PAIS 2022
A. Passerini and T. Schiex (Eds.)

doi:10.3233/FAIA220064

Do You Like Dancing Robots? AI Can Tell You Why

Allegra De Filippo [1], Paola Mello and Michela Milano
Department of Computer Science and Engineering, University of Bologna

Abstract. Humanoid robots have been successfully used in artistic research areas, and many works have studied and implemented systems for robotic dance. However, only few works take into account the human evaluation of these artistic outputs. This work makes a step in the direction of addressing the complex task of defining criteria for the evaluation of robotic dance performances. For this aim, in the context of a Master course on Fundamentals of Artificial Intelligence (AI), we have organized a challenge among our students and the winner is decided on the basis of a questionnaire we defined for robotic dance evaluation. In addition, we created a public dataset that maps the features of each choreography to the judgements provided by audience with different backgrounds on several evaluation targets. Then, we tested various Machine Learning models for predicting the audience evaluation, and we propose a choreography features importance analysis to help both human choreographers and AI algorithms to create dance performances with a major impact on the audience. We also suggest new directions for future interdisciplinary research.

Keywords. Artistic Evaluation, Robotic Dance Choreography, Educational AI

1. Introduction

Dance is considered to be an important part of social lives and entertainment. The development of dancing robots for potential social interactions between humans and machines gained attention in the last ten years. Many researchers have focused on how to automate several aspects of dance [1,2,3], but only few studies have focused on how to define criteria for objective evaluation of dance performances [4,5]. There is an emerging need of exploring the relationships between Artificial Intelligence (AI) and evaluation of robotic dance performance. In this work we take into account human evaluation of AI artistic outputs, and we use AI techniques to discover relations between dance choreography features and audience evaluations.

We propose a contribution that starts from a set of robotic choreographies with AI techniques developed by students of a Master course of Fundamentals of Artificial Intelligence and Knowledge Representation[2] at the University of Bologna. Then, these artistic outputs are evaluated by difference audiences with an ad-hoc questionnaire that merges state-of-the-art evaluation methodologies. This evaluation phase leads first to the

[1] Corresponding Author; E-mail: allegra.defilippo@unibo.it
[2] https://www.unibo.it/it/didattica/insegnamenti/insegnamento/2021/446566

definition of a public training dataset, and finally to the testing and analysis of Machine Learning (ML) models for the evaluation of these AI-based choreographies. All these phases have been conducted in the context of a student competition. Students have been challenged, on a voluntary base, to develop an AI software able to produce a robotic choreography by exploiting their skills on planning and on informed/uninformed complete/heuristic search strategies. At the end of the course, the performances have been evaluated by the audience, so as to establish the competition winner. In order to have a more complete picture of the evaluation process, we considered as audience both students with a STEM[3] background and students with artistic background. Clearly, the chorerographies created use techniques that have been covered during the course. The objective of the project is not to create the best choreography ever, but to showcase the skills obtained during the course.

We can summarize the contributions of this work as follows: (1) a public repository with different implementations of AI techniques for robotic dance creation[4]; (2) an evaluation questionnaire for robotic choreographies; (3) the definition of a set of features that describe different aspects of robotic choreographies; (4) a public dataset for mapping dance choreography features with audience evaluation targets (over different audience backgrounds); (5) an evaluation of different ML models for predicting the evaluation targets; (6) a choreography feature importance analysis to help both human choreographers and AI algorithms to create dance choreographies with a major impact on the audience; (7) an innovative teaching tool providing new methodological perspectives.

2. Related Work

2.1. Computational Dance Automation

Many researchers have attempted to automate several aspects of dance, from dance notation to choreography, and from dance capturing to dance generation. [6,7] present systematic literature reviews by exploring the relations between computer science, information technology and the art of dance. These works cover different aspects of dance automation, such as dance representation, dance capturing, dance semantics, and dance generation. In our work, we focus on the automation of the dance generation process through basic AI techniques, and humanoid robots as performers. However, our aim is not only the automatic generation of robotic choreographies, but it is essentially the comprehension of the evaluation process of these performances.

