A Review of Action Anticipation in Human-Robot Collaboration

Pedro Amaral
DETI
Universidade de Aveiro
Aveiro, Portugal
pedro.amaral@ua.pt

***Abstract*—The increase in the diversity of products on sale due to the evolution of technology and standard of life results in a growing demand for flexible manufacturing that can meet the necessary production, especially in small companies. Although the usual solution for these needs is to use human operators, which provide the necessary flexibility and precision, this comes at a greater cost. In contrast, industrial robots offer a relatively smaller price and show more value in repetitive and heavy tasks. This is where Human-Robot Collaboration (HRC) comes into action since it complements the flexibility of a human worker with the strength and lower cost of the robotic worker in the same workspace. However, to achieve true collaboration it is not enough to react to the partner's movements and intentions, the robot must anticipate them. Inside HRC, Action Anticipation is a technique used to predict the actions of the human workers so that the robot can better plan its movements, increasing manufacturing efficiency and safety. This article reviews the research in this field, including the commonly used data sources and algorithms with a particular focus on machine learning methodologies. The nature of anticipation and the mechanisms that support it remains open questions in the field of HRC.**

***Index Terms*—Human-Robot Collaboration, Machine Learning, Action Anticipation, Predictive Model, Robot Controller, Collaborative Robot**

I. Introduction

The Third Industrial Revolution was characterized by a focus on automating repetitive and heavy tasks on the assembly lines. Still, this created a problem: whenever the manufacturers needed the robots to work in a different assembly process, they needed to be reprogrammed by an expert. The Fourth Industrial Revolution, also known as Industry 4.0, refers to the current trend of the manufacturing sector to become more intelligent and achieve greater automation. This trend takes advantage of the recent developments in artificial intelligence, the Internet of Things, and autonomous robots to pave the way for more efficient and flexible production processes. With Industry 4.0, robots are expected to be more adaptable and perform more actions without constant explicit programming.

The concept of Human-Robot Collaboration (HRC) emerges as part of Industry 4.0 and involves the research of mechanisms that allow humans and robots to work together to achieve a shared goal. Some of the most relevant topics in recent research include collision avoidance and human-aware planning of robot motions. However, to achieve true collaboration, it is not enough to react to the partner's movements and intentions, the robot must anticipate them.

The concept of anticipation has been studied in several research fields, such as biology, psychology, and artificial intelligence. One of the most cited definitions in the last decades and across the various fields is Rosen's [1]:

> An anticipatory system is a system containing a predictive model of itself and/or its environment, which allows it to change state at an instant in accord with the model's predictions pertaining to a later instant.

In general terms, anticipation is viewed as the impact of predictions on the current behavior of a system, be it natural or artificial. A prediction model provides information about the possible future state of the environment and/or system. This perspective of looking to the future is related to the purpose of incorporating that information into a decision-making or planning process. Accordingly, the system becomes anticipatory when it incorporates such a model and, simultaneously, when it uses the model to change its current behavior.

Over the last few decades, experimental evidence of the existence of anticipatory biological processes at different levels of organization have been reported [2]–[4]. The ability to modify behavior in anticipation of future events offers an adaptive advantage to living organisms with an impact on behavioral execution and learning. Anticipation is also considered one of the required abilities of cognitive robots operating in dynamically changing environments. The role of anticipation is to connect the robot's action in the present to its final goal, helping the design of robots with an increased level of autonomy and robustness.

The fundamental aspects of anticipation lie at the intersection of concepts such as time and information, involving abilities such as perception and prediction. The above definition of anticipation contains a temporal element that provides a key division between anticipatory and non-anticipatory robots. Anticipatory robots make decisions based on current states and predicted future states using predictive models of the environment. At the other extreme of the spectrum are the robots that live in the present based on the current state of the observed environment, which are usually called reactive robots (e.g., the Braintenberg's vehicles [5]). However, the behavior of a purely reactive robot is limited by its temporal horizon since they have no memory of the past to build a model of the

world. Most of the current robots present a behavior influenced either by the current perception as well as by the memory of past perceptions but still lacking a perspective of the future.

The nature of anticipation and the mechanisms that support it are considered open questions in AI and robotics. In the context of this article, anticipation is considered a combination of prediction and decision-making, as illustrated by the blocks diagram in Fig. 1. The prediction model offers the possibility of incorporating action selection in their planning through a decision-making block, while the planning module relates to the robot's actions. These modules can be developed separately, or an end-to-end learning technique could be used where the model learns the different parts from the perception to the feedback control.

