eric Ragan, « Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy », in : *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2020, p. 63-72.

[266] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier et Markus Rudy, « Ilastik : interactive machine learning for (bio) image analysis », in : *Nature Methods* 16.*12* (2019), p. 1226-1232.

[267] Daniel Ortiz-Martinez, Ismael Garcia-Varea et Francisco Casacuberta, « Online learning for interactive statistical machine translation », in : *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, p. 546-554.

[268] Jesus Gonzalez-Rubio, Daniel Ortiz-Martinez et Francisco Casacuberta, « Active learning for interactive machine translation », in : *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, p. 245-254.

[269] Alvaro Peris, Miguel Domingo et Francisco Casacuberta, « Interactive neural machine translation », in : *Computer Speech & Language* 45 (2017), p. 201-220.

[270] Tsz Kin Lam, Julia Kreutzer et Stefan Riezler, « A reinforcement learning approach to interactive-predictive neural machine translation », in : *arXiv preprint arXiv :1805.01553* (2018).

[271] George Foster, Pierre Isabelle et Pierre Plamondon, « Target-text mediated interactive machine translation », in : *Machine Translation* 12.*1* (1997), p. 175-194.

[272] Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang et Shuming Shi, « Transmart : A practical interactive machine translation system », in : *arXiv preprint arXiv :2105.13072* (2021).

[273] Joseph Malloch, Carla F Griggio, Joanna McGrenere et Wendy E Mackay, « Fieldward and pathward : Dynamic guides for defining your own gestures », in : *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, p. 4266-4277.

[276] Laurent Itti, Christof Koch et Ernst Niebur, « A model of saliency-based visual attention for rapid scene analysis », in : *IEEE Transactions on pattern analysis and machine intelligence* 20.*11* (1998), p. 1254-1259.

[277] Jonathan Harel, Christof Koch et Pietro Perona, « Graph-based visual saliency », in : *Advances in neural information processing systems* 19 (2006).

[278] Olivier Le Meur, Patrick Le Callet, Dominique Barba et Dominique Thoreau, « A coherent computational approach to model bottom-up visual attention », in : *IEEE transactions on pattern analysis and machine intelligence* 28.*5* (2006), p. 802-817.

[279] Neil Bruce et John Tsotsos, « Saliency based on information maximization », in : *Advances in neural information processing systems* 18 (2005).

[280] Tilke Judd, Krista Ehinger, Fredo Durand et Antonio Torralba, « Learning to predict where humans look », in : *Proceedings of the 12th IEEE Conference on computer vision*, IEEE, 2009, p. 2106-2113.

[281] Wenguan Wang, Jianbing Shen, Yizhou Yu et Kwan-Liu Ma, « Stereoscopic thumbnail creation via efficient stereo saliency detection », in : *IEEE transactions on visualization and computer graphics* 23.*8* (2016), p. 2014-2027.

[282] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan et Ming-Hsuan Yang, « Saliency detection via graph-based manifold ranking », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, p. 3166-3173.

[283] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng et Shipeng Li, « Salient object detection : A discriminative regional feature integration approach », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, p. 2083-2090.

[284] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu et Philip HS Torr, « Deeply supervised salient object detection with short connections », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 3203-3212.

[285] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang et Huchuan Lu, « A stagewise refinement model for detecting salient objects in images », in : *Proceedings of the IEEE international conference on computer vision*, 2017, p. 4019-4028.

[286] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang et Xiang Ruan, « Amulet : Aggregating multi-level convolutional features for salient object detection », in : *Proceedings of the IEEE international conference on computer vision*, 2017, p. 202-211.

[287] Eleonora Vig, Michael Dorr et David Cox, « Large-scale optimization of hierarchical features for saliency prediction in natural images », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, p. 2798-2805.

[288] Srinivas SS Kruthiventi, Kumar Ayush et R Venkatesh Babu, « Deepfix : A fully convolutional neural network for predicting human eye fixations », in : *IEEE Transactions on Image Processing* 26.*9* (2017), p. 4446-4456.

[289] Xun Huang, Chengyao Shen, Xavier Boix et Qi Zhao, « Salicon : Reducing the semantic gap in saliency prediction by adapting deep neural networks », in : *Proceedings of the IEEE international conference on computer vision*, 2015, p. 262-270.

[290] Nian Liu, Junwei Han, Tianming Liu et Xuelong Li, « Learning to predict eye fixations via multiresolution convolutional neural networks », in : *IEEE transactions on neural networks and learning systems* 29.*2* (2016), p. 392-404.

[291] Junting Pan, Elisa Sayrol, Xavier Giro-i-Nieto, Kevin McGuinness et Noel E O'Connor, « Shallow and deep convolutional networks for saliency prediction », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 598-606.

[292] Wenguan Wang et Jianbing Shen, « Deep visual attention prediction », in : *IEEE Transactions on Image Processing* 27.*5* (2017), p. 2368-2378.

[293] Vijay Mahadevan et Nuno Vasconcelos, « Spatiotemporal saliency in dynamic scenes », in : *IEEE transactions on pattern analysis and machine intelligence* 32.*1* (2009), p. 171-177.

[294] Dashan Gao, Vijay Mahadevan et Nuno Vasconcelos, « The discriminant center-surround hypothesis for bottom-up saliency », in : *Advances in neural information processing systems* 20 (2007).

[295] Chenlei Guo et Liming Zhang, « A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression », in : *IEEE transactions on image processing* 19.*1* (2009), p. 185-198.

[296] Dmitry Rudoy, Dan B Goldman, Eli Shechtman et Lihi Zelnik-Manor, « Learning video saliency from human gaze using candidate selection », in : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, p. 1147-1154.

[297] Hae Jong Seo et Peyman Milanfar, « Static and space-time visual saliency detection by self-resemblance », in : *Journal of vision* 9.*12* (2009), p. 15-15.

[298] Xiaodi Hou et Liqing Zhang, « Dynamic visual attention : Searching for coding length increments », in : *Advances in neural information processing systems* 21 (2008).

[299] Yuming Fang, Zhou Wang, Weisi Lin et Zhijun Fang, « Video saliency incorporating spatiotemporal cues and uncertainty weighting », in : *IEEE transactions on image processing* 23.*9* (2014), p. 3910-3921.

[300] Sayed Hossein Khatoonabadi, Nuno Vasconcelos, Ivan V Bajic et Yufeng Shan, « How many bits does it take for a stimulus to be salient ? », in : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, p. 5501-5510.

[301] Victor Leboran, Anton Garcia-Diaz, Xose R Fdez-Vidal et Xose M Pardo, « Dynamic whitening saliency », in : *IEEE transactions on pattern analysis and machine intelligence* 39.*5* (2016), p. 893-907.

