Optimization of Competitive Strategies for Humanoid Fighting Robots Based on Dynamic Evaluation, Game Theory and Sequential Decision Making

Wenqi Jiang[1,*]

[1]North China University of Technology, Beijing 100144, China

[*]Corresponding author

Abstract

Combat-related research can provide a compact research benchmark for whole body dynamics, contact stability, and tactical decision-making. This paper has adjusted the mathematical modeling report of the PM01 humanoid robot competition to CPMI-style engineering research content. This paper proposes a four-layer strategic optimization framework that can cover different decision-making aspects of the robot fighting process. First of all, this paper will use Newton- Euler recursive dynamics, extended zero-point stability analysis, and seven standard decision matrices to evaluate the existing 13 attack actions one by one to ensure that the actual performance of each action can be accurately measured. Before using the AHP entropy-weighted TOPSIS model, this paper did the work of Pareto pre-screening, and also introduced the Mahanobis distance to reduce the computational bias caused by the relevant indicators, so that the final evaluation results are more relevant to the actual situation. Next, to deal with defense-related issues, this article uses the 13- by 22 attack defense effectiveness matrix to describe the correspondence between the attack and defense, and then transforms the problem into a limited zero-sum game to solve, and find a relatively reasonable defense strategy. Then there are the issues related to single-pos motion switching, which is described in this article as a limited Hoylon Markov decision-making process, and then training through the Q-learning method in self-game mode, so that the logic of action switching is more suitable for real fighting scenes. Finally, this article uses random dynamic programs to deal with the best three resource scheduling problems in random failure scenarios to reduce the impact of unexpected failures on fighting performance. The results obtained in this article show that the combination of punch, side kick and boxing and kick combination of these three actions, got the highest attack utility, scoring 0.812, 0.784 and 0.761, respectively, all the best performance in all attack actions. Nash's mixed-defensive strategy got 0.76 game points, and the training of a single-game strategy, in 10,000 Monte Carlo simulation tests, won 68.7%, outperforming many conventional decision-making strategies. The research in this paper also found that practical robot fighting strategies cannot rely on a single indicator to select actions, but also to optimize the factors of impact, poor stability, continuity, recovery ability and resource status, in order to select the most suitable action for the current scene.

Keywords

Humanoid robot,Newton-Euler dynamics, ZMP stability,TOPSIS, zero-sum game,Q-learning, stochastic dynamic programming.

1. Introduction

This article now tests the actual performance of humanoid robots in multiple mission scenarios, most of which require robots to have good balance, timely contact response capabilities, and

planning capabilities to cope with short-term situations. The battle scene with high competition intensity is a demanding test scene, in which the robot needs to complete the attack, defense, state recovery and adjust the action direction under very tight time constraints. If the action with a high instantaneous impact force pushes the pressure center beyond the support polygon, the final role will be greatly reduced, or even completely out of the expected effect. Conversely, if the action is too stable, the output may not be strong enough to change the score of the battle. Therefore, the strategic problems in such scenarios cannot be directly simplified into a single ranking of forces to judge the advantages and disadvantages [1].

The PM01 humanoid platform used in this paper has a total mass of 42 kg, and the platform itself adopts a multi-joint-driven structural design, which can complete a lot of flexible limb movements. There are 13 attack actions and 22 defensive actions in the rules of the game, and these actions and actions are based on clear competition standards. The actions that players can choose include boxing, kicking, combination of action, dodging offense, blocking and counterattack, and the applicable scenes and scoring rules of different actions are different [2]. The opponent will take what action will be no way to determine in advance, so even if a certain best response is now calculated, it is easy to fail in the actual battle [3]. The system also has three modes of limited recovery resources available, that is, the value of the reset, suspension or repair operation will change with the score change of the entire series, and will not always maintain a fixed value.

The modeling solutions we do at the beginning contain a very detailed derivation process for each problem. In response to the relevant requirements submitted by the meeting, the materials prepared need to be replaced with a more compact technical description [4]. This article retains all the main computational design ideas and two core digital results when adjusting the content, and removes the original long game-related descriptions and complete code contents in the appendix [5]. The rewrite version focuses on the construction process of the model, the specific equations used, the transfer logic of the parameters, and the corresponding engineering level interpretation. The aim is to make the entire presentation more in line with the requirements of the CPCI conference paper and not look more like a regular math competition report.