There are different situations in which an anticipatory response seems to be an essential ability for effective robot behavior. In an attempt to distinguish different types of anticipatory behaviors, three contexts in which a robot can operate are categorized below and the respective task requirements are presented as follows:

- **Time synchronization**. The interception of moving objects is central to several benchmark robotic tasks such as ball-catching and playing table tennis [6], [7]. These tasks are challenging due to the demanding spatial–temporal constraints, which require continuous coordination between visual, planning and control systems. On one hand, frequent repredictions of the target location are required as new observations become available. On the other hand, this progressive refinement imposes an online re-planning of robot motion such that the goal is achieved in time.
- **Preventive safety**. Systems that manage risk require some form of anticipatory mechanism such that the robot can adapt its behavior when an undesired situation occurs. Autonomous driving is an example of how predicting future events and reacting properly are important abilities to mitigate risk. Modeling behavior and predicting the future intentions of pedestrians are core elements to ensure that the driver stops the car safely.
- **Coordinate joint activities in human-robot interaction**. Humans have the ability to coordinate their actions when carrying out joint tasks with other partners [8], [9]. In the same line of thought, anticipation can enhance the ability of a robot in its interaction with a human partner by predicting their actions (or intentions) before selecting its own action plan. In collaborative contexts such as those that occur during manufacturing or assembly tasks, the main challenge is combining anticipation and planning in a context of high uncertainty due to the variability of human behavior in complex industrial environments. Anticipation seems to have a significant potential for a more fluid and natural interaction with an impact on safety and cycle time.

This article aims at reviewing previous work relevant to the topic of action anticipation to enhance human-robot collaboration in industrial settings. This collaborative scenario is one in which the robot observes the actions of the human operator, makes predictions about the human's intention and reacts accordingly by either waiting for more observations or executing a physical action. Fig. 2 illustrates an example of a wood box being assembled. The collaborative robot's main function will be to assist with the assembly task by providing a wooden plank or a certain tool while coordinating its actions with those of the human operator who is focused on the assembly process.

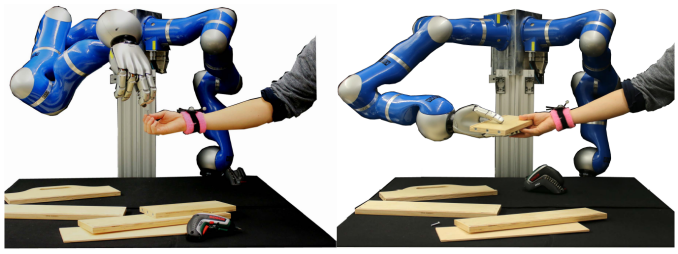

Fig. 2. Situation where a robot anticipates that the worker will need a wooden plank and hands it over to him [10]

The remainder of the document is organized into four chapters. Chapter II describes background concepts about collaborative robotics, including safety in HRC. Chapter III reviews the data sources and sensors used in previous work on Action Anticipation. Chapter IV contains a review of previous work on Action Anticipation in HRC, focusing on the methods. Chapter V concludes the document by stating the main points derived from this study.

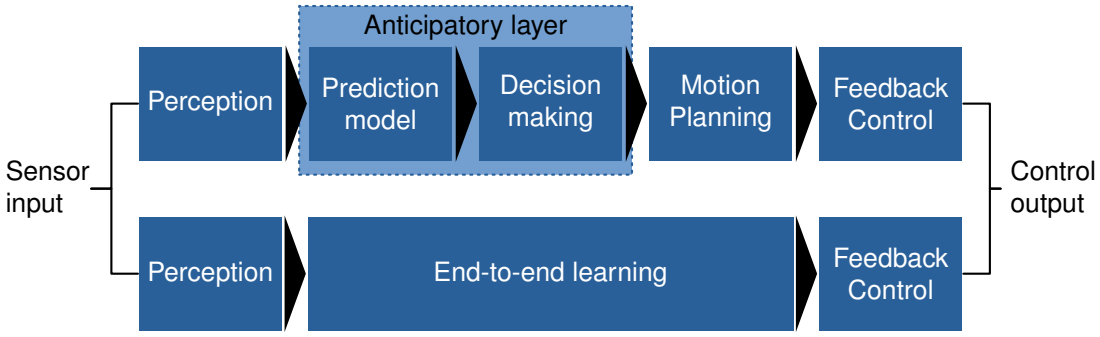

Fig. 1. Functional blocks of an anticipatory robotic system considering two alternative approaches: modules developed separately vs. end-to-end learning.

II. Collaborative Robotics

Human-robot collaboration (HRC) consists of robots and humans working in the same workspace towards a common goal. Classical industrial robots are usually automated to perform repetitive tasks that require high physical strength. On the other hand, tasks that require cognitive knowledge, flexibility, and precision are better suited for humans, even if they are physically weaker. HRC aims to take advantage of both of their strengths and complement each others' weaknesses to increase manufacturing efficiency.