[302] Roland Mech et Michael Wollborn, « A noise robust method for segmentation of moving objects in video sequences », in : *1997 IEEE International conference on acoustics, speech, and signal processing*, t. 4, IEEE, 1997, p. 2657-2660.

[303] Du-Ming Tsai et Shia-Chih Lai, « Independent component analysis-based background subtraction for indoor surveillance », in : *IEEE Transactions on image processing* 18.*1* (2008), p. 158-167.

[304] Berthold KP Horn et Brian G Schunck, « Determining optical flow », in : *Artificial intelligence* 17.*1-3* (1981), p. 185-203.

[305] Cagdas Bak, Aysun Kocak, Erkut Erdem et Aykut Erdem, « Spatio-temporal saliency networks for dynamic saliency prediction », in : *IEEE Transactions on Multimedia* 20.*7* (2017), p. 1688-1698.

[306] Ting Zhao et Xiangqian Wu, « Pyramid feature attention network for saliency detection », in : *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, p. 3085-3094.

[307] Lai Jiang, Mai Xu et Zulin Wang, « Predicting video saliency with object-to-motion CNN and two-layer convolutional LSTM », in : *arXiv preprint arXiv :1709.06316* (2017).

[308] Yi Tang, Wenbin Zou, Zhi Jin et Xia Li, « Multi-scale spatiotemporal ConvLSTM network for video saliency detection », in : *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, p. 362-369.

[309] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen et Kin-Man Lam, « Pyramid dilated deeper convlstm for video salient object detection », in : *Proceedings of the European conference on computer vision (ECCV)*, 2018, p. 715-731.

[310] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng et Jianbing Shen, « Shifting more attention to video salient object detection », in : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 8554-8564.

[311] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang et Liang Lin, « Flow guided recurrent neural encoder for video salient object detection », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 3243-3252.

[312] Souad Chaabouni, Jenny Benois-Pineau et Chokri Ben Amar, « Transfer learning with deep networks for saliency prediction in natural video », in : *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, p. 1604-1608.

[313] Loris Bazzani, Hugo Larochelle et Lorenzo Torresani, « Recurrent mixture density network for spatiotemporal visual attention », in : *arXiv preprint arXiv :1603.08199* (2016).

[314] Volodymyr Mnih, Nicolas Heess et Alex Graves, « Recurrent models of visual attention », in : *Advances in neural information processing systems* 27 (2014).

[315] Siavash Gorji et James J Clark, « Going from image to video saliency : Augmenting image salience with dynamic attentional push », in : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, p. 7501-7511.

[316] Meijun Sun, Ziqi Zhou, Qinghua Hu, Zheng Wang et Jianmin Jiang, « SG-FCN : A motion and memory-based deep learning model for video saliency detection », in : *IEEE transactions on cybernetics* 49.*8* (2018), p. 2900-2911.

[317] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang et Heung-Yeung Shum, « Learning to detect a salient object », in : *IEEE Transactions on Pattern analysis and machine intelligence* 33.*2* (2010), p. 353-367.

[318] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada et Sabine Susstrunk, « Frequency-tuned salient region detection », in : *Proceedings of the IEEE Conference on computer vision and pattern recognition*, IEEE, 2009, p. 1597-1604.

[319] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr et Shi-Min Hu, « Global contrast based salient region detection », in : *IEEE transactions on pattern analysis and machine intelligence* 37.*3* (2014), p. 569-582.

[320] Wenguan Wang, Jianbing Shen et Fatih Porikli, « Saliency-aware geodesic video object segmentation », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, p. 3395-3402.

[321] Wenguan Wang, Jianbing Shen et Ling Shao, « Video salient object detection via fully convolutional networks », in : *IEEE Transactions on Image Processing* 27.*1* (2017), p. 38-49.

[322] Ali Borji, Ming-Ming Cheng, Huaizu Jiang et Jia Li, « Salient object detection : A benchmark », in : *IEEE transactions on image processing* 24.*12* (2015), p. 5706-5722.

[323] Hadi Hadizadeh, Mario J Enriquez et Ivan V Bajic, « Eye-tracking database for a set of standard video sequences », in : *IEEE Transactions on Image Processing* 21.*2* (2011), p. 898-903.

[324] Laurent Itti, « Automatic foveation for video compression using a neurobiological model of visual attention », in : *IEEE transactions on image processing* 13.*10* (2004), p. 1304-1318.

[325] Stefan Mathe et Cristian Sminchisescu, « Actions in the eye : Dynamic gaze datasets and learnt saliency models for visual recognition », in : *IEEE transactions on pattern analysis and machine intelligence* 37.*7* (2014), p. 1408-1424.

[326] Parag K Mital, Tim J Smith, Robin L Hill et John M Henderson, « Clustering of gaze during dynamic scene viewing is predicted by motion », in : *Cognitive computation* 3.*1* (2011), p. 5-24.

[327] Jianwen Xie, Ming-Ming Cheng, Haibin Ling et Ali Borji, « Revisiting Video Saliency Prediction in the Deep Learning Era », in : *IEEE transactions on pattern analysis and machine intelligence* (2020).

[328] Mikel D Rodriguez, Javed Ahmed et Mubarak Shah, « Action mach a spatio-temporal maximum average correlation height filter for action recognition », in : *2008 IEEE conference on computer vision and pattern recognition*, IEEE, 2008, p. 1-8.

[329] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Fredo Durand, Aude Oliva et Antonio Torralba, « Mit saliency benchmark », in : *MIT Press* (2015).

[330] Chris Stauffer et W Eric L Grimson, « Adaptive background mixture models for real-time tracking », in : *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, t. 2, IEEE, 1999, p. 246-252.

[331] Olivier Barnich et Marc Van Droogenbroeck, « ViBe : A universal background subtraction algorithm for video sequences », in : *IEEE Transactions on Image processing* 20.*6* (2010), p. 1709-1724.

[332] Jianfang Dou, Qin Qin et Zimei Tu, « Background subtraction based on circulant matrix », in : *Signal, Image and Video Processing* 11.*3* (2017), p. 407-414.

[333] Rita Cucchiara, Costantino Grana, Massimo Piccardi et Andrea Prati, « Detecting moving objects, ghosts, and shadows in video streams », in : *IEEE transactions on pattern analysis and machine intelligence* 25.*10* (2003), p. 1337-1342.

[334] Thierry Bouwmans, « Background subtraction for visual surveillance : A fuzzy approach », in : *Handbook on soft computing for video surveillance* 5 (2012), p. 103-138.

[335] Sandeep Singh Sengar et Susanta Mukhopadhyay, « Foreground detection via background subtraction and improved three-frame differencing », in : *Arabian Journal for Science and Engineering* 42.*8* (2017), p. 3621-3633.