The main contributions are summarized as follows:

1. A dynamics-based attack evaluation model is coupled with an extended ZMP stability margin,

2. This article uses Pareto screening and AHP entropy TOPSIS model with Mahalanobis distance to carry out the ranking of attack actions, and the adaptability of related methods has also been initially reasonable, which can be adapted to the ranking needs of different types of attack actions under the current research scene, and the ranking process will also synchronize the characteristics of each attack action to ensure that the final ranking results can truly reflect the actual threat level of different attack actions.

3. Defensive decision making is solved by zero-sum game theory and dynamic programming, and

4. The single matching discussed in this article also has two different problems of tandem-level decision-making, which can be related by two commonly used research methods of reinforcement learning and random dynamic programming.

2. Methods

A. Dynamics-based attack evaluation

In this paper, each attack action can be fully represented by the corresponding movement chain, and the more common is the movement chain corresponding to the upper limb movement of the shoulder, elbow, wrist and then fist, as well as the movement chain of the hips, knees, and

ankles corresponding to the upper limb movement.TheDenavitt-Hartenbergformulation, modified in this paper, can be used to calculate the uniform transformation process and specific values from the base frame to the terminal strike point. The speed of motion of the terminal can be calculated by the geometric Jacobian, and the acceleration of the terminal is to include the Jacobian time derivative in the calculation range to obtain accurate results [6].

$$\begin{bmatrix}\mathbf{v}_{end}\\ \boldsymbol{\omega}_{end}\end{bmatrix} = J(\mathbf{q})\dot{\mathbf{q}}, \qquad \dot{\mathbf{v}}_{end} = J(\mathbf{q})\ddot{\mathbf{q}} + \dot{J}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}}. \tag{1}$$

After the model has enough momentum, the application of the Newton-Ule recursive method is selected to complete the next calculation steps. This article first carries out the calculation process of forward passing, mainly used to propagate the angular speed generated during the movement, as well as the angular acceleration parameters and the acceleration parameters corresponding to the quality center. After completing the forward propagation step, this paper is calculated backwards to obtain the force and torque parameters distributed along the various positions of the movement chain. The impact parameter of the terminal position will be restored from the already obtained joint torque by translating Jacobian's pseudo-reverse. For various movements driven by the waist, such as swing punching and round-house kicking commonly used in daily training [7], this paper also adds an additional rotational dynamic term to the calculation requirements for such actions on the basis of the original calculation model.

$$F_{total,i} = |(J_v^T)^+ \boldsymbol{\tau}_i| + \frac{I_{waist}\,\omega_{waist,i}^2}{2l_i}. \tag{2}$$

This treatment in this paper can clearly separate the two factors that would have been mixed together in a simple score, which are the equivalent of amazing quality and acceleration generated by the joint chain [8]. This article also shows the specific reasons why combined impulses have a high force value, but do not automatically become the best action [9]. This kind of combination impulse will mobilize more body quality to participate, but its subsequent recovery state and movement continuity are much weaker than the shorter duration of the combination of action.

B. Extended ZMP stability margin

This article mainly evaluates the stability of high-speed movements, and will use the extended zero-hour standard to do the relevant decision work. Compared with the simplified version of the ZMP calculation method, the extended form of computing will include the inertial tensor of the link and the angular acceleration terminology. This extended calculation method has a great effect on the analysis of the side kick and the rotational kick, and the main risks in the execution of such actions are not only forward displacement, but also the additional effects of the angular effect generated by the torso swing and leg swing [10].

$$DSM_i = \min_t dist((x_{ZMP}(t),\, y_{ZMP}(t)),\ \partial S) \cdot sgn(\text{membership}).$$

The poor positive dynamic stability discussed in this article shows that the ZMP trajectory can still remain within the range of the supporting polygon. The occurrence of negative values indicates that the corresponding action may cause a risk reduction. In the subsequent decision matrix, as long as the margin is negative, additional penalties can be applied to convert the unstable situation into the corresponding cost [11]. In this way, actions that are more influential but do not have a poor balance will not be overestimated.