In an HRC scenario, robots need to be different from the traditional ones, given that they will work in the same workspace as humans. According to Castro *et al.* [11], "Collaborative robots need to be endowed with a set of abilities that enable them to act in close contact with humans, such as sensing, reasoning, and learning. In turn, the human must be placed at the centre of a careful design where safety aspects and intuitive physical interaction need to be addressed as well.". In [12], it is stated that nowadays, collaborative robots are developed to be compact, easy to install and program, flexible, mobile, consistent and precise. Additionally, they positively impact employees since they are responsible for monotonous and dangerous actions and reduce the production cost for the company.

A. Safety

Safety is one of the most critical topics in collaborative robotics and the first step toward establishing a collaborative environment. According to [12], collaborative robots (or cobots) are able to safely work with people because they have sensitive sensors that can detect the human interrupting them, causing them to stop their actions, while traditional robots would potentially injure the worker. However, given that there are tasks that require the robot to move very close to the worker, some norms were implemented: ISO 10218-1 and 10218-2. From these two standards, Castro *et al.* [11] and Villani *et al.* [13] describe the four criteria from which at least one must be met as:

1) **Safety-rated monitored stop**: when a human enters the cobot's workspace, it completely stops;
2) **Hand guiding**: when an operator manually moves the cobot, it is compliant;
3) **Speed and separation monitoring**: as the human moves closer to the cobot, it becomes gradually slower;
4) **Power and force limiting**: the cobot has its operation restricted in terms of force and torque.

III. Data Sources and Sensors

This section covers which data is generally used in human action anticipation and which sensors can be used to capture that data from the environment. Humans and robots can communicate through several methods, which can be direct such as using a console or a remote, or indirect, resulting from data captured from sensors. Indirect communication can be further divided into methods that work in a more active way, such as voice commands, and those that work passively. In action anticipation, passive methods are used because the user should not need to do anything for the robot to act, the robot must be able to understand the worker's body language, such as his involuntary pose, gestures or gaze. Based on [11], [14], [15], the main kinds of data in indirect and passive communication can be seen in the diagram in Fig. 4 and can be described as follows:

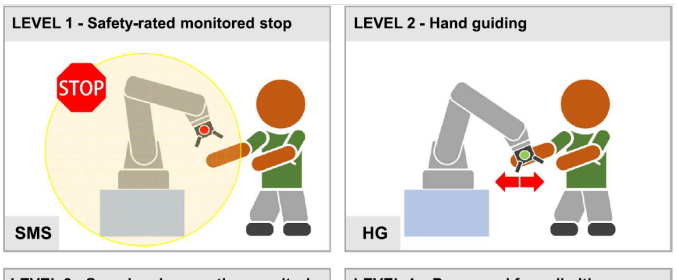

Fig. 3. The four collaborative operative modes identified by robot safety standards ISO modes 10218-1/2 [13]

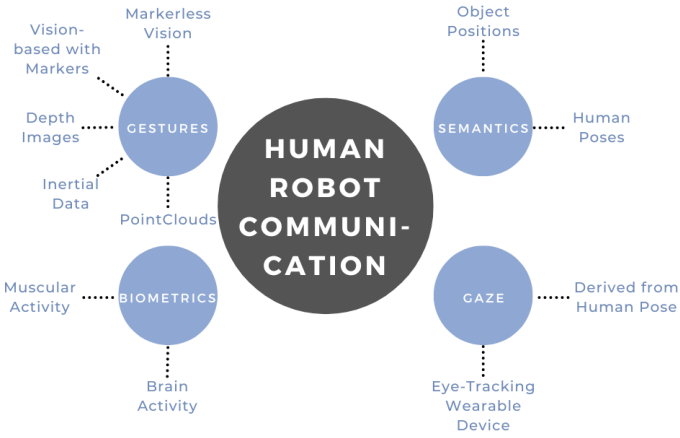

Fig. 4. Data sources common in action anticipation

- **Gestures**: these are one of the main ways humans communicate, whether through simple movements or formal sign language. In the literature about HRC, gestures can also commonly be found since they have the advantage of resisting ambient noise. Usually, gestures are captured with vision-based methods with either an RGB or RGB-D camera, so there is no need for unnatural movements. With vision, it is possible to include markers, but these may lead to occlusions and hinder the worker's movements. Consequently, there is also work in the literature that uses markerless vision to allow more unrestricted

movements. Another way to capture the movements of the human worker would be to use wearable inertial sensors, which contain accelerometers and gyroscopes, but, once again, wearables can hinder the worker's movements. Finally, capturing point clouds using a LIDAR presents another possibility of capturing gestures without restricting the worker's motion.