2.2. Humanoid Robotic Dance

Humanoid robots are successfully used in several research areas. In the arts and especially in dance, where physical movement is the key factor, the use of robots is continually expanding as they have a physical body. Many works have studied and implemented systems for robotic dance. [2] proposes live demonstration of a humanoid robot performing fine-balanced dance movements with a human performer. [3,8] propose a humanoid dance robot system for a hip–hop dance sequence, based on a professional dancer basic

[3]STEM stands for Science, Technology, Engineering and Math.
[4]https://github.com/ProjectsAI/NAOPlanningChallenge

movements. [9] has experimented with the motions of a robot that are co-coordinated automatically to the music beat, by using a real-time music signal thus enabling a humanoid robot to dance autonomously. [10] proposes a dance motion imitation for humanoid robots through visual observation, for generating dance movements adequate to the music rhythm. All these works are focused on the implementation of robotic dances and on human-robot interaction, however they do not take into account a human evaluation of these artistic outputs.

2.3. Evaluation of Robotic Dance Performances

Defining a criterion for objective evaluation in the dance performance has been recognized as a very complex task, and measurement tools for the evaluation of qualitative aspects of dance performance such as Aesthetic Competence Evaluation (ACE) [11] or Performance Competence Evaluation Measure (PCEM) [4] have been developed for modern dance. These measurement tools are focused on evaluating human skills in dance, and they represent a standard in this context. [11] considers aspects, such as technique, space, time and energy, phrasing and presence; while in [4], the authors add a focus on the physical and motor qualities of the performer, including full use of the body, coordination, articulation of the body parts and motor skills, by defining three different levels of judgment for each area of evaluation.

However, as dance is entertaining and it is also considered to be an important part of social lives, the development of dancing robots for the potential social interactions between humans and machines gained attention. Dancing robots are more and more able to perform many kinds of robotic dances [12] due to their humanoid aspect. In this perspective, starting from the above measurement tools focused on evaluating human skills in dance, some works have focused also on evaluation metrics in robotic dance. Some recent works tackle the issue of how to bridge the gap between robotic dance and human perception, by mainly focusing on the perception of robotic body and movements [13,14,15]. [5] proposed a framework for evaluating robotic dance performances based on a Likert [16] questionnaire. Within this assessment, the areas taken into consideration are: 1) the harmony with the music; 2) the variety of movements; 3) the flexibility of the robot control application; 4) the human characterization of performance; 5) the entertainment for the public; 6) the application of the robot in the educational field. [17] presents a system that learns a set of movements for a creative dancer robot. An audience then evaluates the performance of the robots by giving a score from 0 to 5 on time and movements, naturalness of the dance and overall judgment. With a major focus on robotic poses, [18] presents an automatic aesthetic evaluation of robotic dance poses, in order to improve choreography creation: each robotic dance pose is expressed as a mix of features, and tagged with an aesthetic rating (good/bad). [19] focuses on the combination of human and robotic performers, by proposing a collaboration model in which an artificial intelligence system mediates the interaction between a human artist and an artificial artist. The artistic performance is evaluated using a Likert questionnaire with three parameters: 1) the harmony with the music; 2) the harmony between pianist and robot performer; 3) the overall judgment. All the above mentioned works share the issue of focusing on specific evaluation parameters for robotic dances. The most horizontal approach is [5] that considers different evaluation metrics of robotic dance performances. Differently from [5], in this work we propose an evaluation questionnaire that also includes aspects such as the overall theatricality of the dance choreography, the possibility of human reproducibility (related to the field of human-robot interaction), and the use of the surrounding space in the overall dance performance. Moreover, we also evaluate performances with audience having different backgrounds. More details are provided in Section 4.

3. Robotic Choreography Creation

The first part of the competition is related to the implementation of automatic and creative generation of complex movements in robots, such as dance choreographies. The students can test their acquired skills on the case study of the humanoid robot NAO[5].

3.1. Interdisciplinary Context

With the aim of exploring the relationships between AI and arts, an interdisciplinary group, focused on performing robots[6], was born in 2018. The group defines its field of investigation around the artificial body (human and robotic) dealing with case studies that affect the multiple disciplinary approaches expressed by the team members. Part of the attention is payed to robotics applications in performances. Exactly in this context, for three years a competition for the creation of robotic choreographies has been held within the course of Fundamentals of AI for the Master Degree of Artificial Intelligence at the University of Bologna[7]. One of the fundamental aspects of this competition aims at creating innovative teaching tools, in which technologies contribute to reconstruct the creation processes adopted by artists in the field of performing arts, thus providing the user with new methodological perspectives. Through this competition, the students can develop automatic tools to support performative creation by using AI techniques.