[336] Yan Zhang, Stephen J Kiselewich, William A Bauson et Riad Hammoud, « Robust moving object detection at distance in the visible spectrum and beyond using a moving camera », in : *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, IEEE, 2006, p. 131-131.

[337] Manjunath Narayana, Allen Hanson et Erik Learned-Miller, « Coherent motion segmentation in moving camera videos using optical flow orientations », in : *Proceedings of the IEEE International Conference on Computer Vision*, 2013, p. 1577-1584.

[338] Peter Ochs, Jitendra Malik et Thomas Brox, « Segmentation of moving objects by long term video analysis », in : *IEEE transactions on pattern analysis and machine intelligence* 36.*6* (2013), p. 1187-1200.

[339] Artem Rozantsev, Vincent Lepetit et Pascal Fua, « Detecting flying objects using a single moving camera », in : *IEEE transactions on pattern analysis and machine intelligence* 39.*5* (2016), p. 879-892.

[340] Rui Liang, Lei Yan, Pengqi Gao, Xu Qian, Zhongjian Zhang et Huabo Sun, « Aviation video moving-target detection with inter-frame difference », in : t. 3, 2010, p. 1494-1497, doi : 10.1109/CISP.2010.5646303.

[341] Taichi Nakashima et Yoshito Yabuta, « Object detection by using interframe difference algorithm », in : *2018 12th France-Japan and 10th Europe-Asia Congress on Mechatronics*, IEEE, 2018, p. 98-102.

[342] Jiale YIN, Lei LIU, He LI et Qiankun LIU, « The infrared moving object detection and security detection related algorithms based on W4 and frame difference », in : *Infrared Physics & Technology* 77 (2016), p. 302-315.

[343] Sandeep Singh SENGAR et Susanta MUKHOPADHYAY, « A novel method for moving object detection based on block based frame differencing », in : *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, IEEE, 2016, p. 467-472.

[344] Chun-Ming TSAI et Zong-Mu YEH, « Intelligent moving objects detection via adaptive frame differencing method », in : *Asian Conference on Intelligent Information and Database Systems*, Springer, 2013, p. 1-11.

[345] Muyun WENG, Guoce HUANG et Xinyu DA, « A new interframe difference algorithm for moving target detection », in : *2010 3rd international congress on image and signal processing*, t. 1, IEEE, 2010, p. 285-289.

[346] Jiangjian XIAO, Hui CHENG, Harpreet SAWHNEY et Feng HAN, « Vehicle detection and tracking in wide field-of-view aerial video », in : *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2010, p. 679-684.

[347] Yuan Hang CHENG et Jing WANG, « A motion image detection method based on the inter-frame difference method », in : *Applied Mechanics and Materials*, t. 490, Trans Tech Publ, 2014, p. 1283-1286.

[348] Mengjuan FEI, Jing LI et Honghai LIU, « Visual tracking based on improved foreground detection and perceptual hashing », in : *Neurocomputing* 152 (2015), p. 413-428.

[349] Aminou HALIDOU, Xinge YOU, Mahamadou HAMIDINE, Roger Atsa ETOUNDI et Laye Hadji DIAKITE, « Fast pedestrian detection based on region of interest and multi-block local binary pattern descriptors », in : *Computers & Electrical Engineering* 40.*8* (2014), p. 375-389.

[350] Antonio FERNANDEZ-CABALLERO, Jose Carlos CASTILLO, Javier MARTINEZ-CANTOS et Rafael MARTINEZ-TOMAS, « Optical flow or image subtraction in human detection from infrared camera on mobile robot », in : *Robotics and Autonomous Systems* 58.*12* (2010), p. 1273-1281.

[351] Gian Luca Foresti, Christian Micheloni et Claudio Piciarelli, « Detecting moving people in video streams », in : *Pattern Recognition Letters* 26.*14* (2005), p. 2232-2243.

[352] Jianfang Dou et Jianxun Li, « Modeling the background and detecting moving objects based on Sift flow », in : *Optik* 125.*1* (2014), p. 435-440.

[353] Jean-Yves Bouguet, « Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm », in : *Intel corporation* 5.*1-10* (2001), p. 4.

[354] Zhenxiong Xu, Danhong Zhang et Lin Du, « Moving object detection based on improved three frame difference and background subtraction », in : *2017 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*, IEEE, 2017, p. 79-82.

[355] Seungwon Lee, Nahyun Kim, Inho Paek, Monson H Hayes et Joonki Paik, « Moving object detection using unstable camera for consumer surveillance systems », in : *2013 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, 2013, p. 145-146.

[356] Yurong Yang, Huajun Gong, Xinhua Wang et Peng Sun, « Aerial target tracking algorithm based on faster R-CNN combined with frame differencing », in : *Aerospace* 4.*2* (2017), p. 32.

[357] Seyed Mehdi Mohtavipour, Mahmoud Saeidi et Abouzar Arabsorkhi, « A multi-stream CNN for deep violence detection in video sequences using handcrafted features », in : *The Visual Computer* 38.*6* (2022), p. 2057-2072.

[358] Mennatullah Siam, Heba Mahgoub, Mohamed Zahran, Senthil Yogamani, Martin Jagersand et Ahmad El-Sallab, « MODNet : Motion and Appearance based Moving Object Detection Network for Autonomous Driving », in : 2018, p. 2859-2864, doi : 10.1109/ITSC.2018.8569744.

[361] Nicholas J Butko, Lingyun Zhang, Garrison W Cottrell et Javier R Movellan, « Visual saliency model for robot cameras », in : *2008 IEEE International Conference on Robotics and Automation*, IEEE, 2008, p. 2398-2403.

[362] Peng Zhang, Tao Zhuo, Hanqiao Huang et Mohan Kankanhalli, « Saliency flow based video segmentation via motion guided contour refinement », in : *Signal Processing* 142 (2018), p. 431-440.

[363] Ken Fukuchi, Kouji Miyazato, Akisato Kimura, Shigeru Takagi et Junji Yamato, « Saliency-based video segmentation with graph cuts and sequentially updated priors », in : *Proceedings of the IEEE Conference on Multimedia and Expo*, IEEE, 2009, p. 638-641.

[364] Yangyu Chen, Weigang Zhang, Shuhui Wang, Liang Li et Qingming Huang, « Saliency-based spatiotemporal attention for video captioning », in : *2018 IEEE fourth international conference on multimedia big data (BigMM)*, IEEE, 2018, p. 1-8.

[365] Huiyun Wang, Youjiang Xu et Yahong Han, « Spotting and aggregating salient regions for video captioning », in : *Proceedings of the 26th ACM international conference on Multimedia*, 2018, p. 1519-1526.