- **Semantics**: semantic information about the objects can also help the global workflow. Human actions can be represented semantically by obtaining the poses of the human as a specific set of limbs, even if only partially. During action prediction, this, coupled with the object positions, can be used to know which objects the worker can interact with. Having semantic information about the pose of the human body also helps in the path-planning phase of the robot since it can use this information to avoid the worker and prevent collisions.
- **Gaze**: this can be used to determine where the user's attention resides, giving a considerable amount of information that can trigger some action. There are two options to obtain the user's gaze. Wearable sensors can provide better results but are expensive and intrusive. On the other hand, algorithms that detect head pose and assume the gaze from it can also be used, which is a cheaper and non-intrusive solution.
- **Biometrics**: Electromyography (EMG) sensors can measure electrical signals generated by muscle contractions, while electroencephalography (EEG) sensors are commonly used in brain-computer interfaces (BCIs).

Regarding the sensors used to capture the raw data, most literature suggests using an RGB camera. However, the captured images may be used in the following different ways:

- directly used as input to models which can extract features from the images;
- used as input to frameworks that receive an image, process it, and return the key points, such as the skeleton joints of the person in the image; these key points can also then be used to assume the gaze of the human in the image such as in Canuto *et al.* [16] where the authors used OpenPose[17]–[20][1], an open-source project that aims to detect key points in the human body, face, hands, and feet from images as shown in Fig. 5, to obtain not only the skeleton joints but also the worker's gaze;
- used to process the optical flow [21]–[24];
- if the human was wearing markers, the image can be used to obtain the positions of the markers obtaining gestures from the sequence of those positions [10];

Besides RBG cameras, some works, such as the one described in Moutinho *et al.* [25], indicate the use of an RGB-D camera to capture both the color and the depth images, which contain the gestures and pose of the worker. Other than cameras, in Tortora *et al.* [26] IMU and EMG data was used as input to capture the gestures and anticipate the worker's action. When it comes to obtaining the worker's gaze, it is possible to do so from the RGB images, as mentioned before, but it is also possible to use wearable sensors to capture it, such as in Schydlo *et al.* [27].

[1]OpenPose documentation:
https://cmu-perceptual-computing-lab.github.io/openpose

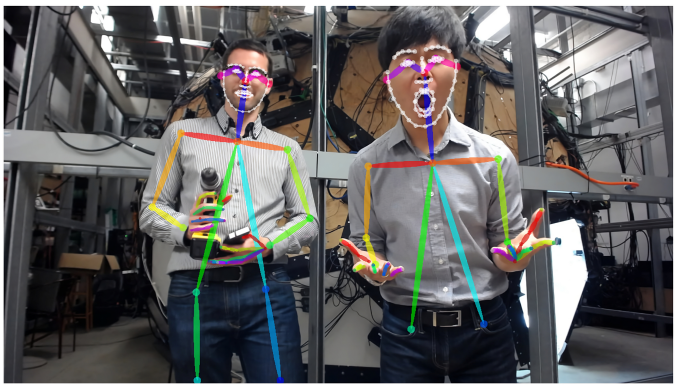

Fig. 5. OpenPose Example [17]

IV. Methods

After exploring which kinds of data are usually captured and provided to an algorithm, this section covers algorithmic solutions, particularly those that make use of machine learning methodologies.

Artificial intelligence has significantly evolved in the last few years. With the increase of computational power, machine learning, a subset of AI, has become an increasingly promising method to deal with complex and multidimensional data like images and text, heavily contributing to areas such as visual perception and speech recognition. Machine learning's ability to learn from data with minimal human intervention and make predictions or decisions from new data it has never seen before makes it a prime candidate to solve many problems in robotics and, in particular, action anticipation in collaborative environments. The most common strategies in this field are Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

- **Supervised Learning**: the models are trained using a dataset of labeled data. According to Sarker [28], these models must generalize the knowledge from the dataset's input-output pairs to correctly deal with a new input they have never seen before. The models from this group are further divided into classification, where the new input is assigned a discrete output class, and Regression, where it is returned a real number from the continuous output space. Currently, RNNs and CNNs are two of the most common classification approaches.
- **Unsupervised Learning**: the datasets involved have no labels. According to Sarker [28], these algorithms aim to find patterns and structure in the data. This makes them valuable in tasks such as clustering based on common characteristics, density estimation, identifying anomalies and outliers, dimensionality reduction, feature learning and finding association rules.

- **Reinforcement Learning**: machine learning approach different from the previous ones because it does not need a dataset. According to Alom *et al.* [29], the agent learns how to act in an unknown environment by interacting with it. After the agent's action, the environment returns an observation and a certain reward to the agent, depending on the quality of the action. The agent uses the reward to update its internal model named policy improving its future performance and the cycle repeats. This type of learning by trial and error has a certain resemblance to how humans gain knowledge, and it is useful when there is a need for an agent to make decisions in an environment that has considerable complexity, such as controlling a robot or playing a game.

