IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 15, 2020, accepted September 22, 2020, date of publication September 28, 2020,
date of current version October 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027152

A Systematic Review on Reinforcement Learning-Based Robotics Within the Last Decade

MD. AL-MASRUR KHAN[1], MD RASHED JAOWAD KHAN[1], ABUL TOOSHIL[1], NILOY SIKDER[2], M. A. PARVEZ MAHMUD[3], ABBAS Z. KOUZANI[3], AND ABDULLAH-AL NAHID[1]

[1]Electronics and Communication Engineering Discipline, Khulna University, Khulna 9208, Bangladesh
[2]Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh
[3]School of Engineering, Deakin University, Geelong, VIC 3216, Australia

Corresponding author: Abdullah-Al Nahid (nahid.ece.ku@gmail.com)

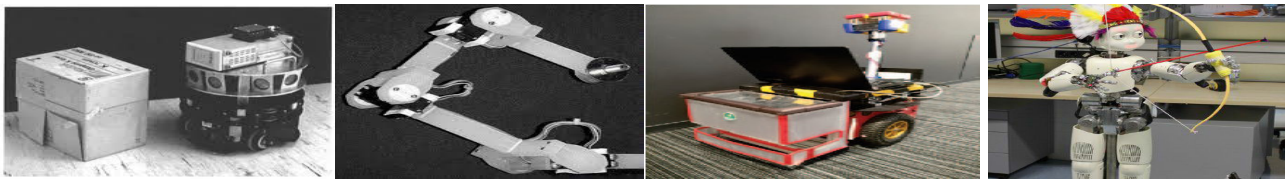
ABSTRACT Robotics is one of the many tools that is making a substantial difference as the world is experiencing the fourth industrial revolution. To ease control over this engineering marvel substantially, Reinforcement Learning (RL) has paved its way in recent years quite remarkably. RL enables robots to become self-aware towards carrying out a specific task followed by user operations. For decades of rigorous endeavor, this research field has gone through numerous groundbreaking developments and it will be the same for the coming days. Therefore, this paper steps in to enlighten the scientific community with a systemic review of the published research papers within the past decade. The bibliographic data that is extracted from the papers are analyzed using an automated tool named Vosviewer with respect to some parameters. Substantial excerpts from the most influential papers are highlighted in this work. Furthermore, this paper points out the global research practice in this field. The paper also generates some intriguing questions and answers them in regards to the research topic. After reading this paper, future researchers will have a firm idea in the RL-based robotics and will be able to incorporate in their own research.

INDEX TERMS Bibliometric analysis, reinforcement learning, robotics, systematic review.

I. INTRODUCTION

The concept and utilization of autonomous systems have eased the daily life activities from a different perspective. The concept of autonomy is first introduced by D.S Harder in 1936 [1]. An autonomous agent who can take decisions in real-time without any human intervention is an elementary component of automation technology. With the development of technology, autonomous agents are working alongside people in various industrial and domestic settings over the past decades. The electricity, behind the automation technology, is Machine Learning (ML) [2] which enables a machine to perceive the world or the agent's working environment as a human does by learning and improving the experiences gained from the circumstance. Different ML techniques have been utilized over the years to make the automation technology more developed. A subset of ML called Deep Learning (DL) [3] has proved itself as a promising approach to the field of automation by integrating features on automation processes like computer vision, image recognition, behaviour learning, etc. to improve the perception tasks of the machines. DL can be classified in three different categories such as; Supervised Learning (SL) [4], Unsupervised Learning (UL) [5], Reinforcement Learning (RL) [6]. In SL a set of labeled training data is used as input to teach the machines. UL looks for an undetected pattern with no labels in a dataset. Here RL is quite different from these two; as RL helps an autonomous machine to adjust the behaviour of it with the new challenges without any kind of training dataset rather by having contact with the environment and by employing the experiences attained from the environment based on trial and error methods. As autonomous agents are replicating humans in different areas and working in a multidimensional field; the complexity of these agents control system is increasing day by day in a way that it is becoming a hard task for the engineers to design the agents for making them adaptive with their

The associate editor coordinating the review of this manuscript and approving it for publication was Yangmin Li.



(a) OBELIX Robot (b) Weightlifter Robot (c) Autonomous Vehicle (d) Humanoid Robot

FIGURE 1. The representation of a small specimen of robots which are based on RL. (a) The OBELIX robot is a value-function based robot which was taught to nudge boxes [7], (b) The weightlifter robot is a policy-search based robot which was developed by Rosenstein *et al.* [13], (c) The autonomous vehicle based on DDQN was able to manoeuvre independently in a human crowded environment [12], (d) The 53-DOF humanoid iCub learned the skill of archery based on ARCHER algorithm [17].

working environment on all possible circumstances. So, RL is a possible solution to overcome all these kinds of barriers. Because of its advantages over the traditional learning algorithms, RL has become very popular among the researchers in recent times. The RL has a wide variety of application fields on autonomous technology. Among the technologies, robotics is a prominent one as robotics has become an indispensable part of automation technology over the last decades. Robotics has undergone a serious renaissance as RL has been started to utilize in this field. Literature shows that, the early works on RL-based robotics dated back to 1992 and 1995 [7], [8]. From then, RL methods have been broadly applied to various robotics control task from manipulation [9], [10] to navigation [11], [12], industrial manufacturing, production [13], [14] and autonomous vehicle control [15], [16].

RL based methods in robotics are getting more and more popular, quite a few research papers have been published on this particular topic. However, no systematic review paper has been published solely on the applications of RL-based methods in robotics. Therefore, new researchers in this area might find it difficult to understand the existing literature. This deficit motivated us to list the existing works since 2010 in a systematic manner and present them in a review paper. The primary contributions of this article are in the following sequence:

1) Provides an overview of a few states of the art RL techniques.
2) Provides a bibliographic analysis of a few selected papers on ''reinforcement learning'' and ''robotics'' based on data mining techniques, specific emphasis on the terms a) Keyword analysis, b) Citation analysis, c) Bibliographic coupling analysis.
3) This paper highlights a deep analysis of the topic ''reinforcement learning'' and ''robotics'' based systematic review techniques, and finally provides a qualitative and quantitative analysis of a few selected papers.

The residual of the paper is embodied as follows section II provides a short survey of the existent state of the art papers on RL based robotics. Section III presents an overview of RL algorithms. Section IV discusses the text mining-based bibliographical analysis techniques of existing works on RL-based robotics. Section V provides a systematic review of the selected papers of the RL-based robotics research domain. The concluding remarks are depicted in section VI.

II. LITERATURE REVIEW

RL has been around the corner as a research field for decades but extensively used in robotics not from very long ago. However, today it has been an optimal part of robotic learning. Innovation, development, and research are remarkably fast-paced and are getting better every day. In a broader sense, now RL is a captivating research interest for engineers, researchers, and academicians. Many surveys have been published in the domain of RL in the past to facilitate experts. This section summarizes some key aspects from previous surveys and points out the characteristics of the current papers that makes itself a noteworthy addition to the field. The first survey that we analyzed in this review dates back to the year 2000 [19], encompasses RL at an introductory level. The authors also highlighted RL can be a useful inventory in robotic soccer. In the next five years, researches were carried out in this field. But in 2008, Argall *et al.* [19] published their survey which was a breakthrough in the survey of RL in robotics. They concluded that learning from demonstration can be a solution for many challenges which a robotic learning system encounters. Biology inspired RL systems were reviewed in the following years. In 2012, Kirumasi *et al.* [23] remarked in their review that many researchers were interested in bio-inspired optimal adaptive control. Another remarkable review was done by Wang and Babuska [28] where the authors discussed several learning algorithms and their applicability to bipedal walking robots. In more recent years, [24] and [26] are published in 2017. The authors discussed various aspects of robot learning including emotion and put more emphasis on humanoid robot interpretation. In the field of robotic grasp detection, [25] provides many insights. The paper suggests from analyzing previous works that the sliding window approach is the most effective robotic grasp system to date. These review papers were not systematic review papers. However, a systematic review paper is necessary to outline the works done in this field. In this paper, we are going to try to give an overview of RL-based robotics research articles and point out some questionnaires and their answers as well. Table 1 summarizes the review articles published on RL based robotics. These papers can be

TABLE 1. Summary of previous review on RL-based robotics.

Ref	Year of publication	Keywords	Name of the Journal/Conference	Major contributions along with future research direction
[19]^B	2000	Multiagent systems, machine learning, survey, robotics, intelligent agents, robotic soccer, pursuit domain, homogeneous agents, heterogeneous agents, communicating agents.	Autonomous Robots	• Providing an overview on multi-agent systems in the framework of Machine Learning including Reinforcement Learning. • Main focus on robotic tasks rather other multi-agent tasks. • Presenting robotic soccer as an efficient field for the study of Multi-agent Systems (MAS).
[22]^B	2004	Mobile robots, machine learning, reinforcement learning, learning from demonstration.	CiteSeerX	• Providing a survey on the system of multi-robots based on nulti-agent RL techniques.. • Analysing the progression and barriers in the advancement of this field. • Future research direction points to learning robot's coordination, state and action abstraction of multi-robot system.
[20]^B	2006	Multi-agent systems, reinforcement learning, game theory, distributed control.	2006 9th International Conference on Control, Automation, Robotics and Vision	• Surveying the challenging issues in multi-agent RL. • Discussing the learning goal of multi-agent systems. • Future research into robotic RL algorithms might focus on the combination of multistep returns along with bootstrapping. • Claiming that more acute cross-fertilization between ML and control theory can bring unprecedented advancement in multi-agent systems.
[19]^D	2008	Learning from demonstration, robotics, machine learning, autonomous systems.	Robotics and Autonomous Systems	• Providing a review on the topic of robot Learning From Demonstration (LFD). • Finding few optimum parameters for LFD. • Mentioning the limitations of LFD.
[27]^A	2011	Model learning, robot control, machine learning, regression.	Cognitive Processing	• Reviewing the articles that focus on the advancement of model learning in case of controlling a robot. • Conducting a study on the model's architecture, discussing the challenges regarding implementation of these models in robotics, and finally mentioning some successful scenarios where model learning has been used. • Enabling learn continuously as well as get adapted to a new assignment is mentioned as a future research topic.
[28]^A	2012	Bipedal walking robots, learning control, reinforcement learning, supervised learning, unsupervised learning.	IEEE Transactions on Systems, Man, and Cybernetics	• Surveying recent progress of different ML algorithms including RL in relation to biped robot control. • Presenting a brief overview of the outcomes as well as drawbacks of these learning approaches and discussing the comparisons among these methods. • Mentioning that building hierarchical learning architectures is a prominent future research topic for overcoming the bipedal robot's complexities from control perspective.
[71]^A	2012	Reinforcement learning, ADP, Q-learning, optimal adaptive control	Annual Reviews in Control	• Providing a summarization of the literature that concentrates on the combination of RL and Optimal Adaptive Control Techniques to use in the field of robotics. • Highlighting prime techniques for the combination of these two research fields which are presented by famous researchers. • Implementing a Q-learning based Optimal Adaptive Controller in the arm of the humanoid Bristol-Elumotion-Robotic-Torso II (BERT II)
[88]^D	2013	Reinforcement learning, learning control, robot, survey.	The International Journals for Robotic Research	• Giving a review on those literature that focuses on RL for behavior generation in robotics. • Highlighting both prime complexities and mentionable success in this field. • Consulting about the contributions behind this success.
[21]^D	2015	Reinforcement learning, risk sensitivity, safe exploration, teacher advice.	Journal of Machine Learning Research	• Presenting a brief review on Safe RL. • Dividing the Safe RL technique into two classes and highlighting the existing literature based on these classifications as well as presenting a list of safe RL methods with advantages and drawbacks. • Future research direction includes developing the learning algorithms which can be directly utilized for robotics tasks.
[18]^D	2017	-	Journal of Sensors	• Giving a brief elucidation of Deep learning techniques. • Presenting a survey on latter usage and appliances of DL in Unmanned Aerial Vehicles (UAVs). • Pointing out the complexities of DL appliances in UAVs. • Designing more proficient DL architecture is pointed out as a future research direction.
[23]^A	2017	Autonomy, data-based optimization, reinforcement learning (RL).	IEEE Transactions On Neural Networks and Learning System	• Highlighting and reviewing the literature which discuss about RL techniques that solve optimal control problems. • Discussing about speeding up the learning process by using experience reply technique. • Discussing core algorithms of continuous-time and discrete-time systems. • Claiming the usage of deep RL approaches is a prominent future research direction for non-linear Optimal Feedback Controllers (OPFB).

TABLE 1. ***(Continued.)*** **Summary of previous review on RL-based robotics.**

Ref	Year of publication	Keywords	Name of the Journal/Conference	Major contributions along with future research direction
[24]^C	2017	Reinforcement learning, emotion, motivation, agent, robot.	Machine Learning	• Presenting the very first survey on the mathematics of RL-based robot emotions. • Providing a systematical overview of underlying computational models from which emotions can be derived in robots. • Pointing out the challenges for implementing emotions in RL robots. • Mentioning the translation of psychological concepts in mathematical expressions as a leading future research direction.
[26]^C	2017	Deep learning, human-level agents, reinforcement learning, robotics, survey, system design.	IEEE Access	• Giving a literature survey on the DRL techniques for building human-level agents. • Finding out the most suitable models and also examining them in case of human-level agents. • Giving an overview for constructing frameworks as well as providing a list of existing frameworks for this type of autonomous systems.
[89]^C	2017	Intelligent robotics, machine learning, model-based reinforcement learning, robot learning, policy search, transition models, reward functions.	Journal of Intelligent and Robotic Systems	• Presenting a survey on the robots built on the basis on model-based RL. • Dividing the model-based methods in terms of return function, optimal policy, and the learned tasks. • Discussing the advantage of model-based methods over model-free methods and also the applicability of model-based methods on various robots such as Rethnik Robotics Baxter, Universal Robots UR10.
[25]^D	2018	Deep learning, deep convolutional neural networks, dcnn, convolutional neural networks, cnn, robot learning, transfer learning, robotic grasping, robotic grasp detection, human-robot collaboration.	Multimodal Technologies and Interaction	• Surveying Deep Learning techniques including Deep RL for grasp detection task of robotics. • Identifying the most prominent DL approaches for robotic grasp detection and mentioning the one-shot method as the top one. • Pointing out the usage of RL techniques in grasp detection problem as a future research topic.
[29]^A	2018	Spiking neural network, brain-inspired robotics, neurorobotics, learning control, survey.	Frontiers in Neurorobotics	• Giving a review of the literature that focuses on controlling robots using biologically-inspired Spiking Neural Networks (SNNs.) • Classifying and analyzing SNN-based robotic appliances as well as mentioning some platforms for SNN-based robotic interactions. • Developing complex brain models to embed into real robots is a leading future research topic.

broadly categorized into 4 groups based on their commonalities. Table 2 represents the groups and their corresponding papers.

TABLE 2. **Representation of the categorized group and the corresponding survey papers.**

Group	Reference
A (Control of robots)	[23], [27], [28], [29], [71]
B (Multi-agent system)	[22], [20], [19]
C (Different RL models)	[24], [89]
D (Miscellaneous)	[18], [25], [88], [26], [21], [19]

III. OVERVIEW OF REINFORCEMENT LEARNING

RL is a sub-set of ML that is a learning process through association and navigation for controlling a system. An RL setup is composed of a decision-maker called an agent that learns by interacting with an environment consisting of different states $s_t \in S$ (existing situations returned by the environment) rather than being taught by any explicit agent. The agent interacts with the environment as like as solving a Markovian Decision Problem (MDP) by taking available actions $a_t \in A$ (a set of activities that an agent does in its environment) randomly (exploration) or after sometimes by it's experience gained from the environment as a probability distribution of a policy such as $\pi(A|S) = P_r(a_t = A|s_t = S)$ to increase the rewards with less hurdles (exploitation) in a particular moment t and reach to the following state s_{t+1} of the environment. Through taking any action; the agent acquires reward $r_t \in R$ from the environment at a particular time step t which describes the success of any particular action where penalties may be represented by negative numbers. The key determination of the RL method is to augment the reward by trial and error method or utilizing a model over the long run. Figure 2 illustrates the agent-environment interaction protocol. The primary aim of RL is to augment the reward by selecting actions based on an algorithm. A comprehensive variety of RL algorithm exists to solve RL problems.