[366] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra et Rita Cucchiara, « Paying more attention to saliency : Image captioning with saliency and context attention », in : *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.*2* (2018), p. 1-21.

[367] Anwesan Pal, Sayan Mondal et Henrik I Christensen, « " Looking at the Right Stuff"-Guided Semantic-Gaze for Autonomous Driving », in : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, p. 11883-11892.

[368] Fahad Lateef, Mohamed Kas et Yassine Ruichek, « Saliency heat-map as visual attention for autonomous driving using generative adversarial network (gan) », in : *IEEE Transactions on Intelligent Transportation Systems* (2021).

[369] Richard Roberts, Duy-Nguyen Ta, Julian Straub, Kyel Ok et Frank Dellaert, « Saliency detection and model-based tracking : a two part vision system for small robot navigation in forested environment », in : *Unmanned Systems Technology XIV*, t. 8387, International Society for Optics et Photonics, 2012, 83870S.

[370] Chin-Kai Chang, Christian Siagian et Laurent Itti, « Mobile robot vision navigation & localization using gist and saliency », in : *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2010, p. 4147-4154.

[371] Tong Yubing, Faouzi Alaya Cheikh, Fahad Fazal Elahi Guraya, Hubert Konik et Alain Trémeau, « A spatiotemporal saliency model for video surveillance », in : *Cognitive Computation* 3.*1* (2011), p. 241-263.

[372] Zhenfeng Shao, Linggang Wang, Zhongyuan Wang, Wan Du et Wenjing Wu, « Saliency-aware convolution neural network for ship detection in surveillance video », in : *IEEE Transactions on Circuits and Systems for Video Technology* 30.*3* (2019), p. 781-794.

[373] Inyong Yun, Cheolkon Jung, Xinran Wang, Alfred O Hero et Joong Kyu Kim, « Part-level convolutional neural networks for pedestrian detection using saliency and boundary box alignment », in : *IEEE Access* 7 (2019), p. 23027-23037.

[374] Jiayu Ji, Ke Xiang et Xuanyin Wang, « SCVS : blind image quality assessment based on spatial correlation and visual saliency », in : *The Visual Computer* (2022), p. 1-16.

[375] Jianping Shi, Qiong Yan, Li Xu et Jiaya Jia, « Hierarchical image saliency detection on extended CSSD », in : *IEEE transactions on pattern analysis and machine intelligence* 38.*4* (2015), p. 717-729.

[376] Yulin Xie et Huchuan Lu, « Visual saliency detection based on Bayesian model », in : *2011 18th IEEE International Conference on Image Processing*, IEEE, 2011, p. 645-648.

[377] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin et Anne Guerin-Dugue, « Modelling spatio-temporal saliency to predict gaze direction for short videos », in : *International journal of computer vision* 82.*3* (2009), p. 231-243.

[378] Manon Bohic et Victoria E Abraira, « Wired for social touch : the sense that binds us to others », in : *Current Opinion in Behavioral Sciences* 43 (2022), p. 207-215.

[380] Leslie Rice, Eric Wong et Zico Kolter, « Overfitting in adversarially robust deep learning », in : *International Conference on Machine Learning*, PMLR, 2020, p. 8093-8104.

[381] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong et Wang-chun Woo, « Convolutional LSTM network : A machine learning approach for precipitation nowcasting », in : *Advances in neural information processing systems* 28 (2015).

[382] Karen Simonyan et Andrew Zisserman, « Very deep convolutional networks for large-scale image recognition », in : *arXiv preprint arXiv :1409.1556* (2014).

[383] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei, « Imagenet : A large-scale hierarchical image database », in : *Proceedings of the IEEE Conference on computer vision and pattern recognition*, Ieee, 2009, p. 248-255.

[384] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao et Zulin Wang, « Deepvs : A deep learning based video saliency prediction approach », in : *Proceedings of the european conference on computer vision (eccv)*, 2018, p. 602-617.

[385] Timothee Jost, Nabil Ouerhani, Roman Von Wartburg, Rene Muri et Heinz Hugli, « Assessing the contribution of color in visual attention », in : *Computer Vision and Image Understanding* 100.*1-2* (2005), p. 107-123.

[386] Benjamin W Tatler, Roland J Baddeley et Iain D Gilchrist, « Visual correlates of fixation selection : Effects of scale and time », in : *Vision research* 45.*5* (2005), p. 643-659.

[387] Robert J Peters, Asha Iyer, Laurent Itti et Christof Koch, « Components of bottom-up gaze allocation in natural images », in : *Vision research* 45.*18* (2005), p. 2397-2416.

[388] Elin H Williams, Fil Cristino et Emily S Cross, « Human body motion captures visual attention and elicits pupillary dilation », in : *Cognition* 193 (2019), p. 104029.

[389] Dwarikanath Mahapatra, Stefan Winkler et Shih-Cheng Yen, « Motion saliency outweighs other low-level features while watching videos », in : *Human Vision and Electronic Imaging XIII*, t. 6806, SPIE, 2008, p. 246-255.

[390] Kamal Sehairi, Fatima Chouireb et Jean Meunier, « Comparative study of motion detection methods for video surveillance systems », in : *Journal of Electronic Imaging* 26.*2* (2017), p. 023025.

[391] Na Zhao, Yingjie Xia, Chao Xu, Xingmin Shi et Yuncai Liu, « APPOS : An adaptive partial occlusion segmentation method for multiple vehicles tracking », in : *Journal of Visual Communication and Image Representation* 37 (2016), p. 25-31.

[392] Omid Motlagh, Danial Nakhaeinia, Sai Hong Tang, Babak Karasfi et Weria Khaksar, « Automatic navigation of mobile robots in unknown environments », in : *Neural Computing and Applications* 24.*7* (2014), p. 1569-1581.

[393] Jorge Garcia, Alfredo Gardel, Ignacio Bravo, Jose Luis Lazaro, Miguel Martinez et David Rodriguez, « Directional people counter based on head tracking », in : *IEEE Transactions on Industrial Electronics* 60.*9* (2012), p. 3991-4000.

[394] Yimeng Zhang, Xiaoming Liu, Ming-Ching Chang, Weina Ge et Tsuhan Chen, « Spatio-temporal phrases for activity recognition », in : Springer, 2012, p. 707-721.

[397] M Ozan Tezcan, Prakash Ishwar et Janusz Konrad, « BSUV-Net 2.0 : Spatio-temporal data augmentations for video-agnostic supervised background subtraction », in : *IEEE Access* 9 (2021), p. 53849-53860.

[398] Diederik P Kingma et Jimmy Ba, « Adam : A method for stochastic optimization », in : *arXiv preprint arXiv :1412.6980* (2014).

[399] Markus Bosch, « Deep learning for robust motion segmentation with non-static cameras », in : *arXiv preprint arXiv :2102.10929* (2021).