C. Improved TOPSIS decision model

This article has built a seven-dimensional vector for each attack action to do the corresponding description, the specific seven dimensions are total impact force, attack coverage, reverse execution time, reverse stability cost, reverse energy cost, continuity and difficulty avoidance. The first thing to do in this article is to convert all the used indicators into revenue-type values, so as to facilitate subsequent unified calculation and comparison. Before formally calculating the scores for each action, this article first uses the Pareto screening method to eliminate those actions that dominate both strength and stability to avoid these special circumstances affecting the subsequent evaluation results [12]. The remaining attack operations, this article uses the combined weight vectors calculated in advance to make corresponding assessments one by one, and get the final score of each operation [13].

$$w_j = \frac{\sqrt{w_{AHP,j} \cdot w_{Entropy,j}}}{\sum_l \sqrt{w_{AHP,l} \cdot w_{Entropy,l}}}. \tag{3}$$

The final weight vector calculated in this paper is specifically w = (0.30, 0.13, 0.13, 0.21,0.09, 0.08, 0.06). The two elements of the corresponding weight in the first and fourth place can be explained that strength and stability are the most important considerations in the entire evaluation system, and this paper also completely retains all the remaining evaluation criteria, so that different sports performance can be based on speed, continuity or tactical coverage of the actual situation to obtain the corresponding additional rewards. In order to better handle the distance between the symlinearity problems in the data and the ideal solution, the distance between the ideal solution and the ideal solution is used in this article [14].

$$\mathrm{U_i} = \frac{\mathrm{d_i^-}}{\mathrm{d_i^+ + d_i^-}}, \tag{4}$$

$$d_i^{\pm} = \sqrt{(\mathbf{v}_i - \mathbf{z}^{\pm})^T \Sigma^{-1} (\mathbf{v}_i - \mathbf{z}^{\pm})}. \tag{5}$$

D. Defensive game and dynamic decision modules

The defense module designed in this article starts with a matrix R of 13 by 22. Each element in the matrix evaluates a specific defensive action individually, and can actually respond to a specific attack. The final calculated score combines the probability value of trajectory matching, as well as the possible instability risk of the defense action itself, and the content of the size of the counterattack window that is actively opened after the implementation of the defensive action [15]. If in the actual battle scene, the defender can not know the specific attack resources allocation of the opponent in advance, it will choose to solve the mixed strategy to develop a response plan, and will not only focus on the current seemingly best single response action.

$$\max_{\mathbf{q}, v} \; v, \tag{6}$$

subject to

$$R\,\mathbf{q} \;\geq\; v\,1, \qquad 1^T\mathbf{q} = 1, \qquad \mathbf{q} \geq 0. \tag{7}$$

The complete competition process studied in this article is divided into 60 independent decision steps. The state corresponding to each decision step contains information on multiple

dimensions, including the current game time, the score difference between the two teams, the overall energy level of the athletes, the stability of the current arena, and the team's recent actions [16]. We chose the Q-learning algorithm with the epsilon-greedy exploration strategy, and combined with the self-match training method to do model training. The set of reward function is very sensitive to the changes in the stage of the game, if the action performed in the early stage of the game can be put on the opponent will get a higher reward, the medium-term action will refer to the difference in the score and energy consumption to calculate the reward [17], to the final stage of the game, the reward calculation will pay more attention to the risk control effect of the action. When faced with the top three special problems, the state contains additional points for the series, and the remaining available resources are available. In the event of a possible failure in the game, we use the Poisson process to model, and then solve the corresponding treatment scheme through backward sensing [18].

3. Experimental Settings

The numerical experiments carried out in this paper refer to the actual physical scale of the PM01 robot to set the relevant parameters. The total mass of the robot is set at 42 kg, which is consistent with the real robot parameters. Corresponding to the quality of the forearm chain we set to 2.4 kg, the quality of the calf and the foot chain is set to 4.5 kg, and the quality of the trunk part is set to 18 kg, and the parameters of each part match the mass distribution of the actual structure. The rotational inertial parameters of the waist are set to 1.8 kg*m^2, which is in line with the actual motion characteristics of this type of robot. In the Q-learning model, we made a total of 10,000 Monte Carlo competitions to evaluate its actual performance and ensure the reliability of the evaluation results. The resource scheduling model used in this article considers three different resource types, namely manual reset, tactical pause and emergency repair, which basically covers the resource call scenarios that may be encountered in practical applications.

This article redraws the two numbers that need to be retained in English, and the layout details are specially adjusted when drawing, so as to facilitate the unified format requirements of the meeting. Figure 1 summarizes and summarizes the impact force of the 13 attacks involved in this study, and also adds the impact comparison between different actions. Figure 2 details the final calculation of the TOPSIS utility ranking results, the ranking also included with the score difference of each program. This article ultimately chose to show these two numbers because they clearly explain the core logic of the transition from mechanical action assessment to tactical action proposal throughout the study, which can help readers to clarify the research derivation context more quickly.