A. LEARNING APPROACHES

The learning approaches of an RL agent or RL algorithms can be categorized into two classes:

- Model-based RL or Indirect learning: A predictive model is employed by the agent for learning about the control policy from the environment by a relatively reduced number of interactions and then the agent

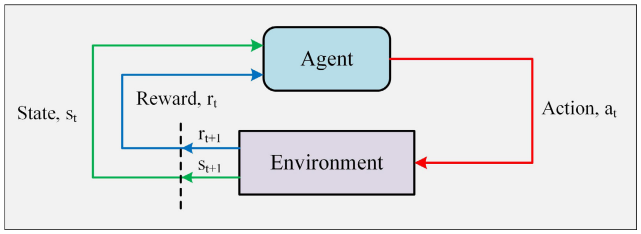


FIGURE 2. Agent-environment interaction protocol.

utilizes the model to the following episodes for getting the rewards.

- Model-free RL or Direct learning: The agent learns about the control policy from the environment by trial and error (i.e. experience) to maximize rewards without any model.

Figure 3 shows a visual representation of the RL algorithm classification and the mathematical description of a few RL algorithms shown in Figure 3 has been described in the appendix section.

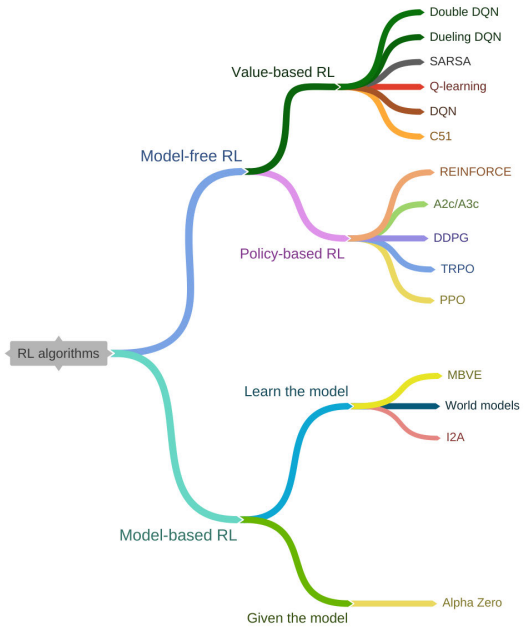


FIGURE 3. Classification of RL algorithms.

In comparison model-based RL with model-free RL, model-free RL has proved itself as a promising approach in the field of robotics [30]. Table 3 presents a quick view on the various RL algorithms used in the development of robots in different studies.

IV. BIBLIOMETRIC ANALYSIS

Research on RL-based robotics is expanding swiftly, so it is a tough task to conduct reviews manually. So, to reduce labor, this work has taken the benefit of keywords and citations, bibliographical data and utilized a text-mining

TABLE 3. RL algorithms used in for robot development and its applications in various studies.

Algorithm	Corresponding article
Q-learning	[31], [32]
SARSA	[33], [34]
DQN	[35], [36]
Double DQN	[37], [38]
Dueling DQN	[39], [40]
C51	[41], [42]
A2C/A3C	[43], [44]
DDPG	[45], [46]
TRPO	[47], [48]
PPO	[49], [50]

software named VOSviewer. VOSviewer is a prominent automated tool [57] to generate several bibliometric maps on the research field as well as for analyzing a large number of articles efficiently from a different perspective. The functionalities of the software which we have used for the process are (1) Import information about publications such as publication year, corresponding journal/conference, number of citations and global distribution of the publications. (2) Create a co-occurrence map using the keywords and visualize it in a network form (3) Retrieve data for citation analysis (4) Visualize bibliometric coupling analysis using three units (sources, authors, countries).

A. PUBLICATION COLLECTION

The bibliographic dataset which is used for this research has been collected from the Web of Science (WOS) repository on 07 April 2020 using the keywords "reinforcement learning" and "robotics". This research limits the searching duration in between the years 2010 to 2020. By utilizing this search string, this work downloaded some publication information like title, abstract, keywords, source of journals, etc. in the form of a text file that is suitable for VOSviewer software. A total of 372 papers on RL based robotics were retrieved by using this process. Then, this work analyzed the papers in imitation of their publication rate of the year, geographical distribution, journal source, and we analyzed the papers deeply in terms of the keywords, citation, bibliographic coupling which provided a more extensive understanding of this research area.

B. PUBLICATION ANALYSIS

This work employed three analysis techniques in this study to bring forth a primary outcome on the advancement and future propensity of robotics research based on RL. The first technique this work used is keyword analysis. In this method, the software takes the keywords into account which can be found in abstracts as well as titles of the articles and results in a scientific landscape by which the development of RL topics and recent research terms can be revealed with which future researchers can pursue their studies. In the second technique, this work analyzed the highly-cited articles which are playing

a prominent role by working as a stock of knowledge in the arena of RL-based robotics research. The third one is bibliographical coupling analysis based on sources, authors, and the countries of the publications. This analysis results in scientific landscapes that show relatedness between two papers which cited a third publication on their reference. The techniques are depicted below.

1) KEYWORD ANALYSIS

For visualizing the scientific landscape in a networked form all the keywords were excerpted from the titles and abstracts of the articles of the dataset which was downloaded from WOS. Initially, 1535 keywords were extracted from the dataset. Then this work experimented to observe how the number of keywords varied with different numbers of co-occurrence. The co-occurrence of keywords means that the keywords are present in a single document. Finally, this work filtered the keywords with a minimum of co-occurrence of 10 words and the VOSviewer software used it's text mining function to generate a scientific landscape of co-occurrence map. The keywords were divided into three different coloured clusters (set of keywords) according to their relatedness and distinct types. The keywords which were used to create the co-occurrence map are different in size. The big keywords have a higher weight that means the keywords have occurred many times. And if the size is small it means that the occurring frequency of that specific keyword is low. Another important term to analyze is the distance between the keywords. If the distance between the keywords is small, it means the link strength between the keywords is high and they co-occur frequently. But if the distance is long, then it represents that the keywords do not co-occur frequently.

2) CITATION ANALYSIS

Citation analysis is another renowned technique to analyze the influential publications on any field and which establish links between the articles based on researchers, journal, countries, etc. To find out the most influential works on our chosen field over the last ten years, this work developed a table by using the bibliographical information which was extracted by VOSviewer from our collection of those publications. The papers were cited 4448 times in total; however, this work has selected the papers which have cited a minimum 30 times for this analysis.

3) BIBIOGRPAHICAL COUPLING ANALYSIS

One more analysis technique this work has considered in the presented study is the bibliographical coupling. Bibliographical coupling occurs when two scientific published documents say x and y use another publication z as their reference. In this work, this work has analyzed bibliographical coupling by using articles, sources of articles, authors, and countries as the unit of analysis. Like keyword analysis, in this case, this work also conducted an experimental study by varying the minimum number of citations of a document, the minimum number of documents of any sources, authors, and countries to observe the variation tendency of the articles, sources, authors and countries which allow to fix the minimum number of citations.

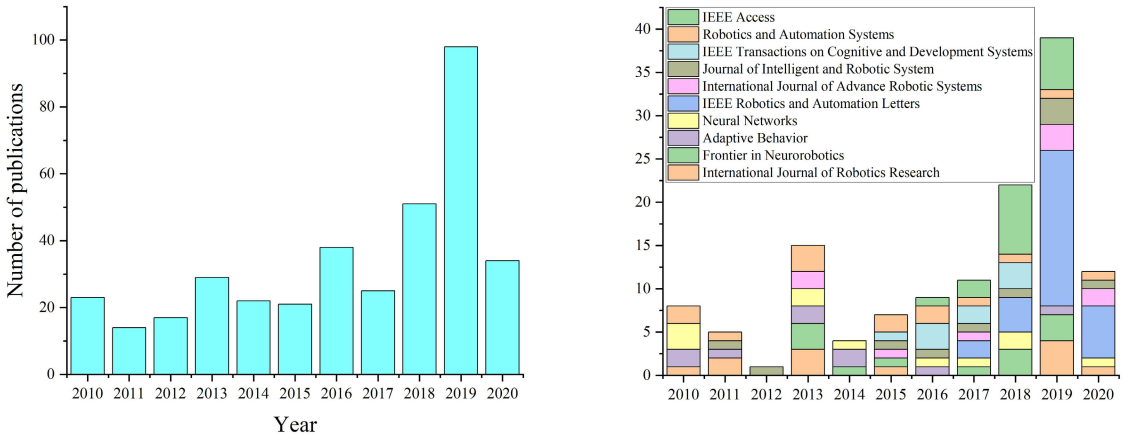
C. RESULTS

1) OVERVIEW OF THE PUBLICATIONS

By using the search strings, this work was able to retrieve a total of 372 papers between 2010 and 2020 which is illustrated in Figure 4 (a) (using WOS). From this figure, it can be seen that the number of papers fluctuates from the year, 2010 to 2016. After then from 2017 to 2019 it follows an upward trend; a total of 174 papers were published in this time which is about 46.77 % of all the published papers during our year range. In 2019 the publication number boomed as it reaches 91 which is the highest in any single year. The research papers published in 2019 consists of 26.34% of all the published papers in RL based robotics studies. Till April of 2020 already 34 papers have been published which is a clear indication of increasing publication rate during this year. In total 145 major publication sources were found in which a total of 372 papers were published. From them, this work has extracted the top 10 sources which are responsible for around 35% (133 papers) of all the papers; where each journal publishes 3.5 papers on an average. The rest 62% (235 papers) were published from other 135 sources where each source carries a publication rate of 1.77 on average. Figure 4 (b) describes the historical advancement of those top publication sources. From Figure 4 (b) it can be seen that the journal ''IEEE Robotics and Automation Letters'' has published the highest papers (30 papers) over the years. ''IEEE Robotics and Automation Letters'' and ''IEEE access'' are the two publication sources where RL based robotics papers have been published in recent years. While ''Neural Networks'', ''Journal of Intelligent and Robotic System'', ''Robotics and Autonomous Systems'' published papers almost all of the years.

2) GLOBAL DISTRIBUTION OF PUBLICATIONS ON RL BASED ROBOTICS STUDIES

From the collection of publications on RL based robotics, this work found that the articles were published from 49 different countries. Figure 5 describes the distribution of the papers in different parts of the world. If this work divides all the countries in some region like the European region, North and Latin American region and the Asian region, then it can be seen that Europe is the leading region for publishing scientific documents; because the European region consists of around 41% of all the publications, the American region is responsible for 27.5% of the papers and the third region Asian region published 24.76% of all the publications. And the rest 6.74% is from the other countries of the world. If this work goes through a deeper analysis, then this work will get to know that Germany is the leading country in the European region by publishing 23% papers of the European region. USA rules the North American region where about 55.7%



(a) Number of RL-based publication over the years (2010-April 2020).

(b) Historical development of RL-based robotics related studies at different journals

FIGURE 4. Summary of the number of publications based on RL-based robotics over the years.

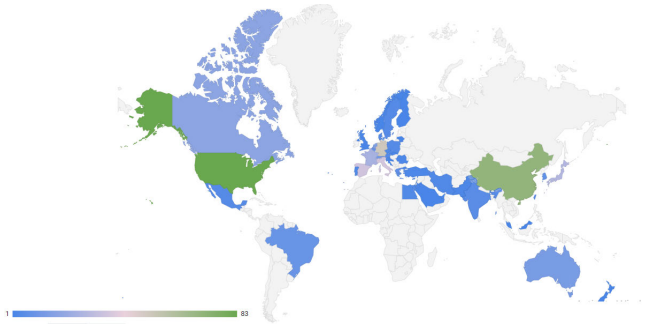


FIGURE 5. Distribution of RL-based robotics publications in the world.

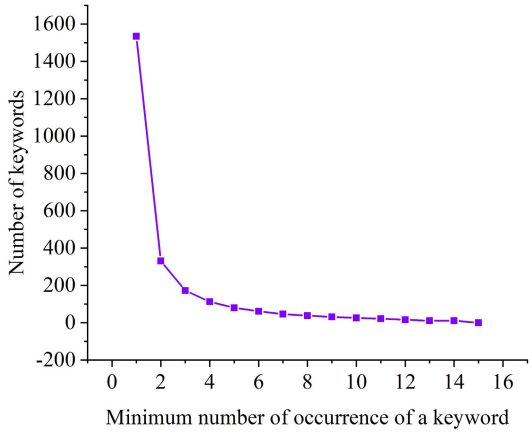


FIGURE 6. Variation of keywords number according to minimum occurrences.

of papers are from the USA. Among the Asian countries, China is the leading one in respect of publishing papers. 52.2% papers come from China of the Asian region. From the analysis, it is clear that researchers from different regions can collaborate with the researchers of the USA, Germany, and China to bring development in the field of RL based robotics in their countries.

3) SCIENTIFIC LANDSCAPE OF KEYWORDS OF RL BASED ROBOTICS

From the dataset, this work utilized for this experiment total 1535 keywords were found. Then this work conducted an experiment which is described in section 1.2.1. Figure 6 shows the variation of keywords number according to the different number of co-occurrences.

Finally, this work decided to filter the keywords with a minimum co-occurrence 10 and got twenty-six keywords in total. The co-occurrence map of the selected keywords from abstracts and titles of the publications is visualized by Figure 7. A keyword or term is presented by each circle on the map. The utilizing frequency of a keyword is indicated by its size. In the terminology of frequency, the ''reinforcement learning'' is the keyword with the largest one (164). Other higher utilizing frequency keywords include: ''robotics'' (60), ''reinforcement'' (31), ''deep reinforcement learning'' (21). In this graph, the distance between the keywords shows the co-occurrence frequency. If the distance between the keywords is small, then the co-occurrence number is high and if the distance is long, the co-occurrence number is low. From Figure 7, it can be seen that there are a total of 4 clusters; Green, Red, Blue, and Yellow and Table 4 summarizes the resulting clusters. The red cluster contains the highest number of keywords. It focuses on the ''optimization'' with close linkage such as ''robot'', ''deep learning in robotics'', ''algorithm''. The highly occurred keyword in the red cluster is ''deep

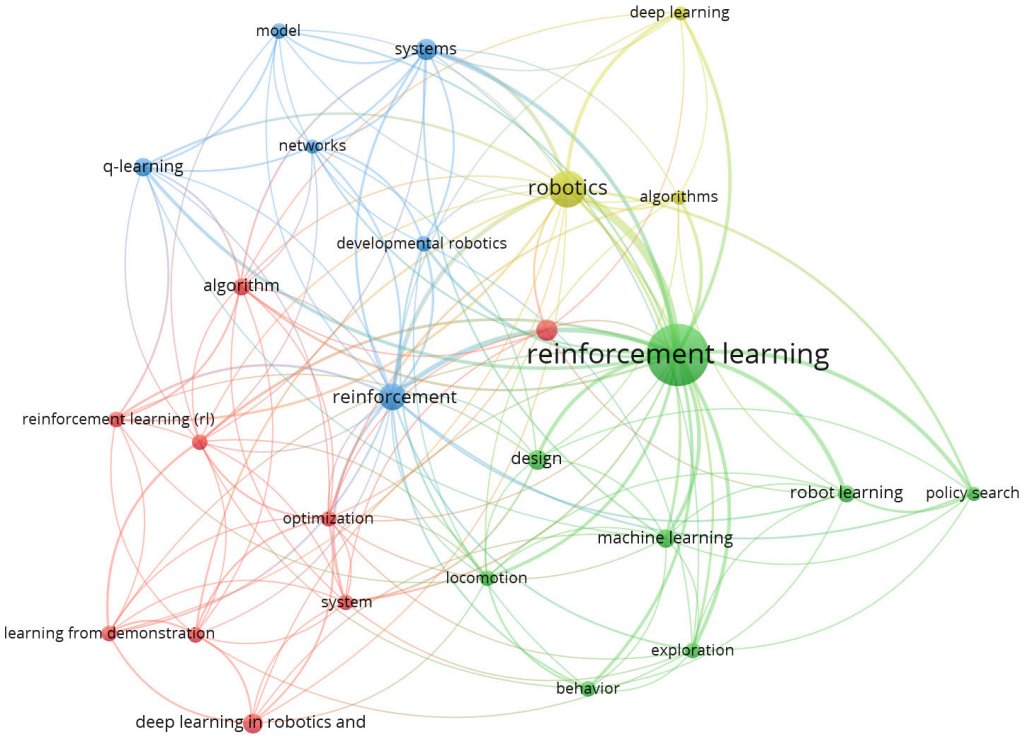


FIGURE 7. Co-occurrence map of keywords.