[400] Alan M Turing, *Computing machinery and intelligence*, Springer, 2009.

[401] Peter H Kahn Jr, Hiroshi Ishiguro, Batya Friedman, Takayuki Kanda, Nathan G Freier, Rachel L Severson et Jessica Miller, « What is a human ? : Toward psychological benchmarks in the field of human–robot interaction », in : *Interaction Studies* 8.*3* (2007), p. 363-390.

[403] Thomas B Sheridan, « Human–robot interaction : status and challenges », in : *Human factors* 58.*4* (2016), p. 525-532.

[404] Sara Kiesler, Aaron Powers, Susan R Fussell et Cristen Torrey, « Anthropomorphic interactions with a robot and robot–like agent », in : *Social Cognition* 26.*2* (2008), p. 169-181.

[405] Jenay M Beer, Arthur D Fisk et Wendy A Rogers, « Toward a framework for levels of robot autonomy in human-robot interaction », in : *Journal of human-robot interaction* 3.*2* (2014), p. 74.

[406] Rebecca Andreasson, Beatrice Alenljung, Erik Billing et Robert Lowe, « Affective touch in human–robot interaction : conveying emotion to the Nao robot », in : *International Journal of Social Robotics* 10 (2018), p. 473-491.

[407] Vincent Duchaine et Clément Gosselin, « Safe, stable and intuitive control for physical human-robot interaction », in : *Proceedings of the IEEE Conference on Robotics and Automation*, IEEE, 2009, p. 3383-3388.

[409] Bernhard Hommel, Jochen Müsseler, Gisa Aschersleben et Wolfgang Prinz, « The theory of event coding (TEC) : A framework for perception and action planning », in : *Behavioral and brain sciences* 24.*5* (2001), p. 849-878.

[410] Anja Jackowski et Marion Gebhard, « Evaluation of hands-free human-robot interaction using a head gesture based interface », in : *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, p. 141-142.

[411] Hannah Pelikan, Frederic Anthony Robinson, Leelo Keevallik, Mari Velonaki, Mathias Broth et Oliver Bown, « Sound in Human-Robot Interaction », in : *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, p. 706-708.

[412] Chien-Ming Huang, Sean Andrist, Allison Sauppé et Bilge Mutlu, « Using gaze patterns to predict task intent in collaboration », in : *Frontiers in psychology* 6 (2015), p. 1049.

[414] Haofeng Li, Guanqi Chen, Guanbin Li et Yizhou Yu, « Motion guided attention for video salient object detection », in : *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, p. 7274-7283.

[415] Natnael, Maelic Wondimu, Antoine Neau, Ubbo Dizet, Cedric Visser et Buche, « Anthropomorphic Human Robot Interaction Framework" Attention Based Approach », in : *Lecture Notes in Artificial Intelligence (LNAI)* x.*x* (2023), p. x-x.

[416] Natnael, Maelic Wondimu, Antoine Neau, Ubbo Dizet, Cedric Visser et Buche, « Efficient Human-Robot Social Interaction through Video Saliency-Based Heuristics for Heavy Machine Learning Models », in : *Lecture Notes in Artificial Intelligence (LNAI) Studies* x.*x* (2023), p. x-x.

[417] Alin Albu-Schaffer, Sami Haddadin, Ch Ott, Andreas Stemmer, Thomas Wimbock et Gerhard Hirzinger, « The DLR lightweight robot : design and control concepts for robots in human environments », in : *Industrial Robot : an international journal* 34.*5* (2007), p. 376-385.

[418] Lakshmi M Hari, G Venugopal et S Ramakrishnan, « Analysis of Isometric Muscle Contractions using Analytic Bump Continuous Wavelet Transform », in :

2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, p. 732-735.

[419] Michael A Gelbart, Jasper Snoek et Ryan P Adams, « Bayesian optimization with unknown constraints », in : *arXiv preprint arXiv :1403.5607* (2014).

[420] Martin L Puterman, *Markov decision processes : discrete stochastic dynamic programming*, John Wiley & Sons, 2014.

[421] Alonso Marco, Felix Berkenkamp, Philipp Hennig, Angela P Schoellig, Andreas Krause, Stefan Schaal et Sebastian Trimpe, « Virtual vs. real : Trading off simulations and physical experiments in reinforcement learning with Bayesian optimization », in : *2017 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2017, p. 1557-1563.

[422] Igor Mordatch, Kendall Lowrey, Galen Andrew, Zoran Popovic et Emanuel V Todorov, « Interactive control of diverse complex characters with neural networks », in : *Advances in Neural Information Processing Systems* 28 (2015), p. 3132-3140.

[423] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali et Giovanni Felici, « Hacking smart machines with smarter ones : How to extract meaningful data from machine learning classifiers », in : *International Journal of Security and Networks* 10 (2015), p. 137-150.

[424] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang et Wenyuan Xu, « Dolphinattack : Inaudible voice commands », in : *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, p. 103-117.

[425] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner et Wenchao Zhou, « Hidden voice commands », in : *25th USENIX security symposium (USENIX security 16)*, 2016, p. 513-530.

[426] Francisco Bernardo, Michael Zbyszynski, Rebecca Fiebrink et Mick Grierson, « Interactive machine learning for end-user innovation », in : *2017 AAAI Spring Symposium Series*, 2017.

[427] Todd Kulesza, Margaret Burnett, Weng-Keen Wong et Simone Stumpf, « Principles of explanatory debugging to personalize interactive machine learning », in : *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, p. 126-137.

[428] Yuan-Ting Hu, Jia-Bin Huang et Alexander G Schwing, « Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation », in : *Proceedings of the European conference on computer vision (ECCV)*, 2018, p. 786-802.

[429] Ascension Gallardo-Antolin et Juan M Montero, « An Auditory Saliency Pooling-Based LSTM Model for Speech Intelligibility Classification », in : *Symmetry* 13.*9* (2021), p. 1728.

[430] Sourav Kumar Mukhopadhyay et Sridhar Krishnan, « Visual saliency detection approach for long-term ECG analysis », in : *Computer Methods and Programs in Biomedicine* 213 (2022), p. 106518.

[431] L Straetmans, B Holtze, S Debener, M Jaeger et B Mirkovic, « Neural tracking to go : auditory attention decoding and saliency detection with mobile EEG », in : *Journal of Neural Engineering* 18.*6* (2022), p. 066054.

[432] Jiazhong Chen, Qingqing Li, Hefei Ling, Dakai Ren et Ping Duan, « Audiovisual saliency prediction via deep learning », in : *Neurocomputing* 428 (2021), p. 248-258.

[433] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang et Xinping Guan, « A multimodal saliency model for videos with high audio-visual correspondence », in : *IEEE Transactions on Image Processing* 29 (2020), p. 3805-3819.