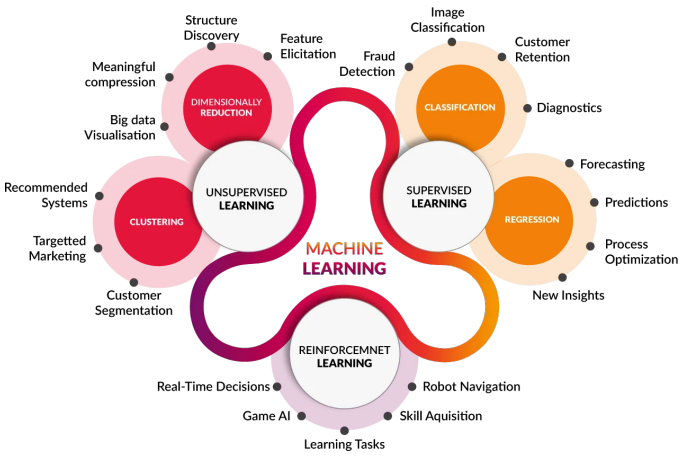

Fig. 6. Use cases of the different types of machine learning [30]

The next subsections mainly contain possible solutions present in previous work, which are described in reasonable detail since they vary not only in algorithms but also in the input data.

A. Predictive Modeling Techniques

Predicting the next action of the worker can be represented as a classification problem since it is possible to use a sequence of images that must be classified as a particular future action class. Using Fig. 7 as an example, the high-five action should be predicted before the frames that contain it are captured.

Fig. 7. Action anticipation using supervised learning diagram [21]

The previous work on predictive models mainly includes CNNs and RNNs, with the latter being the most common and transfer learning also being a frequent technique.

- **Recurrent Neural Network (RNN)**: type of neural network where the output of each time step is fed back into the input at the next time step, allowing the network to remember and incorporate information from previous time steps into its processing of current and future data. This characteristic makes RNNs particularly well-suited to processing sequential data, such as text, speech, or time series data which require context or temporal dependencies. In particular, according to [31], Long Short-Term Memory (LSTM) is an RNN with a more complex architecture that gives it an improved ability to backpropagate the error, making it better to train a model that classifies sequences with several time steps.
- **Convolutional Neural Network (CNN)**: type of neural network made up of several convolutional layers which apply a sliding filter over the input reducing its dimension and obtaining its features. Typically, these layers are followed by one or more fully connected layers that perform the prediction using the mentioned features. This architecture makes CNNs an excellent choice to deal with data in a matrix structure such as an image because this input is too massive for manual feature engineering.
- **Transfer Learning**: technique that makes use of a trained external model. Depending on the goal of its use, these models can be entirely or partially used; optionally, they can also be trained partially or fully. A common use case for this technique in supervised learning is when a small dataset of images is used to obtain a classifier. A standard model cannot generalize from that reduced amount of data. In this case, a model such as VGG-16 and ResNet-50 can be used partially to extract the features with one or more fully connected layers in the end, to perform the desired classification from those features.

In Furnari *et al.* [24], the authors aimed to predict the subsequent actions that someone wearing a camera would perform and the objects he would interact with. They used three datasets containing RBG frames from which they derived the optical flow and the objects in the environment. This data is then passed on to a Rolling-Unrolling LSTM. The Rolling LSTM (R-LSTM) is a network that continuously encodes the received observations and keeps an updated summary of the past. When it is time to make predictions about future actions, the Unrolling LSTM (U-LSTM) is used with its hidden and cell states equal to the current ones of the R-LSTM.

In Schydlo *et al.* [27], the authors used an encoder-decoder recurrent neural network topology to predict human actions and intent where the encoder and the decoder are both LSTM cells. At each step, the decoder returns a discrete distribution of the possible actions making this algorithm able to consider multiple action sequences, which are then subject to a pruning method that reduces them to obtain the right action finally. In their work, these algorithms were tested in two different datasets, one containing RGB images with optical markers and gaze information from wearable sensors and another with RGB-D images.

In Moutinho *et al.* [25], the authors aimed to increase the

natural collaboration between the robot and the human in an assembly station by interpreting implicit communication cues. The data related to the environment was captured using an RGB-D camera. This data was then passed on to a ResNet-34, a pre-trained neural network that extracted the features from the images. These features are used as the input to an LSTM to perform human action recognition.

In Gammulle *et al.* [21], the authors aimed to predict future frames while at the same time predicting the following action. In their implementation, they used public datasets with videos from which they obtained RGB images and optical flow streams. To consider both data sources, they also used two ResNet-50's, which are pre-trained networks, one to get the input features from the image and another from the optical flow, and 2 LSTMs to take into account both sequences of inputs. Then the two results are merged into a final classification. They also used two Generative Adversarial Networks (GAN) to generate the subsequent frames, but this is different from the focus of the analysis.