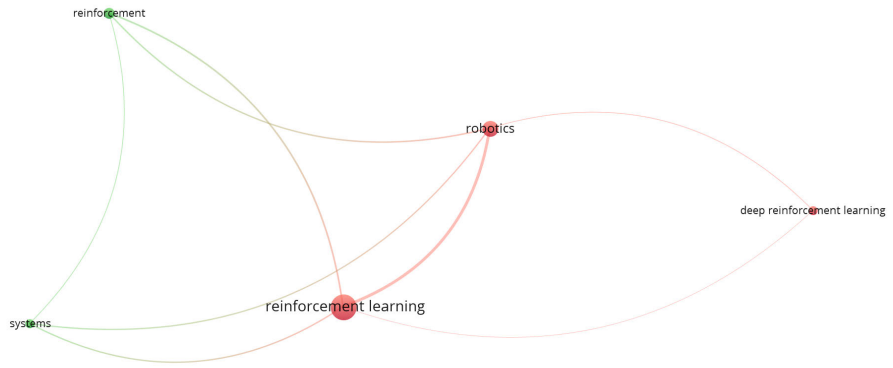


FIGURE 8. Co-occurrence map of top 5 keywords.

reinforcement learning'' but it is connected with only three keywords; ''algorithm'', ''optimization'', ''system''. The green cluster reunites ''reinforcement learning'' and which is the most occurred keyword in our publications with keywords such as ''policy search'', ''behavior'', ''exploration'' highlighting the various important terms of RL. The blue cluster's core is on the ''reinforcement'' with a close connection with the ''q-learning'', ''developmental robotics'' highlighting the branches of the reinforcement learning. Finally, the yellow cluster consists of only three keywords; ''robotics'' is a highly occurred keyword with ''deep learning'' and ''algorithms'' in this cluster. After that, this work extracted the top 5 keywords of our dataset, and visualize the co-occurrence map in Figure 8 by using VOSviewer. The top 5 keywords are ''reinforcement learning'' (164), ''robotics'' (64), ''reinforcement'' (31), ''deep reinforcement learning'' (21), ''systems'' (18). The keywords are divided into '2' clusters, contain 8 inks with a total link strength of 79. Table 5 summarizes the resulting cluster.

Among the keywords ''reinforcement learning'' and ''robotics'' is connected with all other keywords, so they have 4 links and their link strength are 58, 55 respectively.

TABLE 4. Result of keyword clustering by the publication.

Cluster color	Observed keywords	No. of keywords
Green	reinforcement learning, design, locomotion, behavior, exploration, machine learning, robot learning, policy search	8
Blue	reinforcement, systems, model, q-learning, networks, developmental robotics	6
Red	robot, algorithm, optimization, learning from demonstration, system, reinforcement learning (RL), deep learning in robotics and, learning and adaptive systems, deep reinforcement learning	9
Yellow	robotics, algorithm, deep learning	3

TABLE 5. Strengths of keywords.

Clusters	Keywords	Links	Total link strength	occurences
Red	reinforcement learning	4	58	164
	robotics	4	55	60
	deep reinforcement learning	2	3	21
Green	reinforcement	3	24	31
	systems	3	18	20

''reinforcement'' and ''system'' are connected with 3 keywords except ''deep reinforcement learning'' so they have 3 links. And ''deep reinforcement learning'' has 2 links. It is connected with the keyword of its belonging cluster. In the following section this work propulsion a citation analysis to identify puissant research articles and their contribution to the expanding reinforcement-learning based robotics research.

4) TOP CITED PAPER IN RL BASED ROBOTICS

Total 34 published articles which are the most influential based on their citation were found by utilizing the methodology described in the citation analysis section. Those papers were cited in a total of 6473 times from Google scholar, 2923 times from researchgate, and VOSviewer shows 2498 times. The most highly cited publications are shown by their title, publication source, corresponding author's name, corresponding author's country, publication year, total citation (Google Scholar, Researchgate, VOSviewer), and norm. citation in Table 8. The highly cited article was a review paper entitled ''Reinforcement Learning in Robotics: a Survey'' authored by ''Jens Kober'' in ''International Journal of Robotics Research'' with 1454 citations in Google scholar, 1089 citations in researchgate, 405 citations in VOSviewer and 10.4 norm citations. The paper which took second place was published in ''Machine Learning'' entitled ''Policy Search of Motor Primitives in Robotics'' which was also authored by ''Jens Kober'' with 595 Google scholar citation, 279 researchgate citation, and 105 VOSviewer citation. Among the 34 papers; 20 papers (58.8 %) were from the European region which is proof of the tremendous influence of the European region on RL based robotics. If this work goes for a deep analysis, Germany (6) and England (5) are the leading countries of the European region in this field. In contrast, 5 articles came from the USA and the rest were from a few other countries (e.g Canada, China, Netherlands). The highly influential scientific research journal on this field is ''International Journal of Robotics Research'' published the highest number of articles (5) among the top 34 papers. The article which received the least number of citation (40) was authored by ''G. Nores'' entitled ''Reinforcement Learning of Self-regulated Sensory-motor beta-oscillations Improves Motor Performance'' in ''Neuroimage''. After analyzing the publications based on citations, this work has also extracted three different lists of top 10 citations receiving sources, authors, and countries. Table 6 summarizes the lists.

From Table 6 (a) it can be seen that the ''International Journal of Robotics Research'' received the highest citation (694) from 12 documents. One interesting thing in this table to notice that ''IEEE Transactions On Systems man and cybernetics part-c applications and review'' acquired 209 citations from only 2 documents and ''Trends in Cognitive Science'' got 155 citations from only one document, whereas ''Neural Networks'' and ''Robotics and Autonomous System'' got 168 and 157 citations respectively from 11 and 13 papers. This is a clear indication that the ''IEEE Transactions On Systems man and cybernetics part-c applications and review'' and ''Trends in Cognitive Science'' have a high citation rate per document. Table 6 (b) shows that Jan Peters has the highest citation (705) from 16 documents. Jens Kober claims second place with 568 citations. Though J. Andrew Bagnell has fewer documents (2) than Robert Babuska (5) and Marc Peter Deisenroth (4), he received more citations (445). From Table 6 (c) it can be seen that the USA and Germany have the highest citation 1795 and 1156 respectively. Among the 5 countries with high citation; 4 countries are from Europe; which indicates the influence of Europe on this field. In this table, there is no anomaly between the document number and the received citation. Countries with higher documents have received higher citations.

5) SCIENTIFIC LANDSCAPE OF BIBLIOGRAPHIC COUPLING

Bibliographic coupling describes the relatedness of two articles based on their virtue of referencing the same article. This work has conducted some experimental study on the units (articles, journals, authors) which were used for bibliographic coupling analysis in this study by varying the minimum number of citations of the articles, the minimum number of documents of a journal, and an author. Figure 9 shows the variation tendency of the articles, journals, and authors according to their threshold measurement item. For a better analysis of articles, this work considered a threshold of minimum 60 citations of a document and found 15 (4 % of total publications) which met the threshold. Table 7 (a) shows the publication with the highest indices of bibliographic

TABLE 6. Citation based on sources, authors and countries.

(a)

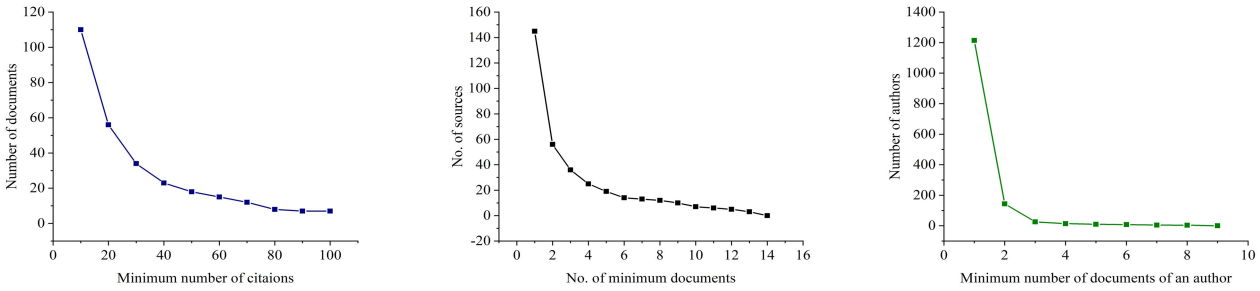
Sources	Documents	Citations
International Journal of Robotics Research	12	694
IEEE Transactions On Systems man and cybernetics part-c applications and review	2	209
Neural Networks	11	168
Robotics and Autonomous Systems	13	157
Trends in Cognitive Science	1	155

(b)

Author	Documents	Citation
Jan peters	16	705
Jens Kober	6	568
J. Andrew Bagnell	2	445
Robert Babuska	5	225
Marc peter deisenroth	4	286

(c)

Country	Documents	Citations
USA	83	1795
Germany	51	1156
England	39	715
Italy	36	500
France	25	498



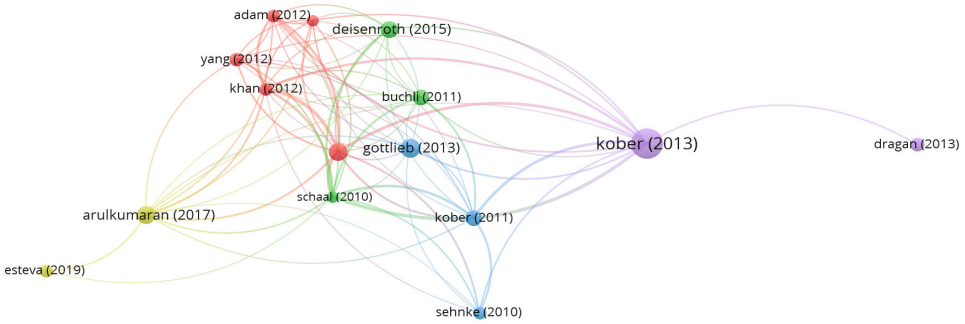
(a) Variation of articles (b) Variation of sources (c) Variation of authors

FIGURE 9. Diversification of articles, sources, and authors.

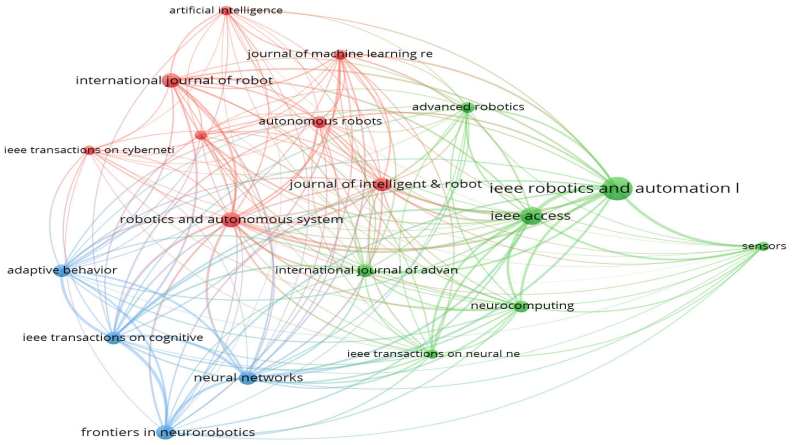
coupling. The Top five publications of the highest indices from Table 7 (a) are:

- J. Kober, J. Andrew Bagnell, Jan Peters (2013). Reinforcement Learning in Robotics: A Survey. The International Journal of Robotics Research, 32(11), 1238-1274. [88]
- Ivo Grondman, Luvican Busoniu (2012). A Survey of Actor-critic Reinforcement Learning: Standard and Natural Policy Gradients. IEEE Transactions on Systems Man and Cybernetics-Part C- Applications and Reviews, 42(6), 1291-1307. [93]
- Stefen Schall, G. Cristopher Atkeson (2010). Learning Control in Robotics Trajectory-based Optimal Control Techniques. IEEE Robotics & Automation Magazines, 17(2), 20-29. [70]
- G. Said Khan, Guido Lewis, L. Frank, Tony Pipe (2012). Reinforcement Learning and Optimal Adaptive Control: An Overview and Implementation Examples. Annual Reviews in Control, 36(1), 42-59. [71]
- Jens Kober, Jan Peters (2011). Policy Search for Motor Primitives in Robotics. Machine Learning, 84(1-2), 171-203. [92]

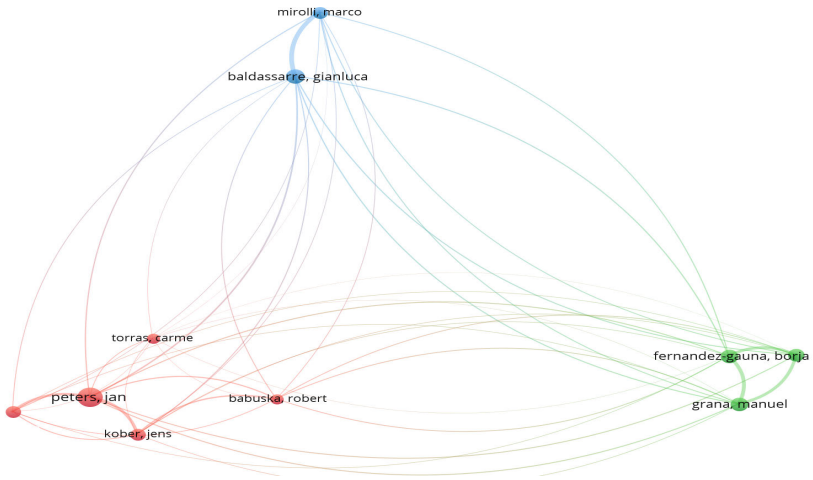
For overall analysis, this work presented the scientific landscape of bibliographic coupling of this indicator (i.e bibliographic coupling of publications. From Figure 10 (a) it can be seen that there are a total of 5 clusters (red, green,



(a) Network diagrams of the bibliographic coupling of articles.



(b) Network diagrams of the bibliographic coupling of sources.



(c) Network diagrams of the bibliographic coupling of authors.