[434] Manfred Lau, Kapil Dev, Weiqi Shi, Julie Dorsey et Holly Rushmeier, « Tactile mesh saliency », in : *ACM Transactions on Graphics (TOG)* 35.*4* (2016), p. 1-11.

[435] Erik Van der Burg, Christian NL Olivers, Adelbert W Bronkhorst et Jan Theeuwes, « Poke and pop : Tactile–visual synchrony increases visual saliency », in : *Neuroscience letters* 450.*1* (2009), p. 60-64.

[436] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba et Frédo Durand, « What Do Different Evaluation Metrics Tell Us About Saliency Models ? », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.*3* (2019), p. 740-757.

[437] Anton Garcia-Diaz, Xose R Fdez-Vidal, Xose M Pardo et Raquel Dosil, « Saliency from hierarchical adaptation through decorrelation and variance normalization », in : *Image and Vision Computing* 30.*1* (2012), p. 51-64.

[438] Jianming Zhang et Stan Sclaroff, « Saliency detection : A boolean map approach », in : *Proceedings of the IEEE international conference on computer vision*, 2013, p. 153-160.

[439] Jiazhong Chen, Bingpeng Ma, Hua Cao, Jie Chen, Yebin Fan, Tao Xia et Rong Li, « Attention region detection based on closure prior in layered bit planes », in : *Neurocomputing* 251 (2017), p. 16-25.

[440] Sudarshan Ramenahalli, « A Biologically Motivated, Proto-Object-Based Audiovisual Saliency Model », in : *AI* 1.*4* (2020), p. 487-509.

[441] Erkut Erdem et Aykut Erdem, « Visual saliency estimation by nonlinearly integrating features using region covariances », in : *Journal of vision* 13.*4* (2013), p. 11-11.

[442] Ali Borji, « What is a salient object ? A dataset and a baseline model for salient object detection », in : *IEEE Transactions on Image Processing* 24.*2* (2014), p. 742-756.

[443] Antoine Coutrot et Nathalie Guyader, « Toward the introduction of auditory information in dynamic visual attention models », in : *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, IEEE, 2013, p. 1-4.

[444] Konstantinos Rapantzikos, Georgios Evangelopoulos, Petros Maragos et Yannis Avrithis, « An audio-visual saliency model for movie summarization », in : *2007 IEEE 9th Workshop on Multimedia Signal Processing*, IEEE, 2007, p. 320-323.

[445] Francois Luus, Naweed Khan et Ismail Akhalwaya, « Active learning with tensorboard projector », in : *arXiv preprint arXiv :1901.00675* (2019).

[446] Thomas Wisspeintner, Tijn Van Der Zant, Luca Iocchi et Stefan Schiffer, « RoboCup@ Home : Scientific competition and benchmarking for domestic service robots », in : *Interaction Studies* 10.*3* (2009), p. 392-426.

[447] M Andriea, M Barange, M Bouadelli, C Buche, A Dizet, D Duhaut, A Legeleux, M NEAU, A Pauchet et S Rasendrasoa, « RoboBreizh 2022 Team Description Paper », in : *ENIB* (2021).

[448] Marcin Marszalek, Ivan Laptev et Cordelia Schmid, « Actions in context », in : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Conference*, IEEE, 2009, p. 2929-2936.

[449] Francois Chollet, *Keras*, `https://github.com/fchollet/keras`, 2015.

[450] Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec et Vasilis Karaiskos, « The AMI meeting corpus », in : *Proceedings of the 5th international conference on methods and techniques in behavioral research*, t. 88, Citeseer, 2005, p. 100.

[451] A KingaD, « A methodforstochasticoptimization », in : *Anon. International Conference on Learning Representations. SanDego : ICLR* (2015).

[452] Soharab Hossain Shaikh, Khalid Saeed et Nabendu Chaki, « Moving object detection approaches, challenges and object tracking », in : *Moving object detection using background subtraction*, Springer, 2014, p. 5-14.

[453] Nishu Singla, « Motion detection based on frame difference method », in : *International Journal of Information & Computation Technology* 4.*15* (2014), p. 1559-1565.

[454] Jian Li, Zhong-Ming Pan, Zhuo-Hang Zhang et Heng Zhang, « Dynamic ARMA-based background subtraction for moving objects detection », in : *IEEE Access* 7 (2019), p. 128659-128668.

[455] Yoshua Bengio et Yann LeCun, « Scaling Learning Algorithms Towards AI », in : *Large Scale Kernel Machines*, MIT Press, 2007.

[456] Geoffrey E. Hinton, Simon Osindero et Yee Whye Teh, « A Fast Learning Algorithm for Deep Belief Nets », in : *Neural Computation* 18 (2006), p. 1527-1554.

[457] Ian Goodfellow, Yoshua Bengio, Aaron Courville et Yoshua Bengio, *Deep learning*, t. 1, MIT Press, 2016.

[458] Hadi Hadizadeh et Ivan V Bajic, « Saliency-aware video compression », in : *IEEE Transactions on Image Processing* 23.*1* (2013), p. 19-33.

[459] Javid Ullah, Ahmad Khan et Muhammad Arfan Jaffar, « Motion cues and saliency based unconstrained video segmentation », in : *Multimedia Tools and Applications* 77.*6* (2018), p. 7429-7446.

[460] Viraj Mavani, Shanmuganathan Raman et Krishna P Miyapuram, « Facial expression recognition using visual saliency and deep learning », in : *Proceedings of the IEEE international conference on computer vision workshops*, 2017, p. 2783-2788.

[461] Zhuoqing Chang, J Matias Di Martino, Qiang Qiu, Steven Espinosa et Guillermo Sapiro, « Salgaze : Personalizing gaze estimation using visual saliency », in : *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[462] Brian J White et Douglas P Munoz, « The superior colliculus », in : *The Oxford handbook of eye movements*, Oxford University Press, 2011, 195-213).

[463] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra et Rita Cucchiara, « Visual saliency for image captioning in new multimedia services », in : *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, 2017, p. 309-314.

[464] Sina Mohseni, Akshay Jagadeesh et Zhangyang Wang, « Predicting model failure using saliency maps in autonomous driving systems », in : *arXiv preprint arXiv :1905.07679* (2019).

[465] Nan Mu, Xin Xu et Xiaolong Zhang, « Finding autofocus region in low contrast surveillance images using CNN-based saliency algorithm », in : *Pattern Recognition Letters* 125 (2019), p. 124-132.

[466] Ijaz Ul Haq, Amin Ullah, Khan Muhammad, Mi Young Lee et Sung Wook Baik, « Personalized movie summarization using deep cnn-assisted facial expression recognition », in : *Complexity* 2019 (2019).