In the actual operation of this article, all the criteria used for judging will be unified and standardized before entering the weighted scoring link. Benefit class standards, we will be in accordance with the actual scope of its coverage to do a specific division, the cost class standard we will first add a small value of the normal number, and then convert it into reverse benefit value, in order to avoid the subsequent calculation when the split instability. This treatment method allows each component table to have a basis for comparison on the parameters of force, time parameters, energy parameters, and stable indicators, without the problem of contrast imbalance due to the difference in the properties of the standard itself.

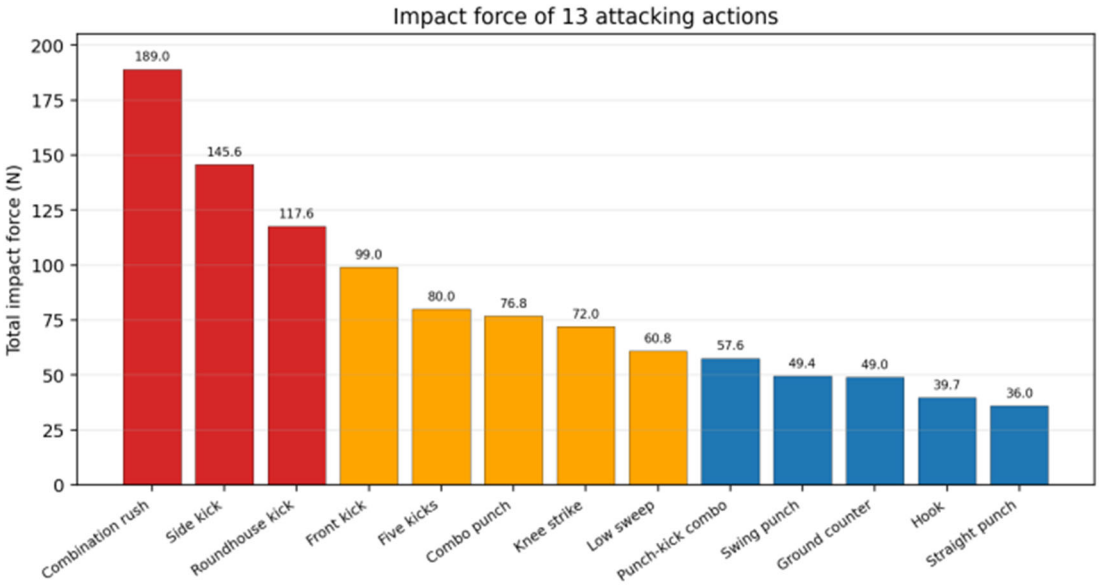

Figure 1. Impact force comparison of 13 attacking actions.

This article statistics out of the action score detailed ranking can be listed one by one, the highest score is the combination of punching, the corresponding value is 0.812, the second place is side kick, the corresponding score is 0.784, the third place is the punch combination, the score value is 0.761, the fourth place is the knee strike, the fourth place is the knee strike, the fifth place is the front kick, the corresponding score is 0.715, the sixth place is the low sweep, the score value is 0.703, the seventh place is the five kick, the score is 0.681, the eighth place is the combination of 0.665, the eighth place is the combination of 0.665, the eighth place is 0.665, the score is 0.665, the score is 0.6The score value is 0.598, the 11th action is the hook, the score is 0.572, the twelfth place is straight, the corresponding score is 0.541, the last place is the ground counter, the score is 0.421.

4. Results and Discussion

The impact data obtained by this paper can clearly show that the value of the combined impulse is the highest of all test types, and the final measured value reaches 189.0 N. Side kick and ball kicking are two kicks, and the actual measured strength value is also at a relatively high level. However, if you only look at the value of force, the relevant strategies obtained are incomplete, and sometimes it will have a misleading effect that is inconsistent with the actual situation. The explosive force of the combined action mobilizes the muscles of the whole body to participate in the force, its movement trajectory is relatively simple and direct, the opponent is easy to predict the direction of the action in advance, and it is difficult to smoothly connect with other actions to be done in the future. Roundhouse kick has a higher requirement for the body's rotation amplitude, and the negative stability of the body is also more prominent when doing action. These different types of attack actions can only play their due role in specific scenarios that meet the corresponding conditions.