3.2. Competition Domain and Rules

The participants can decide to create teams of at most 2 students. Each group must plan a choreography (sequence of positions) given a problem description. Each group must choose a music suitable for the choreography (by respecting a total time limit) and test it on a simulated NAO robot (using the graphical tool Choregraphe). A final day of voting then takes place, and the winning choreography is decided by the audience, as detailed in Section 4.

In further details, NAO is a humanoid robot developed by SoftBank Robotics[8]. It is distributed with the custom operating system based on Unix-Linux NAOqi[9], on which the desired behaviors can be loaded. It supports different languages, including Python, in a complete and cross-compatible way. Moreover, a software suite is provided for testing complex projects on a simulated robot. This graphical interface allows to use a large amount of basic movements already available, and to enrich them by implementing custom ones through SDKs.

[5] https://www.softbankrobotics.com/emea/en/nao
[6] https://site.unibo.it/performingrobots/en
[7] https://www.unibo.it/en/teaching/course-unit-catalogue/course-unit/2021/446566
[8] https://www.softbankrobotics.com
[9] https://developer.softbankrobotics.com

The problem description is given to the students with an initial and final state description, as represented if Figure 1. The participants are also provided with a repository of robot positions[10] with mandatory and intermediate positions already implemented for the NAO robot (e.g., sit position, stand position, ...).

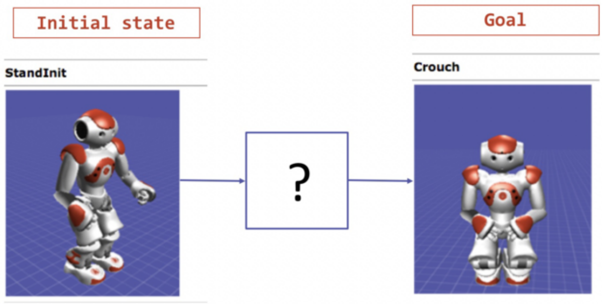

Figure 1. Competition Problem Description

The competition rules are the following (see also Figure 2):

- the choreography must start with the represented initial position and must end in the represented final state;
- the total duration of the choreography must be of 3 minutes at most;
- each choreography must contain at least one repetition of each mandatory position;
- to move from a mandatory position to another, it is possible to use positions from the available set of intermediate positions;
- each choreography must contain at least one repetition of 5 of the intermediate positions provided;
- the sequence of positions must avoid possible incompatibilities between two consecutive positions (e.g., for going to move_forward position from sit position we need intermediate positions).

The final task is to generate an algorithm *A* able to plan the sequence of positions satisfying the given constraints.

The choreographies implemented by students are based on three main AI techniques seen during the course: 1) Planning, 2) Search Strategy, 3) Constraints. All the choreographies are available at the public repository mentioned in Section 1. The implemented system must be able to automate dance choreography given, as input, an initial state, a goal and a set of movements description. The choreographies are built based on the analysis of possible inconsistencies between positions. The implemented algorithms are able to calculate sequences of admissible positions for robots dancer using off-the-shelf planners or search algorithms, and by respecting time/position constraints. The students have mandatory constraints to respect, but they have sufficient degrees of freedom in order to involve creativity in their final artistic output. All the details of the used techniques for each choreography can be found at the repository mentioned above.