FIGURE 10. **Bibliographic coupling of items.**

blue, yellow, violet). That means the paper has been grouped by the clustering technique of Vosviewer mapping software into 5 groups based on their relatedness. As bibliographic coupling occurs between two papers when they cite the same articles as reference, it can be said that the papers have the same research interest or the papers have taken the same research area or domain of the RL techniques into consideration. The distance between the clusters is the parameter for measuring the relatedness between the papers of different clusters, the distance between the papers of any particular cluster is the parameter for measuring the relatedness of the papers situated in the same cluster, the node size of each item represents the number of citations got by any paper, the thickness of the connection among the papers represent the total link strength i.e the number of commonly cited articles by each paper. If the work goes for a detailed interpretation, then it can be observed that the red, blue, and green clusters are closely connected to each other where the violet and yellow clusters are situated far away from those 3 clusters which is a clear indication of the exclusivity of these two clusters. Now, if the work digs deeper and analyzes each cluster individually, it can be noticed that the red cluster contains a total of 5 papers with a total link strength of 140 which is predominantly composed of documents published in 2012. The papers of this cluster have common citation linkage with the papers of each cluster. The most influential paper of this cluster is grondman (2012) which has a total link strength of 65 and is strongly connected with the papers of red cluster and one paper from violet cluster kober (2013). The green cluster holds the second position among the clusters in terms of total link strength (107). This cluster contains three papers and among them, schall (2010) has the common citation linkage. The paper of the green cluster is also connected with all other clusters. In the case of blue cluster, among the three papers, the document kober (2011) plays the centric role for this cluster which contains a total link strength of 42. It can be noticed that this paper is not connected with so many papers; however, it shares a great number of common references with one paper from green cluster schall (2010) and another one from violet cluster kober (2013). This cluster is also connected with all other clusters but has a lower link strength (79) that represents the least number of commonly cited articles. The next one is the violet cluster 12 which consists of only 2 papers and has a total link strength of 70. However, The article with the highest indices kober (2013) is placed on the violet cluster. This paper is connected with almost all the papers except any papers of yellow clusters; which means that there are no commonly cited articles between violet and yellow clusters. So it can be said the research area of these two clusters is totally different. The another paper of this cluster dragan (2013) only have common citation linkage with kober (2013) and the interesting thing is that though this paper is sharing the same cluster with kober (2013), the thickness of the connection between the papers is too low which clearly indicates the unique research domain of this paper. Finally, the yellow cluster consists of only 2 papers with a total link strength of only 35. This cluster is also exclusive in terms of research areas as it has less relatedness with the other clusters.

For analyzing the sources, this work considered a threshold of 5 publications of a source and thus 19 journals (i.e. 13 %) of total sources) were founded. Table 7 (b) shows the journals which are top by the number of documents published are ''IEEE robotics and Automation Letters '' (30 papers), ''IEEE access'' (17 papers), ''Robotics and Autonomous Systems'' (13 papers). The journals with strong link strength or highest indices of bibliographic coupling are at the top of the list. The journals are placed from top to bottom according to indices of bibliographic coupling. The top 5 journals of the list according to the indices are: ''IEEE Transactions on Cognitive and Developmental Systems'', ''IEEE Robotics and Automation Letters'', ''Frontiers in Neurorobotics'', ''Neural Networks'', and ''Robotics and Autonomous systems''. Figure 10 (b) visualizes the bibliographic coupling of journals in networked form. From Figure 10 (b), it can be seen that there are a total of 3 clusters (red, green, blue). If the work goes for a detailed interpretation, then it can be observed that the red cluster is situated between the rest two clusters which indicates that the sources situated in the red cluster publish paper having common research interest with the other sources. However, the green cluster and blue cluster is situated far away from each other. So, it can be said that the sources don't accept papers of the same research interest. Now, if the work digs deeper and analyzes each cluster individually, it can be noticed that the red cluster contains a total of 8 sources and has a total link strength of 4762. The prominent source of this cluster is the journal ''Robotics and Automation System'' which has links with a total of 18 sources and cited common articles with the other sources about 1183 times. The other sources of this cluster have also been linked with a total 18 sources; that means that all the sources of this cluster have cited common papers with all other sources of the rest two clusters. It should be highlighted that the red cluster is the biggest among the clusters but it does not contain any journals of the top 5 in terms of highest link strength. The green cluster contains 7 sources and has a link strength of 5989 which is highest among the clusters. This cluster contains the source ''IEEE Robotics and Automation Letters'' which has published the highest number of documents (30) among the sources and has the highest link strength (1513), which is a clear indication of the versatility of this source. Among the 5 sources of this cluster, 2 sources are in the list of top 5 sources in terms of highest link strength. And finally, the blue cluster contains only 4 sources with a total link strength of 4005. The top source of this cluster is ''IEEE Transactions on Cognitive and Developmental Systems'' which published only 9 documents but cited a vast number of common papers (i.e total link strength 1261) with other sources in those documents. Among the top 5 journals based on indices of bibliographic coupling 3 journals are present in this cluster though it contains only 4 sources. So, it can be said that the research domain of this cluster is quite similar to the rest others. For analyzing the authors, this work

TABLE 7. Bibliographic coupling indices of documents, sources, authors.

(a)

Document	Citation	Total Link Strength
Kober (2013)	405	67
grondman (2012)	138	65
schall (2010)	64	65
khan(2012)	68	56
kober (2011)	105	42
adam (2012)	71	36
arulkumaran (2017)	137	31
Yang (2012)	84	26
deisenroth (2015)	121	22
konar (2013)	61	22
buchli (112)	112	20
gottileb (2013)	155	19
sehnke (2010)	72	18
esteva (2019)	74	4
dragan (2013)	79	3

(b)

Source	Documents	Citations	Total Link Strength
IEEE Robotics and Automation Letters	30	44	1513
IEEE Access	17	69	1400
IEEE Transactions on Cognitive and Developmental Systems	9	57	1261
Robotics and Autonomous Systems	13	157	1183
Neural Networks	11	168	1103
Frontiers in Neurorobotics	12	94	868
Adaptive Behavior	9	74	773
Neurocomputing	8	15	767
Journal of Intelligent & Robotic Systems	10	68	760
Internantional Journal of Robotics Research	12	694	730
International journal of Advanced Systems	9	30	701
Journal of Machine Learning Research	6	28	622
IEEE Transactions on Neural Networks and Learning Systems	5	22	622
Autonomous Robots	8	37	526
Sensors	5	2	506
Advanced Robotics	7	26	486
Engineering Aplications of Artificial Intelligence	5	23	397
Artificial Intelligence	5	52	291
IEEE Transactions on Cybernatics	5	97	253

(c)

Authors	Documents	Citations	Total Link Strength
Jan Peters	16	705	1519
Giancula Baldassare	9	108	1531
Jens Kober	6	568	876
Macro Mirolli	6	100	1352
Girhard Neumman	6	33	703
Borja Fernandez-Gauna	8	83	1583
Manuel Grana	8	79	1571
Joss Manuel Lopez-guede	7	69	1411
Robert Babuska	5	225	494
Carme Torras	16	705	96

considered a threshold of 5 documents of an author and thus this work found that only ten authors are responsible for publishing more than or equal to 5 articles. Table 7 (c) shows the complete list of authors with their bibliographic coupling indices. To complete a thorough analysis of this criterion (i.e bibliographic coupling of authorship) network visualization is presented in Figure 10 (c). It can be noticed from the Figure 10 (c) that there are total 3 clusters (red, green, blue) If the work goes for a detail interpretation, then it can be observed that all the clusters are far away from each other; that means that these authors have common papers in references but their research works are not so much related. If the work goes for cluster by cluster analysis it can be seen that the major cluster of authors is the red cluster containing five authors with a total link strength of 3688. The influential author of this cluster is ''Jan Peters'' (Germany) who has cited common papers with other authors about 1519 times. Among the five authors of this cluster, one author ''Jan Peters'' (Germany), is in the list of top five authors in terms of highest link strength. The green cluster consists of only three authors but the total link strength of this cluster is 4565. The prominent author of this cluster is '' Borja Fernnadez-Gavna'' who has cited common papers with other authors about 1583 times. Among the three authors of this cluster, all three authors are in the list of top five authors in terms of the highest link strength. This clearly indicates that these authors cited more common references than those of the red cluster. The blue cluster consists of only two authors and the total link strength of this cluster is 2883. The prominent author of this cluster is ''Giancula Baldassare (Italy)'' who has cited common papers with other authors about 1531 times. Among the two authors of this cluster, one author ''Giancula Baldassare'' (Italy), is in the list of top five authors in terms of highest link strength. It is noteworthy that, if this work considers the geographical point of view, European scholars showed stronger competence in the field of RL based robotics.

V. SYSTEMATIC REVIEW

In recent years, RL algorithms have acquired popularity because of their prominent success in the field of robotics. For having a better understanding and knowing about recent and previous works on RL, this work has presented a systematic review of RL-based robotics papers. Figure 11 represents the structure of the systematic review.

After eliminating review articles, we ended up with 32 papers. Twelve more papers were also excluded because they were more related to Human-Robot Interaction (HRI), Brain-Machine Interface (BMI), Bio constrained computational model, Human-Robot Collaboration (HRC), Brain-Robot Interface (BRI). Finally, this work was able to retrieve 20 papers on RL-based robots. After retrieving the papers, this work has analyzed those papers one by one based on the problem statement, research methodologies, and experimental results. These 20 papers have been grouped based on the type of RL used in them, i.e. value-based, policy-based, model-free algorithm, others.

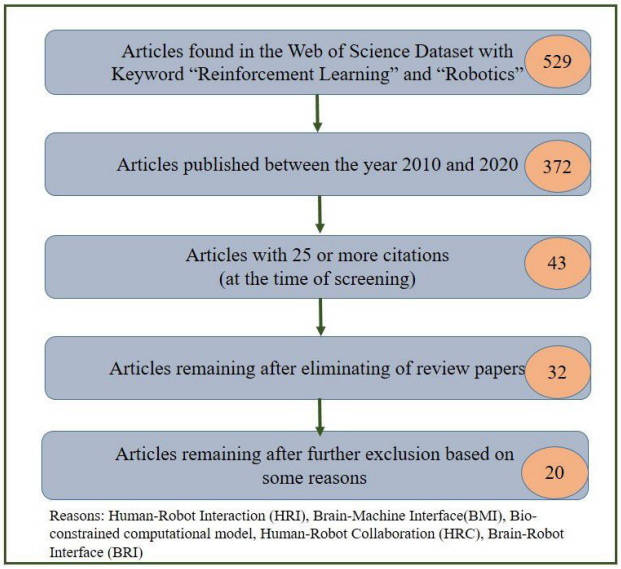


FIGURE 11. Structure of the proposed systematic review.

A. VALUE-BASED ALGORITHM

In [76], Ken *et al.* proposed a biologically inspired rat-like mobile navigator called Psikharpox. The robot was able to perform self-localization and navigate autonomously in an uncouth environment. The Psikharpox robot was implemented based on two navigation strategies with a brain-inspired meta controller for strategy selection. Ken *et al.* utilized a modified version of QL to make the robot enable for choosing the optimal technique to adapt in various situations. This work used ROS middleware for building the software architecture of this robot and evaluated the robot's performance by measuring the learning ability of goal selection and moving towards it on a simulation platform in an artificial environment. They conducted their experiment by utilizing different parts of their model individually to reach a fixed goal as well as by changing the goal. However, the experiment shows that the learning process was quite slow while the new goal was targeted; that means the adaptability of the robot in a new complex environment was not up to the mark like a real rat. However, to overcome this issue, they utilized a context switching mechanism.

In [77], Amit *et al.* presented an Improved Q-learning algorithm (IQL); that is a modified version of the classical Q-learning algorithm (CQL) for performing the path planning performance of a mobile robot. In case of a huge number of (state, action) pairs the space complexity is very high in CQL and thus it is a tough task to store the Q values and update the Q table for selecting an optimum action. The authors considered the problem and modified the CQL into IQL where only the best Q values of any particular state for available states will only be stored. Thus they reduced the problem of time complexity as well as space complexity and increased the performance of path planning tasks. After

TABLE 8. Summary of the top cited papers.

Title	Journal	Corresponding author	Country of corresponding author	Publication year	Cittaon Re-serachgate	Citations Google	Total Citations Vosviewer	Norm. Citations
Online Learning Of Task-driven Object-based Visual Attention Control [65]	Image and Vision Computing	Ali Borji	Iran	2010	55	70	38	2.09
Learning Control in Robotics Trajectory-based Optimal Control Techniques [70]	IEEE Robotics and Automation Magazine	Stefan Schaal	USA	2010	117	160	64	3.51
Imitation and Reinforcement Learning Practical Algorithms for Motor Premitives in Robotics [84]	IEEE Robotics & Automation Magazine	Jens Kober	Germany	2010	83	108	51	2.8
Learning From Demonstration for Autonomous Navigation in Complex Unstructured Terrain [90]	International Journal of Robotics Research	David Silver	USA	2010	73	114	40	2.2
Parameter-exploring Policy Gradients [91]	Machine Learning	Jan Peters	Germany	2010	136	183	72	3.95
Impedance Learning for Robotics Contact Tasks Using Natural Actor-critic Algorithm [95]	IEEE Transactions On Systems Man and Cybernetics Part b-cybernetics	Byungchan Kim	Korea	2010	56	71	43	2.36
Reinforcement Based Mobile Robot Navigation in Dynamic Environment [81]	Robotica and Computer-integrated Manufacturing	Mohammad Abdel Kareem Jaradat	Jordan	2011	64	135	59	1.96
Learning to Pour With a Robot Arm Combining Goal and Shape Learning fo r Dynamic Movement Primitives [85]	Robotics and Autonomous Systems	Minija Tamosiunaite	Germany	2011	71	90	42	1.39
Learning Variable Impedance Control [87]	International Journal of Robotics Research	Jonas Buchli	Italy	2011	-	224	112	3.72
Policy Search for Motor Primitives in Robotics [92]	Machine Learning	Jens Kober	Germany	2011	279	595	105	3.48
Modular Neuronal assemblies Embodied in a Closed-loop Environment: Toward Future Integration of Brains And Machines [64]	Frontier in Neurorobotics	Michela Chiappalone	Italy	2012	56	58	35	1.04
Reinforcement Learning Controller Design for Affine Nonlinear Discrete-time Systems Using Online Apporximators [67]	IEEE Transactions on Systems Man and Cybernetics Part b-cybernatics	Qinmin Yang	China	2012	-	-	84	2.51
Reinforcement Learning and Optimal Adaptive Control: An Overview and Implementation Examples	Annual Reviews in Control [71]	Said Ghani Khan	England	2012	95	129	68	2.03
Cola2: A control Architecture for AUVs [75]	IEEE Journal of Oceanic Engineering	Pere Ridao	Spain	2012	50	44	36	1.07
A Biologically Inspired Meta-control Navigation System For The Psikharpax Rat Robot [76]	Bioinspiration & Biomimetics	Ken Caluwaerts	France	2012	51	57	30	0.89
Experience Reply For Real-time Reinforcement Control [79]	IEEE Transactions On Systems Man and Cybernetics Part c-applications and reviews	Sander Adam	Netherland	2012	-	164	71	2.12
Obstacle Avoidance of Redaundant Manipulators Using Neural Networks Based Reinforcement Learning [80]	Robotica and Computer-integrated Manufacturing	Mihai Duguleana	Romania	2012	-	79	37	1.1
A survey of Actor-critic Reinforcement Learning: standard and natural policy gradients [93]	IEEE Transactions On Systems Man and Cybernetics Part c-applications and reviews	Ivo Grondman	Netherland	2012	230	358	138	4.12
Information-seeking, Curiosity, and attention: Computational and Neural Mechanisms [66]	Trends in Cognitive Sciences	Jacqueline Gottlieb	USA	2013	279	415	155	3.08
A Determinestic Improved Q-learning For Path Planning of a Mobile Robot [77]	IEEE Transaction on Systems Man Cybernatics-System	Sapam Jitu Singh	India	2013	98	115	61	1.57
Intrinsically Motivated Action-outcome Learning and Goal-based Action recall: A system level bio-constrained computational model [78]	Neural Networks	Gianluca Baldassarre	Italy	2013	48	58	33	0.85