This article uses the TOPSIS method to calculate the ranking, can give a reasonable tactical order more in line with the actual combat situation. The combined U score of Combo punch is 0.812, ranking first in all tactical moves tested. Its single strike force is not the largest of all actions, but it can maintain a more positive stability during the confrontation process, the continuity of the action connection is relatively high, and the recovery burden required after completing the action is relatively low. The comprehensive score of the side kick is 0.784, and finally ranks second, it integrates a strong impact force, and the difficulty of the opponent wants

to avoid is also very high, although its negative stability performance is not very good, or got the top ranking. The comprehensive ranking of the Punch-kick combination is in third place, its attributes are more balanced, there is no obvious short board, and after completing the action, it can be related to other tactical actions, and the adaptability in the actual battle is very good.

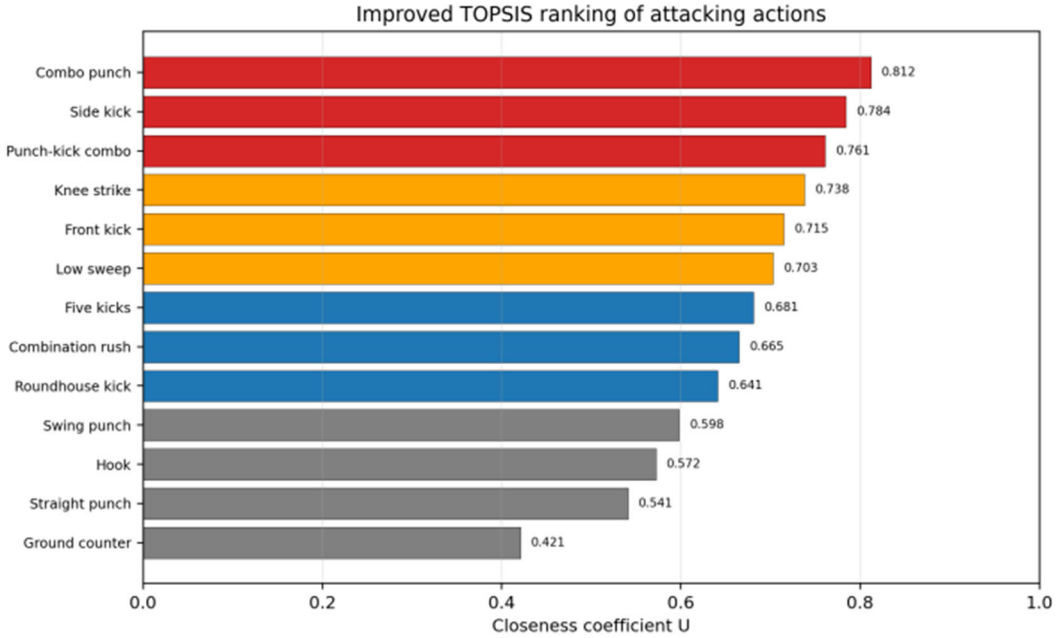

Figure 2. Improved TOPSIS utility ranking of attacking actions.

This paper statistics all the data of the four computing layers, the results are maintained a consistent direction, the relevant calculation of the attack evaluation has a positive effect on the stable continuous action output, the v value of the defensive game link finally gets 0.76, we comb out the best defense chain contains three coherent actions, first the turn to avoid the attack, and then by the center to recover the state after the launch of counterattack, the single-command strategy of the win rate can reach 68.7%, the BO3 scheduling mode of the winning rate is also 68.3%.

The defense results compiled in this article show that there is no single response that can dominate in all attacks. Turning to avoid, cross blocks and low avoidance of these actions, in the mixed strategy can get a relatively large probability of occurrence, the main reason is that these actions can cover different attack directions and different attack heights. When encountering this form of attack, the three action chains with the best effect are the combination of turning, quality compensation, and the round counter. The design of this action chain is reasonable, the first action can avoid the main impact of the attack, the second action can help the user to quickly restore the support state of the body, and the third action can use the opponent to adjust the recovery interval to respond.

The results of this paper through reinforcement learning can also support explanations for non-greed patterns. The policies that the model ultimately learn does not repeat the most profitable attack methods at every attack step. In the early stages of training, this policy prioritizes actions in the pressure-building class, slowly exploring the various responses that opponents may make. After entering the intermediate stage of training, this policy will consider the scoring situation and energy consumption at the same time, and try to keep the two in a relatively suitable state. In the final stage of training, if the current score is in a good position, the policy will tend to choose more secure actions, and if the current score is in a backward state, the

policy will become much more conservative than before. This pattern of behavior that changes with the stage is completely consistent with the actual competition logic in reality.