[467] George Pantazis, George Dimas et Dimitris K Iakovidis, « SalSum : Saliency-based Video Summarization using Generative Adversarial Networks », in : *arXiv preprint arXiv :2011.10432* (2020).

[468] Ripei Zhang, Chunyi Chen, Jiacheng Zhang, Jun Peng et Ahmed Mustafa Taha Alzbier, « 360-degree visual saliency detection based on fast-mapped convolution and adaptive equator-bias perception », in : *The Visual Computer* (2022), p. 1-18.

[469] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian et Vineet Gandhi, « ViNet : Pushing the limits of Visual Modality for Audio-Visual Saliency Prediction », in : *arXiv preprint arXiv :2012.06170* (2020).

[470] Matthias Kummerer, Lucas Theis et Matthias Bethge, « Deep gaze i : Boosting saliency prediction with feature maps trained on imagenet », in : *arXiv preprint arXiv :1411.1045* (2014).

[471] Sudarshan Ramenahalli, « A proto-object based audiovisual saliency map », in : *arXiv preprint arXiv :2003.06779* (2020).

[472] Antigoni Tsiami, Petros Koutras, Athanasios Katsamanis, Argiro Vatakis et Petros Maragos, « A behaviorally inspired fusion approach for computational audiovisual saliency modeling », in : *Signal Processing : Image Communication* 76 (2019), p. 186-200.

[473] Hamed R Tavakoli, Ali Borji, Esa Rahtu et Juho Kannala, « DAVE : A deep audio-visual embedding for dynamic saliency prediction », in : *arXiv preprint arXiv :1905.10693* (2019).

[474] Petros Koutras, Georgia Panagiotaropoulou, Antigoni Tsiami et Petros Maragos, « Audio-visual temporal saliency modeling validated by fmri data », in : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, p. 2000-2010.

[475] Antigoni Tsiami, Athanasias Katsamanis, Petros Maragos et Argiro Vatakis, « Towards a behaviorally-validated computational audiovisual saliency model », in : *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, p. 2847-2851.

[476] Antoine Coutrot et Nathalie Guyader, « An efficient audiovisual saliency model to predict eye positions when looking at conversations », in : *2015 23rd European Signal Processing Conference (EUSIPCO)*, IEEE, 2015, p. 1531-1535.

[477] Jiro Nakajima, Akihiro Sugimoto et Kazuhiko Kawamoto, « Incorporating audio signals into constructing a visual saliency map », in : *Pacific-Rim Symposium on Image and Video Technology*, Springer, 2013, p. 468-480.

[478] Sudarshan Ramenahalli, Daniel R Mendat, Salvador Dura-Bernal, Eugenio Culurciello, Ernst Niebur et Andreas Andreou, « Audio-visual saliency map : Overview, basic models and hardware implementation », in : *2013 47th Annual Conference on Information Sciences and Systems (CISS)*, IEEE, 2013, p. 1-6.

[479] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan et Garrison W Cottrell, « SUN : A Bayesian framework for saliency using natural statistics », in : *Journal of vision* 8.7 (2008), p. 32-32.

[480] Dashan Gao et Nuno Vasconcelos, « Discriminant saliency for visual recognition from cluttered scenes », in : *Advances in neural information processing systems* 17 (2004).

[481] Xiaodi Hou et Liqing Zhang, « Saliency detection : A spectral residual approach », in : *2007 IEEE Conference on computer vision and pattern recognition*, Ieee, 2007, p. 1-8.

[482] Zhuolin Jiang et Larry S Davis, « Submodular salient region detection », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, p. 2043-2050.

[483] Saumya Jetley, Naila Murray et Eleonora Vig, « End-to-end saliency mapping via probability distribution prediction », in : *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 5753-5761.

[484] Wenguan Wang et Jianbing Shen, « Deep cropping via attention box prediction and aesthetics assessment », in : *Proceedings of the IEEE international conference on computer vision*, 2017, p. 2186-2194.

[485] Chenlei Guo, Qi Ma et Liming Zhang, « Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform », in : *2008 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2008, p. 1-8.

[486] Kwok Ching Tsui et Jiming Liu, « Evolutionary diffusion optimization. I. Description of the algorithm », in : *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, t. 1, IEEE, 2002, p. 169-174.

[487] Wenguan Wang, Jianbing Shen, Ling Shao et Fatih Porikli, « Correspondence driven saliency transfer », in : *IEEE Transactions on Image Processing* 25.*11* (2016), p. 5025-5034.

[488] Wenguan Wang, Jianbing Shen, Ruigang Yang et Fatih Porikli, « Saliency-aware video object segmentation », in : *IEEE transactions on pattern analysis and machine intelligence* 40.*1* (2017), p. 20-33.

[489] Ali Borji, Dicky N Sihite et Laurent Itti, « Quantitative analysis of human-model agreement in visual saliency modeling : A comparative study », in : *IEEE Transactions on Image Processing* 22.*1* (2012), p. 55-69.

[490] Yongfang Wang, Peng Ye, Yumeng Xia et Ping An, « A heuristic framework for perceptual saliency prediction », in : *Journal of Visual Communication and Image Representation* 73 (2020), p. 102913.

[491] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa et Padmanabhan Soundararajan, « The CLEAR 2006 evaluation », in : *Multimodal Technologies for Perception of Humans : First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006, Southampton, UK, April 6-7, 2006, Revised Selected Papers 1*, Springer, 2007, p. 1-44.

Sources secondaires

[1] Terrence Fong, Illah Nourbakhsh et Kerstin Dautenhahn, « A survey of socially interactive robots », in : *Robotics and autonomous systems* 42.*3-4* (2003), p. 143-166.

[4] David Vernon, Giorgio Metta et Giulio Sandini, « A survey of artificial cognitive systems : Implications for the autonomous development of mental capabilities in computational agents », in : *IEEE transactions on evolutionary computation* 11.*2* (2007), p. 151-180.

[5] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng et Shi-Min Hu, « Attention mechanisms in computer vision : A survey », in : *Computational Visual Media* 8.*3* (2022), p. 331-368.

[9] Michael I Posner et Steven E Petersen, « The attention system of the human brain », in : *Annual review of neuroscience* 13.*1* (1990), p. 25-42.

[10] Cynthia Breazeal, « Social interactions in HRI : the robot view », in : *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34.*2* (2004), p. 181-186.

[15] Joao Perdiz, Gabriel Pires et Urbano J Nunes, « Emotional state detection based on EMG and EOG biosignals : A short survey », in : *2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG)*, IEEE, 2017, p. 1-4.

[22] Omar Mubin, Catherine J Stevens, Suleman Shahid, Abdullah Al Mahmud et Jian-Jie Dong, « A review of the applicability of robots in education », in : *Journal of Technology in Education and Learning* 1.*209-0015* (2013), p. 13.