TABLE 8. *(Continued.)* **Summary of the top cited papers.**

Title	Journal	Corresponding author	Country of corresponding author	Publication year	Cittaon Re-serachgate	Citations Google	Total Citations Vosviewer	Norm. Citations
A Policy-blending Formalism For Shared Control [82]	International Journal of Robotics Research	Anca Dragan	USA	2013	146	172	79	2.03
Fast Damge Recovery in Robotics With t-resilience Algorithm [83]	International Journal of Robotics Reserach	Jean-Baptiste Mouret	France	2013	47	69	31	0.8
Reinforcement Learning in Robotics: a Survey [88]	International Journal of Robotics Research	Jens Kober	Germany	2013	1089	1454	405	10.4
A Survey of the Ontogeny of Tool Use: From Sensory-motor Experience to Planning [94]	IEEE Transactions On Autonomous Mental Development	Frank Guerin	Scotland	2013	57	69	34	0.87
Using Reinforcement Learning to Provide Stable Brain-machine Interface Control Despite Neural Input Recognization [74]	Plos One	Justin Sanchez	USA	2014	-	51	30	2.58
Gaussian Processes for Data-efficient Learning in Robotics and Control [63]	IEEE Transactions on Pattern Analysis and Machine Intelligence	Marc Peter Deisenroth	England	2015	263	372	121	7.87
Evolutionary Dynamics of Multi-agent Learning: a survey [69]	Journal of Artificial Intelligence Research	Daan Bloembergen	England	2015	97	138	59	3.84
Socially Adaptive Path Planning in Human Environments Using Inverse Reinforcement Learning [62]	International Journal of Robotics	Beomjoon Kim	Canada	2016	54	91	49	5.1
Reinforcement Learning of Self-regulated Sensory-motor beta-oscillations Improves Motor Performance [68]	Neuroimage	Georgious Naros	Germany	2016	26	40	30	3.12
Deep Reinforcement Learning A brief Survey [73]	IEEE Signal Processing Magazine	Marc Peter Deisenroth	England	2017	-	336	137	9.03
Imitation Learning: a Survey of Learning Methods [86]	ACM Computing Surveys	Ahmed Hussein	England	2017	101	132	44	2.9
Survey of Model-based Reinforcement Learning: application on Robotics [89]	Journal of Intelligent & Robotic Systems	Athanasios Polydoros	Denamrk	2017	-	93	31	2.04
A Guide to Deep Learning in Health Care [72]	Nature Medicine	Andre Esteva	USA	2019	202	229	74	28.88

successfully modified the algorithm, they conducted some simulation experiments and real-life experiments also. For simulation experiments, they made a scenario of a 20×20 grid world where each grid represented a state. The assignment for the robot was to reach a targeted goal in the presence of obstacles or a free world. They divided the experiment session into two phases, such as the learning part and the planning part. They conducted the experiments by making some combinations

- Both the learning sessions and planning sessions were without obstacles.
- The learning session was without obstacles but the planning session was with obstacles.
- The learning session was with a few obstacles but the planning session was with more number of obstacles.

After conducting experiments they compare the IQL algorithm with CQL and EQL in terms of time for reaching the goal and number of 90-degree turns. They presented a table that described the effectiveness of IQL over those two algorithms in case of the performance measuring metrics. After the simulation experiments, they deployed the algorithm on a mobile robot named KHEPERA II robot. Then they conducted three different experiments on real-life environments with no obstacles, six obstacles, and four obstacles respectively. They measured the time for reaching the goal in seconds, the count of 90-degree revolve, and sate traversed by the robot for the three algorithms and proved that the IQL algorithm was superior to CQL and EQL by presenting with a table that contains necessary data.

In [79], Adam *et al.* utilized the Experience Reply (ER) approach which is promising in the case of RL techniques. EL learns from a very small amount of data and uses this data repetitively on underlying RL algorithms. Adam *et al.* used Q-learning and SARSA algorithms as the underlying approach for the ER approach. They first created a framework of ER and then combined it with Q-learning and SARSA algorithm to produce ER Q-learning and ER SARSA. The prime target of their work was to prove the effectiveness of ER in real-life environments as ER was previously utilized to simulation problems only. As a consequence, they tested their framework on three different platforms; such as a pendulum swing-up problem, a robotic arm manipulator, and a robotic goalkeeper by not only simulation but also real-life experiments. After conducting experiments they compare their framework with traditional RL algorithms and proved that the ER is an effective approach to use in a real-life environment.

In [80], Mihai *et al.* presented an approach that is capable to resolve obstacles avoiding challenges in the

time of robotic manipulation using the Double Neural Network-based Q learning algorithm. The focus of their work was to control a robotic arm named powercube attached with a mobile robot named powerbot that will be able to perform several tasks while navigating without colliding any obstacles. The authors designed their robotic arm using Computer-aided Three-dimensional Interactive Application (CATIA) software and modeled the Neural Network controller for obstacle avoiding as well as path planning tasks on an unknown complex environment using MATLAB. In this work, they also utilized the concept of Human-Robot Interaction (HRI) and for assessing it they used Virtual Reality (VR) which was simulated in A Cave Automatic Virtual Environment (CAVE) software and they used C++/Object-Oriented Graphics Rendering Engine (OGRE) for building the virtual environment. Then they used Transmission Control Protocol (TCP)/ Internet Protocol (IP) server communication for communicating between the VR environment and the Neural Network (NN) model. To evaluate the performance of their approach authors conducted both simulation and real-life environment experiments. In a simulation case, they conducted three different experiments by placing no obstacles on the working field, by placing a cylindrical obstacle on the working field and by placing four obstacles on the working field. And in the case of real-life experiments, there were several problems like hardware malfunctions, security issues and to overcome these problems they utilized the VR solution.

In [81], Jaradat *et al.* presented a new Q-learning based approach for obstacle avoidance as well as path planning strategy for transportable robots in an unfamiliar environment. Like most of the researchers, the authors didn't consider the static environment, rather they worked with a dynamic environment where obstacles and goals might be dynamic. To overcome the problem of the infinite number of actions and states in a dynamic environment the authors followed a new definition of state spaces and thus reduced the number of actions and states. This modification on Q-learning provided a solution in a high-speed navigation problem. They designed the states for the environment into four categories: Safe States (SS), Non-Safe States (NS), Winning State (WS), and Failure State (FS). The robot was equipped with three actions such as, move forward, turn left, and turn right. This robot moves forward to find the optimum path and take a turn to the left or right for 45 degrees if it encountered any obstacles. The authors set a reward of 2 points for reaching the WS and a penalty of 2 points for colliding with any obstacles. To assess the method they conducted both simulation experiments and real-life experiments. The simulation experiment was based on the MATLAB programming language. The simulation experiment was divided into a training phase and a test phase. In the training phase, they created 4 different scenarios where the target and the obstacle were moving with different velocities. In all those scenarios, the velocity and the initial position of the robot were kept constant but the obstacle was moving from different locations in each scenario. From these experiments, the researchers measured the time to reach the goal and they also found out that in the second scenario the robot was not able to reach the goal. After that, they started the test phase, for the test phase they created two scenarios, where the 1st test scene was the same as the 2nd scenario of the training phase. However, the robot was susceptible to reach the goal in this screenplay, which means the robot was susceptible to earn from its training phase and utilized it on the test phase. The second test scenario was a complex one with two static and four dynamic obstacles. The robot was able to reach the goal where it took 83s which is the maximum among all the scenarios. In the case of real-life environments, they conducted experiments on multiple soccer robots that were able to avoid obstacles and path planning as well as able to update their positions using the proposed Q- learning algorithm. Both from a simulation and real-life environment they proved that their algorithm was an efficient one.

In [85], Tamosiunaite *et al.* designed a robot to pour liquid from one container to another one using RL based Dynamic Movement Primitives (DMP). In this work, they combined based on both goal learning and shape learning concepts. They utilized a value function based approach for goal learning and for the shape learning they utilized both Natural Actor-Critic (NAC) algorithm and policy improvement with Path Integrals (PI^2) algorithm. To assess their strategy the authors performed a pouring simulation as well as implemented the learning on a 7 Degree of Freedom (DOF) based robotic arm named Mitsubishi Pa10 robot. This work also conducted both simulation and real-life experiments on this problem. For experimenting, they brought the robotic arm to an initial position to execute the pouring process from one filled container to the empty container. They set the mass of the 2nd container as the performance metrics to find out the pouring success. After learning from the human demonstration 8-32 trails were needed to learn the whole pouring process for the robot. They compared both NAC and Path PI^2 algorithms in a simulation process and found that the newly developed PI^2 algorithm was more efficient to implement this task. So, in real-life experiments, they utilized a PI^2 algorithm for shape learning and combined it with value function approximation to execute the whole liquid pouring process. They also arranged a relearning session for the robot by changing the container to be poured. They found that the robot was capable to perform the pouring assignment quite efficiently just after the first learning epoch. From this, they proved that RL can be used as a successful approach for a robot to learn in an unfamiliar environment and they also claimed the PI^2 algorithm is an efficient and fast learning approach for shape learning.

In [98], Hung *et al.* presented flocking among Unmanned Aerial Vehicles (UAVs) by using model-free RL. The experiment demonstrates the flocking techniques of UAVs in a non-stationary stochastic environment using leader-follower topology. To do so, the authors formulated flocking in the form of the MDP. UAVs with small-fixed wings that fly using both average speed and fixed altitude can be defined as leader and follower. The whole experiment is executed in

a simulation environment where the manoeuvre is updated once every second. A 6 DOF aircraft model was used as the kinematics for the UAV. A reward function is set up to facilitate flocking for the follower UAVs while two individual updates of the Q-table composed the Q-learning algorithm of this work. Q-table. The evaluation process is carried out by measuring the cost incurred while flocking a leader in the simulation with 1000 random trajectories. The policies are evaluated throughout the learning process as well as when the policies converge.

In [100], Kathryn Elizabeth Merrick proposes a NN architecture combining with RL to integrate various value systems. The paper addresses some challenges in the corresponding field and evaluates four value systems using the architecture proposed in this paper. Her architecture uses four types of neurons such as sensory neurons, observation and motivation neurons, and activation neurons which are organized in layers. This methodology is then tested on a Lego Mindstorms NXT robot. The author then analyzes the posture, color intensity, identifies cyclic behavior, and calculates the number of observations the robot made. The performance matrices for the robot include posture and point cloud matrices, identifying cyclic behavior, behavioral stability, and exploration. Functional equations have been devised to mathematically measure each matrix. An important finding of this paper states that the experimented robots spent more than 50% of their lifetime exploring rather than exploiting learned behavior cycles

Yu *et al.* contributed highly with their paper [101] by coordinating control of multiple biomimetic robotic fish in an aquatic environment. The whole operation works together as scattered subsystems including information processing and decision-making subsystem which receive information and control commands as input respectively. Then the subsystem sends the processed commands to the robotic fish to execute the commands. The posture of the robotic fish is adjusted by a modified proportional guidance law. To evaluate the performance of robotic fishes, a 2v2 water polo game was taken into account. The robot's position and direction was controlled accurately by a stabilization controller. RL is employed here to attain the block behavior of this robot. The performance was measured by total rewards per trial and total time steps needed to win per trial. Here, fuzzy logic was adopted for discretizing the state and action spaces. The whole experiment was conducted in a physical environment.

B. POLICY BASED ALGORITHM

In [67], Yang *et al.* designed an RL-based adaptive critic controller for discrete-time systems with the help of online approximators and evaluated the performance of the model by conducting a simulation experiment on a two-link robotic arm and a pendulum balancing system. The controller is divided into two networks; one is an action network whose main task is to generate signals and the other is a critic network that is designed for testing the enforcement of the action network through estimating the cost-to-go function. The action network as well as the critic network uses a two-layer Neural Network (NN) to predict the uncommon states so that the model doesn't need a separation principle. For conducting simulation experiments the authors used a Proportional Integral (PI) controller and presented the simulation result of the PI controller as well as the simulation result of the observer state and actual system output for the pendulum balancing system. And in the case of the two-link robotic arm, they presented the simulation result of the actual rotation angle and desired rotation angle of the robotic arm.

In [75], palomeras *et al.* proposed a three layer-based control architecture called Component Oriented Layer-based Architecture for Autonomy (COLA2) for Autonomous Underwater Vehicle (AUV) cable tracking. These different layers of the control architecture are the reactive layer, the execution layer, and the mission layer respectively. An RL technique is used to improve the underwater vehicle's versatility to a changeable environment in the reactive layer. The execution layer models the vehicle's primitive execution flow. And finally, the mission layer depicts the mission phases by utilizing a Mission Control Language (MCL). For evaluating the fruition of the control architecture the authors conducted a simulation experiment based on the Natural Actor-critic (NAC) algorithm by making a scenario of a pool where the task of the robot is to detect a cable. The simulation process was an episodic task where each episode contains 150 iterations. After finishing the simulation process, the obtained policy was implemented on an underwater vehicle called Ietineu AUV for testing a real-life environment inside a water tank. Again an online learning process in this step is also conducted to improve the policy which is obtained from the simulation phase. After 20 trails, the algorithm obtains a suitable policy to find the underwater cable.

In [92], Kober *et al.* proposed a new policy-based RL algorithm named Policy Learning by Weighting Exploration with the Returns (PoWER) which is an EM-inspired RL algorithm to solve high dimensional RL problems in the context of motor primitive learning issues. In the work, the authors introduced their new algorithm and compared with its different policy gradient algorithms like Vanilla' Policy Gradients (VPG), Reward-Weighted Regression (RWR), Finite Difference Gradients (FDG), Episodic Natural Actor Critic (eNAC). To evaluate the performance of their algorithm they conducted both simulation experiment and real-time experiment on Barrett WAM^{TM} robot arm by implementing two different tasks: underactuated Swing-Up, Ball-in-a-Cup game. The robot showed unprecedented performance on both the tasks after only a few episodes and they claimed by means of the obtained result that their algorithm was the efficient one than the other algorithms they took into consideration in their work.

In the paper [95], Kim *et al.* presented a novel learning strategy of motor skill. The paper emphasizes on learning methods based on human-motor control ethos as well as machine learning methods. In the paper, the authors developed a simulator using MSC. ADAMS2005 and

MATLAB/Simulink (Mathworks) was utilized to implement the control algorithm. Then these systems are trained as a human learns various functions to carry out tasks. A planar manipulator was used as the agent of this work and the target of the RL section was to optimize a policy instead of maximizing the rewards. To resolve the problems of continuous state-action pairs, the authors considered the RL System based Episodic Natural Actor Critic (eNAC) algorithm. It works in an actor-critic structure. Opening a door, moving point to point, and catching a flying ball are some of the tasks the proposed framework can enhance contact tasks for.

In [97], Tapia *et al.* presented an RL method for aerial cargo delivery tasks in an environment with motionless obstacles. The experiment was conducted in a simulated environment and physical environment with a quadrotor. The approximate value iteration algorithm was used to produce an approximate solution to the MDP. Trajectories with minimal oscillations are found in the subsequent steps. Once the parameterization value was learned, the trajectories are planned using a greedy policy. An important characteristic of this paper is that; after the learning process, an optimized policy will be generated which is robust to noise. The authors used some performance metrics for simulation experiments such as; collision-free path length, the maximum allowed load displacement, maximum swing and maximum derivation from the path, trajectory waypoints after bisections, number of waypoints. On the other hand, the measurement of performance of the physical experiment is based on the calculation of derivation the quadrotor makes from the simulated result. Another top-notch characteristic of this paper is it presented AI in very large action space.