This article does relevant analysis of the three best-performing settings, and the random dynamic program used can clearly show the asymmetric value of the resource itself. If the robot wins the first game, it is often much more cost-effective to keep the repair resources and the relevant resources suspended first than to immediately use all these resources. If the robot loses in the first game, the marginal value of emergency maintenance will increase significantly, after all, if you lose again, the whole series will be directly over. This paper also did the corresponding sensitivity analysis, the final results show that the failure rate is the most influential of all parameters, but also shows that the reliability of the machine itself and fault-tolerant control role, and the tactical level of the scoring ability is exactly the same as the same.

The pipeline designed in this paper also has interpretable features. Each score in the final ranking can find the corresponding physical or tactical reasons, the value of the impact force is calculated from the dynamic characteristics, the value of the stability penalty comes from the calculation of the edge of ZMP, the value of the continuity corresponds to the feasibility of the transition of the action, and the value of the game can reflect the stability of the defense in the worst scene. This interpretability can play a practical role in the preparation phase of the competition, and engineers do not need to retrain the entire system, only need to adjust a single module to achieve the desired modification effect.

5. Conclusion and Future Work

This article mainly introduces the humanoid robot fighting competition format strategy optimization research related content, the research process to build the framework can be connected to dynamic action assessment, stability perception multi-standard ranking, defense game resolution, single-match reinforcement learning and five-game three-win system of random resource scheduling these different modules, we did a number of numerical simulation experiments after the results can be seen, the maximum impact of a single action between the maximum impact can not be directly equated. After many rounds of testing and comparison, we recommend the combination of punching, side kicking and punching as the main attack choice, these types of attack action in the impact, their own stability, movement continuity and tactical difficulty between several dimensions, to find a more appropriate balance, the actual game in the practical practicality is also stronger.

The work to be done in this paper focuses on three different directions. The first direction is that we will gradually identify the corresponding dynamic parameters from the actual PM01 motion log collected. The second direction is that we will put the Q-learning policy in a high-fidelity simulator with contact constraints and actuator restrictions, and conduct multiple rounds of testing to verify its actual performance. The third direction is that we will integrate the resource scheduling module and the online fault diagnosis module, so that we can rely on the actual data returned by the real sensor, rather than setting the fixed failure rate parameters in advance to trigger the corresponding emergency repair decision, so that the rationality of the decision can be improved.

References

[1] Spong, M. W., Hutchinson, S., & Vidyasagar, M. (2006). Robot modeling and control. Wiley.

[2] Craig, J. J. (2005). Introduction to robotics: Mechanics and control (3rd ed.). Pearson.

[3] Luh, J. Y. S., Walker, M. W., & Paul, R. P. C. (1980). On-line computational scheme for mechanical manipulators. Journal of Dynamic Systems, Measurement, and Control, 102(2), 69–76.

[4] Vukobratovic, M., & Borovac, B. (2004). Zero-moment point - thirty five years of its life. International Journal of Humanoid Robotics, 1(1), 157–173.

[5] Kajita, S., Kanehiro, F., Kaneko, K., Fujiwara, K., Harada, K., Yokoi, K., & Hirukawa, H. (2003). Biped walking pattern generation by using preview control of zero-moment point. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 1620–1626).

[6] Siciliano, B., Sciavicco, L., Villani, L., & Oriolo, G. (2009). Robotics: Modelling, planning and control. Springer.

[7] Saaty, T. L. (1980). The analytic hierarchy process. McGraw-Hill.

[8] Hwang, C. L., & Yoon, K. (1981). Multiple attribute decision making: Methods and applications. Springer.

[9] Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics, 18(1), 183–190.

[10] Mahalanobis, P. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India, 2(1), 49–55.

[11] Nash, J. (1951). Non-cooperative games. Annals of Mathematics, 54(2), 286–295.

[12] Osborne, M. J., & Rubinstein, A. (1994). A course in game theory. MIT Press.

[13] Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8(3–4), 279–292.

[14] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

[15] Bertsekas, D. P. (2017). Dynamic programming and optimal control (4th ed.). Athena Scientific.

[16] Puterman, M. L. (1994). Markov decision processes: Discrete stochastic dynamic programming. Wiley.

[17] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Veda, V., Silver, T., Kummaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529, 484–489.

[18] Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. International Journal of Robotics Research, 32(11), 1238–1274.