[23] Michael A Goodrich et Alan C Schultz, « Human–robot interaction : a survey », in : *Foundations and Trends® in Human–Computer Interaction* 1.*3* (2008), p. 203-275.

[24] Francesco Semeraro, Alexander Griffiths et Angelo Cangelosi, « Human–robot collaboration and machine learning : A systematic review of recent research », in : *Robotics and Computer-Integrated Manufacturing* 79 (2023), p. 102432.

[26] Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schaffer, Kazuhiro Kosuge et Oussama Khatib, « Progress and prospects of the human–robot collaboration », in : *Autonomous Robots* 42 (2018), p. 957-975.

[47] Andrea Bauer, Dirk Wollherr et Martin Buss, « Human–robot collaboration : a survey », in : *International Journal of Humanoid Robotics* 5.*01* (2008), p. 47-66.

[89] Aswin K Ramasubramanian, Syed M Aiman et Nikolaos Papakostas, « On using human activity recognition sensors to improve the performance of collaborative mobile manipulators : Review and outlook », in : *Procedia CIRP* 97 (2021), p. 211-216.

[96] Scott A Green, Mark Billinghurst, XiaoQi Chen et J Geoffrey Chase, « Human-robot collaboration : A literature review and augmented reality approach in design », in : *International journal of advanced robotic systems* 5.*1* (2008), p. 1.

[113] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati et Nicolai Marquardt, « Augmented reality and robotics : A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces », in : *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, p. 1-33.

[198] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi et Uwe Aickelin, « A Review on Human–AI Interaction in Machine Learning and Insights for Medical Applications », in : *International Journal of Environmental Research and Public Health* 18.*4* (2021), p. 2121.

[204] Konstantinos G Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson et Dionysis Bochtis, « Machine learning in agriculture : A review », in : *Sensors* 18.*8* (2018), p. 2674.

[205] Douglas J White, « A survey of applications of Markov decision processes », in : *Journal of the operational research society* 44.*11* (1993), p. 1073-1096.

[227] John J Dudley et Per Ola Kristensson, « A review of user interface design for interactive machine learning », in : *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.*2* (2018), p. 1-37.

[228] Angelos Chatzimparmpas, Rafael M Martins, Ilir Jusufi et Andreas Kerren, « A survey of surveys on the use of visualization for interpreting machine learning models », in : *Information Visualization* 19.*3* (2020), p. 207-233.

[239] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay et Debdeep Mukhopadhyay, « Adversarial attacks and defences : A survey », in : *arXiv preprint arXiv :1810.00069* (2018).

[254] Dana Angluin, « Computational learning theory : survey and selected bibliography », in : *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, 1992, p. 351-369.

[265] Roger C Mayer, James H Davis et F David Schoorman, « An integrative model of organizational trust », in : *Academy of management review* 20.*3* (1995), p. 709-734.

[274] Natnael A Wondimu, Cedric Buche et Ubbo Visser, « Interactive machine learning : A state of the art review », in : *arXiv preprint arXiv :2207.06196* (2022).

[275] Laurent Itti et Christof Koch, « Computational modelling of visual attention », in : *Nature reviews neuroscience* 2.*3* (2001), p. 194-203.

[360] Joao Filipe Ferreira et Jorge Dias, « Attentional mechanisms for socially interactive robots–a survey », in : *IEEE Transactions on Autonomous Mental Development* 6.*2* (2014), p. 110-125.

[379] Dima Amso et Gaia Scerif, « The attentive brain : insights from developmental cognitive neuroscience », in : *Nature Reviews Neuroscience* 16.*10* (2015), p. 606-619.

[395] Steven S. Beauchemin et John L. Barron, « The computation of optical flow », in : *ACM computing surveys (CSUR)* 27.*3* (1995), p. 433-466.

[396] Belmar Garcia-Garcia, Thierry Bouwmans et Alberto Jorge Rosales Silva, « Background subtraction in real applications : Challenges, current models and future directions », in : *Computer Science Review* 35 (2020), p. 100204.

[408] Steven E Petersen et Michael I Posner, « The attention system of the human brain : 20 years after », in : *Annual review of neuroscience* 35 (2012), p. 73-89.

[413] Simone Frintrop, Erich Rome et Henrik I Christensen, « Computational visual attention systems and their cognitive foundations : A survey », in : *ACM Transactions on Applied Perception (TAP)* 7.*1* (2010), p. 1-39.

COLLEGE MATHSTIC
DOCTORAL BRETAGNE
BRETAGNE OCEANE

ÉCOLE NATIONALE D'INGÉNIEURS DE BREST

Title: Application of Interactive Machine Learning Models For Human-Robot Interaction: The Human Attention Approach

Keywords: human-robot interaction, saliency prediction, moving object detection, human attention model, heuristic computer vision, intuitive HRI

Abstract: In this thesis, we develop algorithmic and computational techniques to employ enhanced attention models for intuitive and efficient human-robot social interaction. One of the attention models we use in our thesis is interactive video saliency prediction, which we advance through a research contribution. We develop a stacked-convLSTM based video saliency prediction model that uses the Dynamic Human Fixation (DHF1K) dataset to predict human attention. To evaluate our model, we employ Normalized Scanpath Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), AUC-Judd (AUC-J), and shuffled AUC (s-AUC) in adherence to the nature of saliency prediction. Our experimental results show an outstanding performance of our saliency prediction model against the state-of-the-art models.

In addition to saliency prediction, we discuss our contribution in addressing moving object detection and segmentation machine learning problem. We introduce a robust frame differencing technique called XY-shift frame differencing as a major constituent of our interactive moving object detection and segmentation framework. We use a modified version of the change detection network (CDNet-2014) dataset to train and evaluate our model. Our extensive experimental results based on 4-fold cross validation strategy show an outstanding performance of our model against the state-of-the-art moving object detection and segmentation models.

Furthermore, we introduce a novel attention-based anthropomorphic HRI framework. Our research work mainly investigates the use of attention models as heuristic functions and their impact in enhancing the intuitiveness and efficiency of human-robot interaction. We experiment with the characteristics of our framework in both real and virtual environments using Robot Operating System (ROS) and third-party enabling modules as communication and processing modules, respectively. We also use the Linux-based humanoid robot, Pepper, to experiment with the characteristics of our framework in a physical environment. In addition, we use Gazebo, a robot simulation software, and the model of Pepper humanoid robot to evaluate the characteristics of our model in a virtual environment. Our experimental results show the impact of the proposed framework in enhancing human-robot social interaction.

In conclusion, we have identified several potential research directions that may be of interest to scholars in the fields of HRI, machine learning, and computer vision.