Lian *et al.* [99] presented a control technique for Wheeled Motor Robots (WMRs) based on receding-horizon dual heuristic programming (RHDHP) in their acknowledged paper [99]. In the experiment presented, a backstepping kinematic controller was designed to generate the desired velocity profile. Infinite-horizontal control problems were decomposed to a finite-horizontal control problem by the receding horizon strategy. The experiment was carried out in a simulated environment using a mobile robot with differential driven wheels. The mobile robot was studied for an eight-shaped trajectory and an ellipse trajectory. The performance is measured by tracking accuracy and computational burden. The proposed robot successfully outperforms its predecessor under different predictions and control horizons.

C. MODEL FREE ALGORITHM

In [83], Koos *et al.* proposed a new algorithm called T-resilience for damage recovery robotics. They implemented their algorithm on a hexapod robot with an onboard RGB-D sensor and Simultaneous Localization and Mapping (SLAM) algorithm where they trained the robot using only 25 tests which took a total of 20 minutes running time. They experimented with the robot using six different setups. After that, they compare their algorithm with the policy gradient and stochastic policy algorithm, and each time they proved with different data that their algorithm is an effective and fast approach.

In [91], Sehnke *et al.* considered the problem of higher variance gradient on various policy-based RL algorithms. To win out this dilemma the authors introduced a new model-free policy gradient based RL algorithm named Policy gradients with parameter-based exploration (PGPE). After deriving this algorithm they implemented it on several projects and compared the effectiveness of this algorithm to other policy-based algorithms like Stochastic Adaptation (SPSA), REINFORCE, evolution strategies (ES), and episodic natural actor-critic (eNAC). First of all, they implemented their algorithm on a pole balancing task. From experiments, they showed that PGPE with SYS was an effective and fast algorithm for this task. After that the implemented it on a flexcube walking task; a Jordan network with 32 inputs, 10 hidden units and 12 output units were used as the controller of this task. Results showed that among the algorithms PPGE was the fast learning algorithm and was best in case of a final reward. The next simulation experiment they conducted on the open dynamics engine simulator based on a biped robot named Johnnie. The robot was expected to stand on its legs despite being perturbed by some kind of external force. The controller of the robot was a Jordan network having 41 inputs, 20 hidden units, and 11 output units. The experimental result showed that initially the REINFORCE algorithm was the fastest in learning but after 500 training episodes PPGE surpassed it. The fourth experiment was based on a sheep steering task. The authors simulated a scenario of a ship navigating at maximum speed. By results, they showed that the PPGE algorithm was the best among all the algorithms for this task. The last Simulation experiment they conducted on the Consumers Co-operative Refineries Ltd. (CCRL) robot which was also implemented on the open dynamics engine simulator. The objective of this experiment was to grab an object from a different position of a table. They conducted the experiments on four different phases (the object was at the edge of the table, the object was quite away from the edge, the object was at the center of the table, several objects were distributed around the surface of the table. All the phases proved the PPGE algorithm was the most effective one. After obtaining these results they claimed that their method led to lower variance gradient than other algorithms and outperformed the algorithms in case of several robotics problems.

D. OTHERS

In [62], Kim *et al.* developed a robotic wheelchair by utilizing Inverse Reinforcement Learning (IRL) which can reach a predefined destination through the navigation of a human crowded environment in a socially adaptive manner. The characteristic of the robot allows it to interact with pedestrians in different dynamic environments like market, shopping mall, etc. daily. To implement this work, the authors divided their working methodology into three sections, such as Feature extraction module, IRL module, and Path-planning module. This robot navigates the environment with the help of

TABLE 9. Summary of the answers of the questions.

Ref	Q1	Q2			Q3	Q4		Q5					Q6
		Value-based	Policy based	Model-free		S	R	UGV	UUV	UAV	R.A	HR	
[62]	IRL	-	-	-	✗	✗	✓	✓	✗	✗	✗	✗	Path planning in crowded environment
[67]	Actor-critic	✗	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	Manipulate the robot arm to track sine waves.
[75]	NAC	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	Cable tracking
[76]	Modified Q-learning	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	Self-localization and autonomous navigation.
[77]	Improved Q-learning	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	Path planning.
[79]	Q-learning, SARSA	✓	✗	✗	✗	✓	✓	✗	✗	✗	✓	✗	To catch a ball like a goalkeeper.
[80]	Double DQN	✓	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	Obstacle avoiding and grasping object.
[81]	Q-learning	✓	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	Path planning and playing soccer.
[83]	T-resilience	✗	✗	✓	✓	✓	✓	✓	✗	✗	✗	✗	Damage recovery.
[85]	CQL, NAC, PI^2	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✗	Pouring liquid from one glass to another.
[90]	IRL	-	-	-	✗	✗	✓	✓	✗	✗	✗	✗	Autonomous navigation in complex unstructurred environment.
[91]	PGPE		✗	✓	✗	✓	✓	✗	✗	✗	✗	✓	Biped robot standing task, Grasping task.
[92]	PoWER	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓	✗	Playing Ball-in-a-cup game.
[95]	eNAC	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗	Opening a door and Catching a flying ball.
[96]	Hierarchical RL	-	-	-	✗	✓	✓	✓	✗	✗	✗	✗	Rescue victims.
[97]	Monte Carlo methods	✗	✓	✗	✗	✓	✓	✗	✗	✓	✗	✗	Cargo delivery
[98]	Q-flocking	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	Learning how to flock in a leader follower topology.
[99]	Actor-critic	✗	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗	Tracking control.
[100]	Q-learning	✓	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	Self-motivated exploration.
[101]	Fuzzy RL, Q-Learning	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	Coordination control by playing 2 Vs 2 game.

feature extraction by calculating a few features like densities and velocities of humans using an RGB-Depth sensor. The focus of the IRL module is to learn social manners as a cost function while navigating in a crowded place from a human demonstrator. A human expert plans his action sequences in a socially adaptive way and the robot or AI agent learns it offline using IRL to navigate. And the last module called the path planning module consists of three different layers: 1. Global path planning, 2. Local path planning, 3. Obstacle avoidance. Global path planning is used to reach the final goal using a priori map. In contrast, Local path planning is used to reach a sub-goal. An obstacle avoidance layer is made of some hand-crafted rules to avoid collision. The overall framework is built on the Robot Operating System (ROS) and finally deployed on a real-life robotic wheelchair. After that they evaluate the performance of their robot by comparing it with a human demonstration on different scenarios in their lab-like: Pedestrian walking towards the robot, a pedestrian walking horizontally to the robot, multiple pedestrians. They also allowed a human sitting on the chair to measure the human intervention percentage in critical situations. The experiments show that their algorithm was quite close to the human demonstration.

In [90], Silver *et al.* considered the problem of coupling path planning and perception tasks for mobile robotics systems. They claimed that the performance of a mobile robot in an unstructured complex environment depends not only on the individual performance of planning and perception tasks but also on the synchronization of these tasks on the navigation of a mobile robot. However, they presented a solution to their work for this problem on a Crusher autonomous navigation platform by using learning from human demonstration. For interpreting the data of demonstration learning they utilized an algorithm named LEArning to seaRCH (LEARCH algorithm). This robot learned from both online perception (satellite image, lidar) and onboard perception setting. Several experiments were conducted to teach the cost function and they set the average path loss and ratio of two different cost functions as the performance metrics of the robot. After conducting some experiences they found that this approach reduces training and testing time as well as produces efficient and robust systems for mobile robots.

In [96], Doroodgar *et al.* presented an RL-based control architecture that can be utilized on rescue robots. The aim includes learning from the robot's own experience and improve its overall performance. Both physical and

simulation experiments were conducted for this paper. Simulation setup includes 20 by 20 cell environments which constitute approximately 336 square meters in the physical world area. physical experiments were conducted in a cluttered 12 square meter Urban Search and Rescue (USAR) like environment which mimics a disaster scene. The experiment includes a rescue robot with a real-time 3D mapping sensor, a thermal camera, and an infrared sensor. The semi-autonomous control architecture which was presented by the authors in this work was based on Hierarchical Reinforcement Learning (HRL) to fast-track rescue operations. The HRL method used in this experiment is called MAXQ. SLAM module is then to make a 3D model of the environment. MDP is an integral part of the MAXQ technique which is used in the learning method. The authors set the number of victims found by the robot and exploring the ability of the environment by avoiding obstacles as the performance metrics for their work.

After analyzing the papers we have raised some questions

- Q1. Which algorithm is used in an article?
- Q2. What kind of algorithm that is?
- Q3. Do the authors utilized Neural Networks in their paper?
- Q4. Whether they conduct experiments on simulation experiments or real-life environments or both?
- Q5. What types of robot was utilized in a research article?
- Q6. What was/were the main task(s) of the robot(s)?

and summarizes the findings in Table 9. In the table, NN→Neural Network, S→Simulation, R→Real, UGV→ Unmanned Ground Vehicle, UUA→Unmanned Underwater vehicle, UAV→ Unmanned Aerial vehicle, RA→Robotic Arm, HR→Humanoid Robot, ✓→ the feature is present in any particular article, ✗→ the feature is absent in any particular article. From the table, it can be seen that among the 20 papers 9 papers [76], [77], [79]–[81], [98], [100], [101] utilized value-based algorithm, 7 papers [67], [75], [85], [92], [95], [97], [99] utilized policy-based algorithms, 2 papers [83], [91] used model-based learning, [62] and [90] used Inverse Reinforcement Learning (IRL) which learns from human demonstration. Only 25 percent papers [67], [80], [83], [99], [100] used a Neural Network architecture on their work. Nine papers [79]–[81], [83], [85], [91], [92], [96], [97] conducted both simulation and real life experiments to measure the robustness of their works. [62], [76], [77], [80], [81], [83], [90], [96], [99], [100], 6 papers [77], [79], [80], [85], [92], [95] utilized robotic arms and the rest of the papers utilized UUV [75], [101], UAV [97], [98], Humanoid robot [91].

VI. CONCLUSION

In this paper, we have presented a systematic review of the existing literature on Reinforcement Learning-based robotics. RL has become a tantalizing part of robotics research. From making robotics learning swifter and more precise to gain autonomous superiority to help human accessibility, RL is an integral part of modern robotics trend. This field has seen a successful spike in the development curve of robotics applications. This systematic review paper presented a bibliographic analysis of the existing literature within the last decade to figure out the research trend in this domain. Furthermore, the paper reconnoiter possible future research directions by finding out the influential research terms in this domain so that the succeeding researchers can explore in this research concept. Additionally, this paper showcases a thorough survey on RL-based robotics papers as well as it also summarizes many useful insights from a few of the scrutinized papers. Some fascinating research questions over and above have been generated as a sequel to the major findings from the reviewed papers. The paper also provides eloquent answers to those questions for checking the viability and robustness of any particular paper in this research area. Following the preceding history and application of RL algorithms on robotics, the authors of this paper convene on one point that Deep Reinforcement Learning(DRL) is going to be an enthralling research trend in this domain. In terms of practical application, hands-on effectiveness, rapid learning environment, and feasible outcome, DRL can spearhead the RL practice in robotics for the days to come. It will be an elegant step for researchers to extend their research to DRL based mobile robots in their further research persuasion.

APPENDIX

Model-free RL are of two types: (a) Value-based RL, (b) Policy-based RL

A. VALUE-BASED RL

In a value-based RL, an agent makes its decision based on the value function $V(s)$. The value function is the representation of expected maximum reward r_t which will be collected using a fixed stochastic policy π by an agent in a given state s_t, such as

$$V_\pi(s) = \mathbb{E}_\pi[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots\ldots + \gamma r_{t+n} | s_t = S]. \quad (1)$$

where γ lies in 0 and 1 represents the discount factor for future awards. Future rewards are less important for an agent so, they should be discounted. The system will calculate the value functions for every action $a_t \in A$ at a given state $s_t \in S$. Then the agent will select the (s_t, a_t) pair having the biggest value to reach the next step s_{t+1} of the environment and the process will continue until the goal is attained. Some well-known value-based algorithms are as follows: Q-learning, State-Action-Reward-State-Action (SARSA), Deep Q-Network (DQN), Categorical DQN (C51), Double DQN, Dueling DQN, etc. In this article, we have described Q-learning, SARSA, C51, DQN.

1) Q-LEARNING

Q-learning [56] is an off-policy (Learning the value of optimal policy does not depend on the agent's action) value-based

algorithm. In a Q-learning algorithm, a table called Q-table is generated by using the available states (s_t) and actions (a_t) of an environment. Let the environment has m states (s_t) and n actions (a_t), then a table of $m \times n$ size will be generated where each cell of the table will be equipped with a value (i.e Q value) which is represented by a function called Q-function or quality function denoted by $Q(s_t, a_t)$. The agent will take actions by choosing the biggest Q-value from available (s_t, a_t) pairs of any given state (s_t). At the incipiency, of the learning procedure, all the Q-values are inducted with zero, then the learning rate α is high and the agent takes random states (exploration) using the ϵ-greedy policy with a probability of ϵ to explore the environment. As the agent explores the environment with time; the learning rate decreased and the agent starts exploitation and choose action with a probability 1-ϵ by using the updated Q-values. The Q-values get updated through an iterative process using the Bellman equation [58]. If we consider any particular state s_t, then the updated Q-value for any Q-function $Q(s_t,a_t)$ will be

$$Q'(s_t, a_t) = Q(s_t, a_t) + \alpha[r_t + \gamma(maxQ'(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]. \quad (2)$$

The agent becomes more confident about selecting actions by utilizing the updated Q-value and collect more rewards with fewer hurdles. The updating process continues until the learning is stopped and after a successful learning session an updated Q table is generated and the agent utilizes it to choose an optimum control policy.

2) SARSA

SARSA is an on-policy (Learning the value of the optimal policy depends on the agent action) value-based algorithm which was proposed by Rummery and Nirajan in [51] is a modified conception of Q-learning. Akin to Q-learning; the SARSA algorithm does not use necessarily the maximum Q-value from the available (s_t, a_t) pairs of the next state s_{t+1} instead, it uses the same policy to choose the next action which was determined by the original action. So the Q-value update equation of SARSA can be represented as

$$Q'(s_t, a_t) = Q(s_t, a_t) + \alpha[r_t + \gamma(Q'(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]. \quad (3)$$

The updating process continues until the learning is stopped and after a successful learning session an update Q table is generated and the agent utilizes it to navigate on the environment.

3) DQN

The Deep Q-Network or DQN is an off-policy value-based reinforcement algorithm which is a combination of DNN and traditional Q-learning proposed by Milh *et al.* in [52]. To suppress the challenge of having a huge knowledge space or storing a large amount of (state, action) pairs in a Q-table; DQN is introduced to approximate the Q-value using a Neural Network by replacing the Q-table. In a DQN algorithm states (s_t) of the environment are given as input of the network and the DQN predict the Q-values at the output node for all possible actions $(a_t, a_{t+1}\ldots a_{t+n})$ of the given state (s_t). In case of Q-function updating DQN algorithm uses a Neural Network function approximator $Q(s_t,a_t;\theta)$ with weights θ and tries to minimize the square error loss function $L_i(\theta_i)$ (difference between targeted Q-value and predicted Q-value) in each iteration in lieu of using the Bellman equation. The loss function can be defined as

$$L_i(\theta_i) = \mathbb{E}[(x_i - Q(s_t, a_t; \theta_i))^2]$$
$$x_i = \mathbb{E}[r_t + \gamma(maxQ'(s_{t+1}, a_{t+1}; \theta_i^-)]. \quad (4)$$

Now this loss function is being differentiated with respect to θ for optimizing the loss function using Stochastic Gradient Descent (SGD) to decrease the error meaning that the current policy's output or predicted output becomes similar to the targeted output and the following gradient arrives

$$\nabla L_i(\theta_i) = \mathbb{E}[r_t + \gamma(maxQ'(s_{t+1}, a_{t+1}; \theta_i^-) - Q(s_t, a_t; \theta_i))\nabla_{\theta_i} Q(s_t, a_t; \theta_i)]. \quad (5)$$

Thus using this equation, loss function get minimized and Q-values get updated in each iteration. The updating process continue until the learning is stopped and after a successful learning session the agent can navigate on the environment in an optimum way using the updated Q-values.