In Wang *et al.* [32], the authors used video datasets to train a model that would predict a future action from the observed frames. They used three pre-trained neural networks in their work: VGG-16, TS, and ConvNet, to extract features from the images. Then these features were aggregated using a Temporal Transformer module (TTM), and finally, a progressive prediction module (PPM) would anticipate the worker's future action. This article also addresses the issue of specifying what the algorithm should consider as an action. Although most of the literature often implies that the last frames captured by the camera are considered an action, given that those are the frames that contain the last action made by the user, the authors of this article go into greater detail. They tested and evaluated how many frames should be considered as the last action to obtain the best results using a metric from Geest *et al.* [33] named per-frame calibrated average precision (cAP) calculated with (1). In [32] it is defined with

$$cAP = \frac{\sum_k cPrec(k) * I(k)}{P}, \tag{1}$$

"... where calibrated precision $cPrec = \frac{TP}{TP+FP/w}$, $I(k)$ is an indicator function that is equal to 1 if the cut-off frame k is a true positive, P denotes the total number of true positives, and w is the ratio between negative and positive frames. The mean cAP over all classes is reported for final performance.".

In Rodriguez *et al.* [23], the authors aimed to predict the following action by first predicting the following motion images. They used datasets containing videos and then processed them to obtain motion images. These motion images become the input of a convolutional autoencoder network that generates the following motion images. These images are then passed to a CNN that processes them and makes action predictions for the future. The final action prediction is obtained from the results of the previous network and those of a second CNN, which analyzes the original RGB images.

In Wu *et al.* [22], the author's goal was to predict the following action someone wearing a camera would perform after some time. Initially, the optical flow was obtained from the captured images, and both were used as input to the model. The model is comprised of a Temporal Segment Networks (TSN), a CNN, and an LSTM to predict the future frame features and then use them to perform the required classification.

B. From Prediction to Planning

After predicting the next action of the worker, the robot must execute some action as a response to complete the anticipation process. This subsection contains articles that go beyond the predictive model and have relevant details for the integration of the model in a robot controller.

In Canuto *et al.* [16], the authors aimed to predict the following action using an LSTM, one of the most common RNNs. In their work, they used a dataset captured with an RGB camera. From these images, they obtained the objects in the environment, the human skeleton joints extracted over time using OpenPose, and the gaze derived from the joints. Then the three data sources were given to the LSTM as input to perform the desired classification. In this process, the authors use an adaptive threshold on the uncertainty of the recurrent neural network, which makes the model need a certain level of certainty to classify the action as a particular class. This creates a more robust solution since a standard supervised learning algorithm would predict the class with the highest probability even if the model has low certainty about every category.

In Maeda *et al.* [10], the authors aimed to reduce the delay in the robot's response by anticipating the human worker and providing a screw or a plate accordingly. They captured the environment using an RGB camera and tracked the hand using optical markers. Then they predicted the following human action using a look-up table containing different orders for assembly actions. With the nearest neighbor algorithm, the actions of the human would be matched with a particular order. The limitation of this method is that all possible sequences need to be on the table because if they are not there, then the robot will match with a different order which may be undesirable. If the robot eventually notices that it did the wrong action, it would then follow a hard-coded contingency trajectory to return to the pre-grasping position. When performing a handover action, the previously captured data is used to generate possible trajectories, and this is given to the feedback controller as a reference.

In Zhang *et al.* [34], the authors aimed to predict the intention of the human worker to provide him with the required piece. To achieve this, they used an RGB camera to capture the data from the environment. Then the images are given to a convLSTM framework where the CNN part is in charge of extracting features from the input images, and these features are then passed on to the LSTM to predict the intention. This article also tackles the issue of having several possible assembly orders. It solves it by creating a phase at the beginning of the collaboration in which the robot learns the assembly actions and their order from a demonstration. After

the prediction of the intention of the worker, the robot proceeds to fetch the required piece. It uses a CNN to recognize said piece and ROS Open Motion Planning Library (OMPL) to handle the trajectory planning jobs. In terms of safety, the authors defined speed limits for the robot and ensured that the robot would avoid the workspace of the human. Then when it needs to move closer to the user, its speed is reduced to guarantee the user's safety.

In Huang *et al.* [35], the goal of the authors is to make the robot use the anticipated actions of the worker to decide its tasks. It monitors the worker's gaze using a wearable device and uses it to predict his intent using SVM. After predicting it, the robot uses an anticipatory motion planner named "MoveIt!" to plan its motion according to a certain confidence threshold. This means that while it is unsure of what the human wants, the robot starts to move toward the item it thinks he wants but only really moves completely when it surpasses the threshold.