[10]https://github.com/ProjectsAI/NAOPlanningChallenge/tree/main/RobotPositions

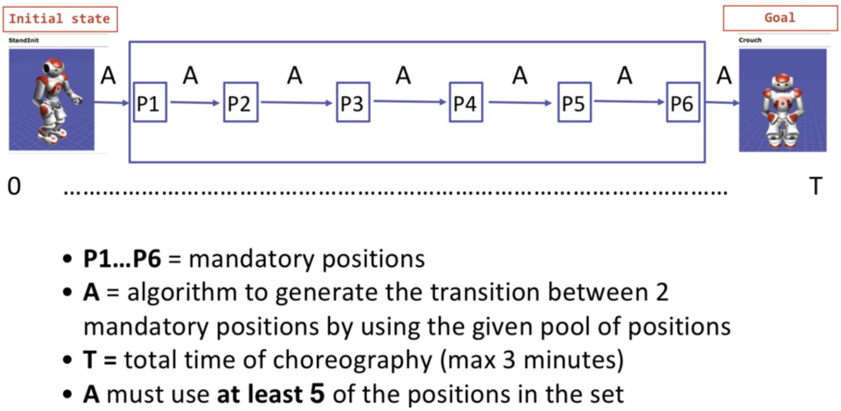

Figure 2. Competition Problem Constraints

4. Robotic Choreography Evaluation

The second phase of the competition is conducted to investigate the reactions and perceptions of the audience during the staging of performances involving robots as performers. The methodological tools used for data collection are participant observation and questionnaires. Based on the state of the art, we define 7 evaluation parameters and we define a survey to evaluate the robotic choreographies implemented by students. We define two macro group of audience: audience with a scientific background, and audience with an artistic background. Both the groups are composed by master students.

4.1. Artistic Evaluation Survey

If dance researchers and choreographers want to measure and understand the effect of various aspects of robotic dance performance on the audience, reliable systems for evaluation need to be developed. Until then, much of the assessment will either remain theoretical, with no consideration for how changes in those aspects may impact the quality of performance. In this perspective, we define the evaluation parameters of the survey questionnaire by merging state-of-the-art methodologies for the evaluation of both human and robotic dance performances. This is the first step to build the training dataset for the ML models illustrated in Section 5.

For the realization of the questionnaire, we follow the Likert structure with a judgment from “strongly disagree” to “strongly agree”. We consider that 5 levels of evaluation is a good approximation for questionnaires, based on [20]. The areas that are taken into consideration here are merged among four main state-of-the-art methodologies [21,4,5,19]:

- The performance embodies a theme or tells a story [21]
- The performance has rhythmic coherence with music [4,5]
- The performer presents fluidity of movement transitions [4]
- The performer is able to involve the public [5]
- The performer extensively uses the surrounding space [4]
- The performer movements have human characterization [5]
- The choreography can be reproduced also by or with a human performer [19]

4.2. Results Collection

The goal of this evaluation phase was to understand how the different audiences perceive the robotic dances. From the literature in robotic perceptions in arts, it has emerged that humans over-estimate the capacity of artificial systems (especially unfamiliar artificial systems) and underestimate the capacity of natural systems (such as humans) [22]. For this reason we consider two types of audience with dramatically different backgrounds: scientific and artistic. The evaluation phase for each type of audience is made through the following steps: 1) the audience views all the robotic performances; 2) the audience has access to the survey questionnaire and can read all the questions; 3) the audience watches again the performance one-by-one and completes the questionnaire. To avoid decision fatigue [23], performances order is randomized. The experimental setting is based on 13 choreographies evaluated by 31 attendants in the scientific audience and 31 attendants in the artistic one.

5. Dataset Construction

For the dataset construction, we analyze and extract the features from the choreographies. Then, we collect all the evaluations for each choreography, for each evaluation target and for each participant. We build two different datasets based on the different backgrounds (artistic and scientific).

5.1. Feature Extraction

The two datasets are composed by 403 records (31 run x 13 instance choreography). For each choreography in input, we encoded 20 features. The features are extracted by analyzing each single choreography. We collect them related to each evaluation target, based on the state-of-the-art analysis (see Section 5.1).

Analyzed aspect	Related works
Music parameters	[19]
Movement parameters	[19,1] [24]
Artistic performance parameters	[24,1]
AI system parameters	[19,25]

Table 1. Areas of extracted features and related works

For each record, the following information are stored:

- *timeDuration* (t): an integer that represents the number of seconds of the coreography duration.
- *nMovements* (n): an integer that represents the number of different movements of the coreography.