4) C51

C51 is an off-policy value-based reinforcement algorithm was first presented in [53]. The key idea behind the C51 algorithm is to use the distribution of future rewards instead of expected future rewards. As a result in this algorithm quality function $Q(s_t,a_t)$ is replaced by a distribution function $Z(s_t,a_t)$ which contains value distribution Z correspondence to Q-value and as a consequence an equation called ''Distribution Bellman Equation'' arises which is equivalent to Bellman equation. The distributional conception of Bellman equation is as follows:

$$Z(s_t, a_t) = r_t + \gamma\ max\ Z(s_{t+1}, a_{t+1}). \quad (6)$$

Now, in case of update the distributional values target distribution Z_{tar} is calculated by scaling the next distribution Z_{t+1} by the discount factor γ and by shifting using the reward r_t. After that, the algorithm tries to minimize the cross-entropy loss function $C.L_i(\theta_i)$ between the $Z(s_{tar}, a_{tar})$ and $Z(s_t, a_t)$ using the weights (θ) of the deep network in each iteration to minimize the difference between targeted value distribution and current value distribution. The loss function can be defined as

$$C.L_i(\theta_i) = (y_i - Z(s_t, a_t; \theta_i))$$
$$y_i = r_t + \gamma maxZ(s_{t+1}, a_{t+1}; \theta_i^-). \quad (7)$$

Now this cross-entropy loss function is being differentiated w.r.t θ to optimize the loss function using SGD to decrease the error means that the difference between the targeted value

distribution and current value distribution becomes similar and the following gradient arrives:

$$\nabla C.Loss_i(\theta_i) = r_t + \gamma maxZ(s_{t+1}, a_{t+1}; \theta_i^-) - Z(s_t, a_t; \theta_i))\nabla_{\theta_i} Z(s_t, a_t; \theta_i)] \quad (8)$$

Thus using this equation the loss function get minimized and distribution values get updated in each iteration. The updating process continues until the learning is stopped and after a successful learning session, the agent can navigate on the environment in an optimum way using the updated distribution values.

B. POLICY-BASED RL

Policy-based RL is another type of model-free RL algorithm. The main objective of a policy-based RL is to improve a policy function π(s) directly without using the value function *V(s)*. The policy π(s) selects the best action (a_t) which should be considered in a particular state (s_t) to increase the reward without calculating the value function. The policy function π(s) can be defined as the probability of selecting an action $a_t \in A$ at a given state $s_t \in S$ and a parameterized vector ζ such as

$$\pi(A|S, \zeta) = P_r(a_t = A|s_t = S, \zeta_t = \zeta) \quad (9)$$

Now, for calculating the performance of the policy a score function J(ζ) is introduced which can be defined as

$$\mathrm{J}(\zeta) = \mathbb{E}_\pi \zeta[\sum \gamma r]. \quad (10)$$

After that, an appropriate parameter (ζ *) will maximize the expected reward using this policy value such as

$$\zeta^* = \arg\max \ \mathrm{J}(\zeta) \quad (11)$$

According to Duan *et al.* [61] some prominent algorithms in robotics are DDPG, TRPO, PPO, etc. So we have discussed only these of the policy-based RL algorithms in our paper.

1) DEEP DETERMINISTIC POLICY GRADIENT (DDPG)

DDPG [59] is a policy-based RL algorithm which simultaneously calculate value-function using Deep Q-learning (DQN) and optimizes policy. It is difficult for a Q-function to select the best action in a continuous action space; so Lillicrap *et al.* presented an algorithm which is based on the Deterministic Policy Gradient (DPG) [60] combining with DQN named DDPG in [59]. The value-function based learning of this algorithm is just like a DQN structure. Now in the case of policy optimization, the authors used the parametrized action function $\mu(s_t|\zeta^\mu)$ and update the policy function with batches using the following policy gradient:

$$\nabla_\zeta \mathcal{J}_\zeta = \frac{1}{M}\sum_i \nabla_a Q(s_t, a_t|\zeta^Q)|_{s_t=s_i, a_t=\mu(s_i)} \nabla_{\zeta\mu}\mu(s_t|\zeta^\mu)|_{s_i} \quad (12)$$

where M is the batch size. The authors also solved the problem of exploration in continuous action space by constructing an exploration policy μ' by adding an independent noise N such as

$$\mu'(s_t) = \mu(s_t|\zeta_t^\mu) + N \quad (13)$$

2) TRUST REGION POLICY OPTIMIZATION (TRPO)

TRPO [54] is a policy-based on policy RL algorithm for optimizing the parameters of a policy. The task of this algorithm is to change/update the previously used policy $\pi_{\zeta_{prev}}$ into a new policy π_ζ at each learning update by solving the optimization problem described in [54] which can be denoted by the following equation

$$\begin{aligned} &\max_\zeta \ \mathbb{E}_{s_t,a_t}\pi_{\zeta_{prev}}[\pi_\zeta(a_t, s_t)A_{\zeta_{prev}}(s_t, a_t)] \\ &s.t \quad \mathbb{E}_{s_t,a_t}\pi_{\zeta_{prev}}[D_{KL}(\pi_\zeta(.|s_t)||\pi_{\zeta_{prev}}(.|s_t))] \ \leq \delta \end{aligned} \quad (14)$$

where, $A_{\zeta_{prev}}$ denotes the advantage function, the ratio between the targeted policy and previous policy $\frac{\pi_\zeta(a_t|s_t)}{\pi_{\zeta_{prev}}(a_t|s_t)}$ is represented by $r_\zeta(a_t|s_t)$, average KL Divergence is denoted by D_KL and δ is used to represent the step-size parameter. Conjugate-gradient algorithm is utilized by TRPO to solve this optimization problem.

3) PROXIMAL POLICY OPTIMIZATION (PPO)

PPO is another policy-based on policy RL algorithm which is a modified version of TRPO was first introduced by Schulman *et al.* in 2016 [55]. To overcome the issue of high computational complexity in case of solving the optimization problem the surrogate objective function KL divergence of TRPO is substituted by a clipped surrogate objective function L_ζ^{clip} containing a penalty term σ for large policy updates

$$L_\zeta^{clip} = \mathbb{E}_{s_t,a_t}\pi_{\zeta_{prev}}[min(r_\zeta(a_t|s_t)\, A_{\zeta_{prev}}(a_t, s_t), \ clip(r_\zeta(a_t|s_t), 1-\sigma, 1+\sigma)\, A_{\zeta_{prev}}(a_t, s_t)] \quad (15)$$

Whenever the probability ratio $r_\zeta(s_t,a_t)$ causes the objective function tends to extent, the computational complexity increases and then the probability ration is clipped in the range of $[1-\sigma, 1+\sigma]$ to decrease the high computational complexity.

REFERENCES

[1] K. Hitomi, "Automation—Its concept and a short history," *Technovation*, vol. 14, no. 2, pp. 121–128, 1994.
[2] A. M. Turing, "Mind a quarterly review of psychology and philosophy," *Comput. Machinery Intell.*, vol. 236, pp. 433–460, Oct. 1950.
[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[4] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Noida, India: Pearson, 2018.
[5] D. L. Wang, "Unsupervised learning: Foundations of neural computation a review," *AI Mag.*, vol. 22, no. 2, 2001.
[6] R. S. Sutton, "Introduction: The challenge of reinforcement learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 225–227, May 1992.
[7] S. Mahadevan and J. Connell, "Automatic programming of behavior-based robots using reinforcement learning," *Artif. Intell.*, vol. 55, nos. 2–3, pp. 311–365, Jun. 1992.
[8] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda, "Vision-based reinforcement learning for purposive behavior acquisition," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 1995, pp. 146–153.

[9] J. Peters, K. Mulling, and Y. Altun, "Relative Entropy Policy Search," in *Proc. 24th AAAI Conf. Artif. Intell. (AAAI)*, 2010, pp. 1607–1612.
[10] M. Kalakrishnan, L. Righetti, P. Pastor, and S. Schaal, "Learning force control policies for compliant manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 4639–4644.
[11] W. C. Anderson, K. Carey, E. M. Sturzinger, and C. J. Lowrance, "Autonomous navigation via a deep Q network with one-hot image encoding," in *Proc. IEEE Int. Symp. Meas. Control Robot. (ISMCR)*, Sep. 2019, pp. A2-2-1–A2-2-6.
[12] Y. Kato, K. Kamiyama, and K. Morioka, "Autonomous robot navigation system with learning based on deep Q-network and topological maps," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Dec. 2017, pp. 1040–1046.
[13] M. T. Rosenstein, A. G. Barto, and R. E. A. Van Emmerik, "Learning at the level of synergies for a robot weightlifter," *Robot. Auto. Syst.*, vol. 54, no. 8, pp. 706–717, Aug. 2006.
[14] S. Mahadevan and G. Theocharous, "Optimizing production manufacturing using reinforcement learning," in *Proc. 11th Int. FLAIRS Conf.*, 1998, p. 377.
[15] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," Oct. 2016, *arXiv:1610.03295*. [Online]. Available: http://arxiv.org/abs/1610.03295
[16] X. Pan, Y. You, Z. Wang, and C. Lu, "Virtual to real reinforcement learning for autonomous driving," in *Proc. Brit. Mach. Vis. Conf.*, 2017, doi: 10.5244/c.31.11.
[17] P. Kormushev, S. Calinon, R. Saegusa, and G. Metta, "Learning the skill of archery by a humanoid robot iCub," in *Proc. 10th IEEE-RAS Int. Conf. Humanoid Robots*, Dec. 2010, pp. 417–423.
[18] A. Carrio, C. Sampedro, A. Rodriguez-Ramos, and P. Campoy, "A review of deep learning methods and applications for unmanned aerial vehicles," *J. Sensors*, vol. 2017, pp. 1–13, Aug. 2017.
[19] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auto. Syst.*, vol. 57, no. 5, pp. 469–483, May 2009.
[20] L. Busoniu, R. Babuska, and B. De Schutter, "Multi-agent reinforcement learning: A survey," in *Proc. 9th Int. Conf. Control, Autom., Robot. Vis.*, 2006, pp. 1–6.
[21] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015.
[22] E. Yang and D. Gu. (Jun. 2014). *Multiagent Reinforcement Learning for Multi-Robot Systems: A Survey*. [Online]. Available: https://www.researchgate.net/publication/2948830
[23] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.
[24] T. M. Moerland, J. Broekens, and C. M. Jonker, "Emotion in reinforcement learning agents and robots: A survey," *Mach. Learn.*, vol. 107, no. 2, pp. 443–480, Feb. 2018.
[25] S. Caldera, A. Rassau, and D. Chai, "Review of deep learning methods in robotic grasp detection," *Multimodal Technol. Interact.*, vol. 2, no. 3, p. 57, Jul. 2018.
[26] N. D. Nguyen, T. Nguyen, and S. Nahavandi, "System design perspective for human-level agents using deep reinforcement learning: A survey," *IEEE Access*, vol. 5, pp. 27091–27102, 2017.
[27] D. Nguyen-Tuong and J. Peters, "Model learning for robot control: A survey," *Cognit. Process.*, vol. 12, no. 4, pp. 319–340, Nov. 2011.
[28] S. Wang, W. Chaovalitwongse, and R. Babuska, "Machine learning algorithms in bipedal robot control," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 5, pp. 728–743, Sep. 2012.
[29] Z. Bing, C. Meschede, F. Röhrbein, K. Huang, and A. C. Knoll, "A survey of robotics control based on learning-inspired spiking neural networks," *Frontiers Neurorobot.*, vol. 12, p. 35, Jun. 2018.
[30] A. R. Mahmood, D. Korenkevych, G. Vasan, W. Ma, and J. Bergstra, "Benchmarking reinforcement learning algorithms on real-world robots," Sep. 2018, *arXiv:1809.07731*. [Online]. Available: http://arxiv.org/abs/1809.07731
[31] L. Khriji, F. Touati, K. Benhmed, and A. Al-Yahmedi, "Mobile robot navigation based on Q-learning technique," *Int. J. Adv. Robotic Syst.*, vol. 8, no. 1, p. 4, Mar. 2011.
[32] Y. Liu, H. Liu, and B. Wang, "Autonomous exploration for mobile robot using Q-learning," in *Proc. 2nd Int. Conf. Adv. Robot. Mechatronics (ICARM)*, Aug. 2017, pp. 614–619.
[33] D. Ramachandran and R. Gupta, "Smoothed sarsa: Reinforcement learning for robot delivery tasks," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 2125–2132.
[34] F. Tavakoli, V. Derhami, and A. Kamalinejad, "Control of humanoid robot walking by fuzzy Sarsa learning," in *Proc. 3rd RSI Int. Conf. Robot. Mechatronics (ICROM)*, Oct. 2015, pp. 234–239.
[35] M. M. Rahman, S. M. H. Rashid, and M. M. Hossain, "Implementation of q learning and deep q network for controlling a self balancing robot model," *Robot. Biomimetics*, vol. 5, no. 1, Dec. 2018, doi: 10.1186/s40638-018-0091-9.
[36] L. Tai and M. Liu, "A robot exploration strategy based on Q-learning network," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot. (RCAR)*, Jun. 2016, pp. 57–62.
[37] W. Zhang, J. Gai, Z. Zhang, L. Tang, Q. Liao, and Y. Ding, "Double-DQN based path smoothing and tracking control method for robotic vehicle navigation," *Comput. Electron. Agricult.*, vol. 166, Nov. 2019, Art. no. 104985.
[38] X. Xue, Z. Li, D. Zhang, and Y. Yan, "A deep reinforcement learning method for mobile robot collision avoidance based on double DQN," in *Proc. IEEE 28th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2019, pp. 2131–2136.
[39] R. Özaln, C. Kaymak, O. Yildirum, A. Ucar, Y. Demir, and C. Guzelis, "An implementation of vision based deep reinforcement learning for humanoid robot locomotion," in *Proc. IEEE Int. Symp. Innov. Intell. Syst. Appl. (INISTA)*, Jul. 2019, pp. 1–5.
[40] X. Qiu, K. Wan, and F. Li, "Autonomous robot navigation in dynamic environment using deep reinforcement learning," in *Proc. IEEE 2nd Int. Conf. Autom., Electron. Electr. Eng. (AUTEEE)*, Nov. 2019, pp. 338–342.
[41] M. A. Fahami, M. Roshanzamir, and N. H. Izadi, "A reinforcement learning approach to score goals in RoboCup 3D soccer simulation for NAO humanoid robot," in *Proc. 7th Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2017, pp. 450–454.
[42] Y. Choi, K. Lee, and S. Oh, "Distributional deep reinforcement learning with a mixture of Gaussians," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9791–9797.
[43] T. Hester, M. Quinlan, and P. Stone, "Generalized model learning for reinforcement learning on a humanoid robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 2369–2374.
[44] J. Kulhanek, E. Derner, T. de Bruin, and R. Babuska, "Vision-based navigation using deep reinforcement learning," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Sep. 2019, pp. 1–8.
[45] C. Guo, W. Luk, S. Q. S. Loh, A. Warren, and J. Levine, "Customisable control policy learning for robotics," in *Proc. IEEE 30th Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, Jul. 2019, pp. 91–98.
[46] K. A. A. Mustafa, N. Botteghi, B. Sirmacek, M. Poel, and S. Stramigioli, "Towards continuous control for mobile robot navigation: A reinforcement learning and slam based approach," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W13, no. 2/W13, pp. 857–863, 2019.
[47] S. Shao, J. Tsai, M. Mysior, W. Luk, T. Chau, A. Warren, and B. Jeppesen, "Towards hardware accelerated reinforcement learning for application-specific robotic control," in *Proc. IEEE 29th Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, Jul. 2018, pp. 1–8.
[48] A. R. Dooraki and D.-J. Lee, "Multi-rotor robot learning to fly in a bio-inspired way using reinforcement learning," in *Proc. 16th Int. Conf. Ubiquitous Robots (UR)*, Jun. 2019, pp. 118–123.
[49] M. Totani, N. Sato, and Y. Morita, "Step climbing method for crawler type rescue robot using reinforcement learning with proximal policy optimization," in *Proc. 12th Int. Workshop Robot Motion Control (RoMoCo)*, Jul. 2019, pp. 154–159.
[50] B. Qin, Y. Gao, and Y. Bai, "Sim-to-real: Six-legged robot control with deep reinforcement learning and curriculum learning," in *Proc. 4th Int. Conf. Robot. Autom. Eng. (ICRAE)*, Nov. 2019, pp. 1–5.
[51] G. A. Rummery and M. Niranjan, "On-line Q-learning using connectionist systems," *CUED/F-INFENG/TR*, vol. 166, Sep. 1994.
[52] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," Dec. 2013, *arXiv:1312.5602*. [Online]. Available: http://arxiv.org/abs/1312.5602
[53] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 449–458.
[54] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," Apr. 2015, *arXiv:1502.05477*. [Online]. Available: http://arxiv.org/abs/1502.05477
[55] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," Aug. 2017, *arXiv:1707.06347v2 [cs.LG]*. [Online]. Available: https://arxiv.org/abs/1707.06347