V. Conclusion

This document presented a review of the problem of anticipating human actions in collaborative environments. Although initially, it was also referred that an end-to-end learning technique could be used, the articles found all pointed towards separately developed modules. Looking at previous work found in the literature, there is a clear predominance of perception using RGB cameras with different ways of preprocessing the captured images. When it comes to the methods, machine learning and, in particular, supervised learning techniques are predominant, given that most work nowadays takes advantage of the progress made in that field. With the continuous evolution of machine learning, it is expected that the algorithms related to the topic in this paper also evolve and, consequently, give rise to even better solutions.

In summary, the results of this study demonstrate that Action Anticipation is still a relatively new concept, but it has much potential to increase the efficiency and safety of collaborative tasks, revolutionizing the world of human-robot collaboration.

VI. References

[1] R. Rosen, *Anticipatory Systems: Philosophical, Mathematical and Methodological Foundations*. Elsevier, 1985, pp. 339–347, ISBN: 9780080311586. DOI: 10.1016/C2009-0-07769-1.

[2] C. Deans, "Biological prescience: The role of anticipation in organismal processes," *Frontiers in Physiology*, vol. 12, Dec. 2021, ISSN: 1664-042X. DOI: 10.3389/fphys.2021.672457. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2021.672457/full.

[3] R. Poli, "The many aspects of anticipation," *Foresight*, vol. 12, R. Miller, Ed., pp. 7–17, 3 Jun. 2010, ISSN: 1463-6689. DOI: 10.1108/14636681011049839. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/14636681011049839/full/html.

[4] A. Louie, "Robert rosen's anticipatory systems," *Foresight*, vol. 12, R. Miller, Ed., pp. 18–29, 3 Jun. 2010, ISSN: 1463-6689. DOI: 10.1108/14636681011049848. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/14636681011049848/full/html.

[5] V. Braitenberg, *Vehicles: Experiments in Synthetic Psychology*. The MIT Press, Feb. 1986, ISBN: 9780262521123.

[6] D. Carneiro *et al.*, "Robot anticipation learning system for ball catching," *Robotics*, vol. 10, p. 113, 4 Oct. 2021, ISSN: 2218-6581. DOI: 10.3390/robotics10040113. [Online]. Available: https://www.mdpi.com/2218-6581/10/4/113.

[7] Z. Wang *et al.*, "Anticipatory action selection for human–robot table tennis," *Artificial Intelligence*, vol. 247, pp. 399–414, Jun. 2017, ISSN: 00043702. DOI: 10.1016/j.artint.2014.11.007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0004370214001398.

[8] N. Sebanz *et al.*, "Joint action: Bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, pp. 70–76, 2 Feb. 2006, ISSN: 13646613. DOI: 10.1016/j.tics.2005.12.009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661305003566.

[9] G. Hoffman *et al.*, "Cost-based anticipatory action selection for human–robot fluency," *IEEE Transactions on Robotics*, vol. 23, pp. 952–961, 5 Oct. 2007, ISSN: 1552-3098. DOI: 10.1109/TRO.2007.907483. [Online]. Available: https://ieeexplore.ieee.org/document/4339531/.

[10] G. J. Maeda *et al.*, "Anticipative interaction primitives for human-robot collaboration," in *AAAI Fall Symposium - Technical Report*, 2016, ISBN: 9781577357759.

[11] A. Castro *et al.*, "Trends of human-robot collaboration in industry contexts: Handover, learning, and metrics," *Sensors*, vol. 21, p. 4113, 12 Jun. 2021, ISSN: 1424-8220. DOI: 10.3390/s21124113. [Online]. Available: https://www.mdpi.com/1424-8220/21/12/4113.

[12] WiredWorkers, *Cobots*, Last accessed 3 January 2023. [Online]. Available: https://wiredworkers.io/cobot/.

[13] V. Villani *et al.*, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, Nov. 2018, ISSN: 09574158. DOI: 10.1016/j.mechatronics.2018.02.009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957415818300321.

[14] D. Mukherjee *et al.*, "A survey of robot learning strategies for human-robot collaboration in industrial settings," *Robotics and Computer-Integrated Manufacturing*, vol. 73, p. 102 231, Feb. 2022, ISSN: 07365845. DOI: 10.1016/j.rcim.2021.102231. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0736584521001137.

[15] F. Semeraro *et al.*, "Human–robot collaboration and machine learning: A systematic review of recent research," *Robotics and Computer-Integrated Manufacturing*, vol. 79, p. 102 432, Feb. 2023, ISSN: 07365845. DOI: 10.1016/j.rcim.2022.102432. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0736584522001156.

[16] C. Canuto *et al.*, "Action anticipation for collaborative environments: The impact of contextual information and uncertainty-based prediction," *Neurocomputing*, vol. 444, pp. 301–318, Jul. 2021, ISSN: 09252312. DOI: 10.1016/j.neucom.2020.07.135. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231220317719.

[17] Z. Cao *et al.*, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172–186, 1 Jan. 2021, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2019.2929257. [Online]. Available: https://ieeexplore.ieee.org/document/8765346/.

[18] T. Simon *et al.*, "Hand keypoint detection in single images using multiview bootstrapping," unpublished, Apr. 2017. [Online]. Available: http://arxiv.org/abs/1704.07809.

[19] Z. Cao *et al.*, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," unpublished, Dec. 2018. [Online]. Available: http://arxiv.org/abs/1812.08008.

[20] S.-E. Wei *et al.*, "Convolutional pose machines," unpublished, Jan. 2016. [Online]. Available: http://arxiv.org/abs/1602.00134.

[21] H. Gammulle *et al.*, "Predicting the future: A jointly learnt model for action anticipation," in *2019 IEEE/CVF International Conference on*

Computer Vision (ICCV), IEEE, Oct. 2019, pp. 5561–5570, ISBN: 978-1-7281-4803-8. DOI: 10.1109/ICCV.2019.00566. [Online]. Available: https://ieeexplore.ieee.org/document/9009844/.

[22] Y. Wu *et al.*, "Learning to anticipate egocentric actions by imagination," *IEEE Transactions on Image Processing*, vol. 30, pp. 1143–1152, 2021, ISSN: 1057-7149. DOI: 10.1109/TIP.2020.3040521. [Online]. Available: https://ieeexplore.ieee.org/document/9280353/.

[23] C. Rodriguez *et al.*, "Action anticipation by predicting future dynamic images," in *Computer Vision – ECCV 2018 Workshops*, Springer, 2019, pp. 89–105, ISBN: 9783030110147. DOI: 10.1007/978-3-030-11015-4_10. [Online]. Available: http://link.springer.com/10.1007/978-3-030-11015-4_10.

[24] A. Furnari *et al.*, "Rolling-unrolling lstms for action anticipation from first-person video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 4021–4036, 11 Nov. 2021, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2020.2992889. [Online]. Available: https://ieeexplore.ieee.org/document/9088213/.

[25] D. Moutinho *et al.*, "Deep learning-based human action recognition to leverage context awareness in collaborative assembly," *Robotics and Computer-Integrated Manufacturing*, vol. 80, p. 102449, Apr. 2023, ISSN: 07365845. DOI: 10.1016/j.rcim.2022.102449. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0736584522001314.

[26] S. Tortora *et al.*, "Fast human motion prediction for human-robot collaboration with wearable interface," in *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, IEEE, Nov. 2019, pp. 457–462, ISBN: 978-1-7281-3458-1. DOI: 10.1109/CIS-RAM47153.2019.9095779. [Online]. Available: https://ieeexplore.ieee.org/document/9095779/.

[27] P. Schydlo *et al.*, "Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction," IEEE, May 2018, pp. 1–6, ISBN: 978-1-5386-3081-5. DOI: 10.1109/ICRA.2018.8460924. [Online]. Available: https://ieeexplore.ieee.org/document/8460924/.

[28] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, p. 160, 3 May 2021, ISSN: 2662-995X. DOI: 10.1007/s42979-021-00592-x. [Online]. Available: https://link.springer.com/10.1007/s42979-021-00592-x.

[29] M. Z. Alom *et al.*, "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, vol. 8, p. 292, 3 Mar. 2019, ISSN: 2079-9292. DOI: 10.3390/electronics8030292. [Online]. Available: https://www.mdpi.com/2079-9292/8/3/292.

[30] A. Gupta, *Introduction to machine learning...* Last accessed 30 January 2023, 2017. [Online]. Available: https://medium.com/analytics-vidhya/introduction-to-machine-learning-c2cde23aded2.

[31] T. Miguel, *How the lstm improves the rnn*, Last accessed 20 January 2023, 2021. [Online]. Available: https://towardsdatascience.com/how-the-lstm-improves-the-rnn-1ef156b75121.

[32] W. Wang *et al.*, "Ttpp: Temporal transformer with progressive prediction for efficient action anticipation," *Neurocomputing*, vol. 438, pp. 270–279, May 2021, ISSN: 09252312. DOI: 10.1016/j.neucom.2021.01.087. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231221001697.

[33] R. D. Geest *et al.*, "Online action detection," unpublished, Apr. 2016. [Online]. Available: http://arxiv.org/abs/1604.06506.

[34] Z. Zhang *et al.*, "Prediction-based human-robot collaboration in assembly tasks using a learning from demonstration model," *Sensors*, vol. 22, p. 4279, 11 Jun. 2022, ISSN: 1424-8220. DOI: 10.3390/s22114279. [Online]. Available: https://www.mdpi.com/1424-8220/22/11/4279.

[35] C.-M. Huang *et al.*, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, Mar. 2016, pp. 83–90, ISBN: 978-1-4673-8370-7. DOI: 10.1109/HRI.2016.7451737. [Online]. Available: http://ieeexplore.ieee.org/document/7451737/.