[56] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[57] N. J. van Eck and L. Waltman, "VOSviewer manual," Univ. Leiden, Leiden, South Holland, Tech. Rep., 2020.
[58] R. Bellman, *The Theory of Dynamic Programming*. Santa Monica, CA, USA: RAND Corporation, 1954.
[59] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," Jul. 2015, *arXiv:1509.02971*. [Online]. Available: http://arxiv.org/abs/1509.02971
[60] D. Silver, G. Lever, N. H. N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, 2014, pp. 387–395.
[61] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," May 2016, *arXiv:1604.06778*. [Online]. Available: http://arxiv.org/abs/1604.06778
[62] B. Kim and J. Pineau, "Socially adaptive path planning in human environments using inverse reinforcement learning," *Int. J. Social Robot.*, vol. 8, no. 1, pp. 51–66, Jan. 2016.
[63] M. P. Deisenroth, D. Fox, and C. E. Rasmussen, "Gaussian processes for data-efficient learning in robotics and control," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 408–423, Feb. 2015.
[64] J. Tessadori, M. Bisio, S. Martinoia, and M. Chiappalone, "Modular neuronal assemblies embodied in a closed-loop environment: Toward future integration of brains and machines," *Frontiers Neural Circuits*, vol. 6, Dec. 2012, doi: 10.3389/fncir.2012.00099.
[65] A. Borji, M. N. Ahmadabadi, B. N. Araabi, and M. Hamidi, "Online learning of task-driven object-based visual attention control," *Image Vis. Comput.*, vol. 28, no. 7, pp. 1130–1145, Jul. 2010.
[66] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: Computational and neural mechanisms," *Trends Cognit. Sci.*, vol. 17, no. 11, pp. 585–593, Nov. 2013.
[67] Q. Yang and S. Jagannathan, "Online reinforcement learning-based neural network controller design for affine nonlinear discrete-time systems," in *Proc. Amer. Control Conf.*, Jul. 2007, pp. 4774–4779.
[68] G. Naros, I. Naros, F. Grimm, U. Ziemann, and A. Gharabaghi, "Reinforcement learning of self-regulated sensorimotor β-oscillations improves motor performance," *NeuroImage*, vol. 134, pp. 142–152, Jul. 2016.
[69] D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers, "Evolutionary dynamics of multi-agent learning: A survey," *J. Artif. Intell. Res.*, vol. 53, pp. 659–697, Aug. 2015.
[70] S. Schall and C. G. Atkenson, "Learning control in robotics trajectory-based optimal control techniques," *IEEE Robot. Automat. Mag.*, vol. 17, no. 2, pp. 20–29, 2010.
[71] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, and C. Melhuish, "Reinforcement learning and optimal adaptive control: An overview and implementation examples," *Annu. Rev. Control*, vol. 36, no. 1, pp. 42–59, Apr. 2012.
[72] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.
[73] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
[74] E. A. Pohlmeyer, B. Mahmoudi, S. Geng, N. W. Prins, and J. C. Sanchez, "Using reinforcement learning to provide stable brain-machine interface control despite neural input reorganization," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e87253.
[75] N. Palomeras, A. El-Fakdi, M. Carreras, and P. Ridao, "COLA2: A control architecture for AUVs," *IEEE J. Ocean. Eng.*, vol. 37, no. 4, pp. 695–716, Oct. 2012.
[76] K. Caluwaerts, M. Staffa, S. N'Guyen, C. Grand, L. Dollé, A. Favre-Félix, B. Girard, and M. Khamassi, "A biologically inspired meta-control navigation system for the Psikharpax rat robot," *Bioinspiration Biomimetics*, vol. 7, no. 2, Jun. 2012, Art. no. 025009.
[77] A. Konar, I. Goswami Chakraborty, S. Jitu Singh, L. C. Jain, and A. K. Nagar, "A deterministic improved Q-learning for path planning of a mobile robot," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 43, no. 5, pp. 1141–1153, Sep. 2013.
[78] G. Baldassarre, F. Mannella, V. G. Fiore, P. Redgrave, K. Gurney, and M. Mirolli, "Intrinsically motivated action–outcome learning and goal-based action recall: A system-level bio-constrained computational model," *Neural Netw.*, vol. 41, pp. 168–187, May 2013.
[79] S. Adam, L. Busoniu, and R. Babuska, "Experience replay for real-time reinforcement learning control," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 2, pp. 201–212, Mar. 2012.
[80] M. Duguleana, F. G. Barbuceanu, A. Teirelbar, and G. Mogan, "Obstacle avoidance of redundant manipulators using neural networks based reinforcement learning," *Robot. Comput.-Integr. Manuf.*, vol. 28, no. 2, pp. 132–146, Sep. 2011.
[81] M. A. K. Jaradat, M. Al-Rousan, and L. Quadan, "Reinforcement based mobile robot navigation in dynamic environment," *Robot. Comput.-Integr. Manuf.*, vol. 27, no. 1, pp. 135–149, Feb. 2011.
[82] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," *Int. J. Robot. Res.*, vol. 32, no. 7, pp. 790–805, Jun. 2013.
[83] S. Koos, A. Cully, and J.-B. Mouret, "Fast damage recovery in robotics with the T-resilience algorithm," *Int. J. Robot. Res.*, vol. 32, no. 14, pp. 1700–1723, Dec. 2013.
[84] J. Kober, B. Mohler, and J. Peters, "Imitation and reinforcement learning for motor primitives with perceptual coupling," in *Studies in Computational Intelligence From Motor Learning to Interaction Learning in Robots*. 2010, pp. 209–225.
[85] M. Tamosiunaite, B. Nemec, A. Ude, and F. Wörgötter, "Learning to pour with a robot arm combining goal and shape learning for dynamic movement primitives," *Robot. Auto. Syst.*, vol. 59, no. 11, pp. 910–922, Nov. 2011.
[86] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–35, 2017.
[87] J. Buchli, E. Theodorou, F. Stulp, and S. Schaal, "Variable impedance control—A reinforcement learning approach," in *Robotics: Science and Systems VI*. MIT Press, 2010.
[88] J. Kober and J. Peters, "Reinforcement learning in robotics: A survey," *Adaptation, Learn., Optim. Reinforcement Learn.*, vol. 12, pp. 579–610, 2012.
[89] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *J. Intell. Robotic Syst.*, vol. 86, no. 2, pp. 153–173, May 2017.
[90] D. Silver, J. A. Bagnell, and A. Stentz, "Learning from demonstration for autonomous navigation in complex unstructured terrain," *Int. J. Robot. Res.*, vol. 29, no. 12, pp. 1565–1592, Oct. 2010.
[91] F. Sehnke, A. Graves, C. Osendorfer, and J. Schmidhuber, "Multimodal parameter-exploring policy gradients," in *Proc. 9th Int. Conf. Mach. Learn. Appl.*, Dec. 2010, pp. 113–118.
[92] J. Kober and J. Peters, "Policy search for motor primitives in robotics," *Mach. Learn.*, vol. 84, nos. 1–2, pp. 171–203, Jul. 2011.
[93] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1291–1307, Nov. 2012.
[94] F. Guerin, N. Kruger, and D. Kraft, "A survey of the ontogeny of tool use: From sensorimotor experience to planning," *IEEE Trans. Auto. Mental Develop.*, vol. 5, no. 1, pp. 18–45, Mar. 2013.
[95] B. Kim, J. Park, S. Park, and S. Kang, "Impedance learning for robotic contact tasks using natural actor-critic algorithm," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 2, pp. 433–443, Apr. 2010.
[96] B. Doroodgar, Y. Liu, and G. Nejat, "A learning-based semi-autonomous controller for robotic exploration of unknown disaster scenes while searching for victims," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2719–2732, Dec. 2014.
[97] A. Faust, I. Palunko, P. Cruz, R. Fierro, and L. Tapia, "Automated aerial suspended cargo delivery through reinforcement learning," *Artif. Intell.*, vol. 247, pp. 381–398, Jun. 2017.
[98] S.-M. Hung and S. N. Givigi, "A Q-learning approach to flocking with UAVs in a stochastic environment," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 186–197, Jan. 2017.
[99] C. Lian, X. Xu, H. Chen, and H. He, "Near-optimal tracking control of mobile robots via receding-horizon dual heuristic programming," *IEEE Trans. Cybern.*, vol. 46, no. 11, pp. 2484–2496, Nov. 2016.
[100] K. E. Merrick, "A comparative study of value systems for self-motivated exploration and learning by robots," *IEEE Trans. Auto. Mental Develop.*, vol. 2, no. 2, pp. 119–131, Jun. 2010.
[101] J. Yu, C. Wang, and G. Xie, "Coordination of multiple robotic fish with applications to underwater robot competition," *IEEE Trans. Ind. Electron.*, vol. 63, no. 2, pp. 1280–1288, Feb. 2016.

MD. AL-MASRUR KHAN received the B.Sc. degree in electronics and communication engineering from Khulna University, Khulna, Bangladesh, in 2020. He is currently conducting individual researches under the supervision of Dr. Abdullah Al Nahid, and also serving as the Remote Intern with the Intelligent Manipulation Laboratory, University of Lincoln, under the supervision of Dr. Amir Ghalamzan Esfahani. His research interests include deep learning, reinforcement learning, and robotics.

MD RASHED JAOWAD KHAN is currently pursuing the B.Sc. degree in electronics and communication engineering (ECE) with Khulna University. He has an avid interest in the thriving research fields of the Internet of Things (IoT), embedded systems, very large scale integration (VLSI), and Blockchain. In the future, he intends to integrate his research interests with a sophisticated real-life application through entrepreneurship and make out-of-reach techs available for the mass population.

ABUL TOOSHIL received the B.Sc. degree in electronics and communication engineering (ECE) from Khulna University, Bangladesh, in 2018. He is currently working with Aamra Technologies Ltd., as a System Engineer. His current research is primarily focused on photonics and optics. In the future, he would like to work on AI and ML.

NILOY SIKDER received the B.Sc. degree in electronics and communication engineering (ECE) from Khulna University, Bangladesh, in 2017, where he is currently pursuing the M.Sc. degree in computer science and engineering (CSE). His current research is primarily focused on developing biomedical, mechanical, and robotics applications using the existing machine learning and computational intelligence algorithms. In the future, he would like to work on the architecture of the existing methods and develop new techniques for better classification and prediction outcomes.

M. A. PARVEZ MAHMUD received the B.Sc. degree in electrical and electronic engineering and the M.Eng. degree in mechatronics engineering. After the successful completion of his Ph.D. degree with multiple awards, he worked as a Postdoctoral Research Associate and Academic with the School of Engineering, Macquarie University, Sydney. He is currently an Alfred Deakin Postdoctoral Research Fellow at Deakin University. He worked at the World University of Bangladesh (WUB) as a Lecturer for more than two years, and at the Korea Institute of Machinery and Materials (KIMM) as a Researcher for about three years. His researches are focused on energy sustainability, secure energy trading, microgrid control and economic optimization, machine learning, data science, and micro/nano scaled technologies for sensing and energy harvesting. He has accumulated experience and expertise in machine learning, life cycle assessment, sustainability and economic analysis, materials engineering, microfabrication, and nanostructured energy materials to facilitate technological translation from the lab to real-world applications for a better society. He was involved in teaching engineering subjects in the electrical, biomedical, and mechatronics engineering courses at the School of Engineering, Macquarie University, for more than two years. He is currently involved in the supervision of six Ph.D. students at Deakin University. He is a key member of Deakin University's Advanced Integrated Microsystems (AIM) Research Group. He has produced over 50 publications, including one authored book, three book chapters, 29 journal papers, and 21 fully refereed conference papers. He received several awards, including the "Macquarie University Highly Commended Excellence in Higher Degree Research Award 2019." Apart from this, he is actively involved with different professional organizations, including Engineers Australia and IEEE.

ABBAS Z. KOUZANI received the B.Sc. degree in computer engineering from the Sharif University of Technology, Iran, the M.Eng.Sc. degree in electrical and electronic engineering from The University of Adelaide, Australia, and the Ph.D. degree in electrical and electronic engineering from Flinders University, Australia. He was a Lecturer with the School of Engineering, Deakin University, and then a Senior Lecturer with the School of Electrical Engineering and Computer Science, University of Newcastle, Australia. He is currently a Professor at the School of Engineering, Deakin University, Australia. He is the Director of Deakin University's Advanced Integrated Microsystems (AIM) Research Group. He provides research leadership in embedded, connected, and low-power devices, circuits, as well as instruments that incorporate sensing, actuation, control, wireless transmission, networking and IoT, data acquisition/storage/analysis, AI, energy harvesting, power management, and fabrication for tackling research questions relating to a variety of disciplines including healthcare, ecology, mining, infrastructure, automotive, manufacturing, energy, utilities, and agriculture. He has authored over 370 publications, including one book, 17 book chapters, 180 journal papers, and 181 fully refereed conference papers. He holds three patents and two pending patents. He has been involved in over $15 million research grants, and has managed projects and delivered research solutions to over 25 Australian and International companies. He has received several awards, including the Outstanding Contribution to Scholarly Publication Award, School of Engineering, Deakin University, in 2019. He has supervised 24 research fellows/assistants, and produced 28 Ph.D. and six Master's by Research completions. He is currently involved in the supervision of 12 Ph.D. students.

ABDULLAH-AL NAHID received the B.Sc. degree in electronics and communication engineering from Khulna University, Khulna, Bangladesh, in 2007, the M.Sc. degree in telecommunication engineering from the Institute for the Telecommunication Research (ITR), University of South Australia (UniSA), Australia, in 2012, and the Ph.D. degree from Macquarie University, Sydney Australia, in 2018. His research interests include machine learning, biomedical image processing, data classification, and smart grid.