- *movementDifficulty* (*md*): integer that represents the level of movement difficulty of the coreography. The considered levels (i.e., low, medium, high) are associated to the interval $md \in [1,3]$.
- *AItechnique* (*ai*): a categorical feature representing the AI technique used to implement the robot choreography, with $ai \in [planning, searchStategy, constraints]$.
- *robotSpeech* (*rs*): a boolean value to indicate the presence of robot speech in the choreography (i.e., 0=absence, 1=presence).
- *acrobaticMovements* (*am*): integer that represents the level of acrobatic movements, $am \in [1,3]$.
- *movementsRepetition* (*mr*): integer that represents the level of movement repetitions, $mr \in [1,3]$.
- *musicGenre* (*m*): a categorical feature representing the music genre used for the robot choreography, with $m \in [Electronic, Rap, Rock, Folk, Indie, Pop]$.
- *movementsTransitionsDuration* (*mtd*): integer that represents the level of movement transition duration, $mtd \in [1,3]$.
- *humanMovements* (*h*): integer that represents the level of human movement presence, $h \in [1,3]$.
- *balance* (*b*): integer that represents the level of balance movements, $b \in [1,3]$.
- *speed* (*s*): integer that represents the level of movement speed, $s \in [1,3]$.
- *bodyPartsCombination* (*bc*): integer that represents the level of different body parts combinations, $bc \in [1,3]$.
- *musicBPM* (*bpm*): integer that represents the number of music Beat Per Minute.
- *sameStartEndPositionPlace* (*pp*): a boolean value to indicate if the robot starts and finishes in the same place of the background (i.e., 0=no, 1=yes).
- *headMovement* (*hm*): integer that represents the level of different body parts combination of movements, $hm \in [1,3]$.
- *armsMovement* (*arm*): integer that represents the level of different body parts combination of movements, $arm \in [1,3]$.
- *handsMovement* (*hdm*): integer that represents the level of hands movement presence, $hdm \in [1,3]$.
- *legsMovement* (*lm*): integer that represents the level of legs movement presence, $lm \in [1,3]$.
- *feetMovement* (*fm*): integer that represents the level of feet movement presence, $fm \in [1,3]$.

The evaluation targets follow the survey questions (see Section 4). In details, they are represented by a range of values between 1 and 5, and they are: 1) Choreography Story Telling, 2) Choreography Rhythm, 3) Choreography Movement Technique, 4) Choreography Public Involvement, 5) Choreography Space Use, 6) Choreography Human Characterization, 7) Choreography Human Reproducibility.

The two datasets are publicly available[11].

[11] https://github.com/ProjectsAI/NAOPlanningChallenge/tree/main/datasets

6. Experimental Analysis

6.1. ML models comparison

As a first experiment, we tried to assess whether the ML algorithms are able to predict the evaluation based on the performance features. In particular, four different ML models were taken into account: (1) linear regressor (LR), (2) decision trees (DT), (3) random forests (RF), (4) gradient boosting (GB). We focused on symbolic and explanatory models, both for the training dataset, which is not excessively vast, and to explore the open issue on interpretable ML.

In all experiments we performed 10-fold cross-validation; we report only the average values over all folds, as the difference between the different folds were negligible. All the categorical input features are managed with one-hot encoding representation as binary vectors, and the numerical features are normalized. The ML models are implemented in the open source Python library scikit-learn [26] with a cross validation matrix to find the best hyper-parameters for each model. We used a grid search to tune the hyperparameters for all the ML methods.

In Table 2 we report the results obtained with the different ML models[12]. The first column represents the regression target; the second column is the ML model used. Then, the table presents two triple columns, reporting Mean Average Error (MAE), Root Mean Squared Error (RMSE), and Standardized Mean Average Percentage Error (SMAPE); for each metric, we distinguish between using the dataset based on the artistic audience, and the scientific one. The reported results are averaged over all test set instances.

The results reported show that the ML models have learned the relationship between input features and targets. For all targets the slightly best models are GBs, closely followed by RFs. Interestingly, the human reproducibility regression target is better predicted in the artistic dataset, while in the scientific one is more or less equal to the Rhythm regression target. Moreover, the human characterization target seems to be more difficult to predict in the artistic dataset, compared to the other regression targets and to the scientific dataset. These observations lead to analyze in more details the feature importance of GBs that is very close to the RFs one.

6.2. Feature Importance

Feature importance indicates the relative importance of each feature when making a prediction. The scores are useful to provide insight about the dataset, e.g., to highlight which features may be most relevant to the target, and which features are the least relevant. This then may be interpreted by a domain expert and could be used as the basis for gathering more/different data.

We show this analysis for the GBs results, since they obtained the best results in predicting targets. All the other results are available in the repository mentioned above. In Table 3 and Table 4 we report the feature importance for each regression target for the GB regression models. The first column represents the feature; then, the table presents a double column, one for each target, reporting the feature importance for each different dataset (A=artistic, S=scientific). The reported results are averaged over all test set instances.

[12] https://github.com/ProjectsAI/RoboticPerformanceArtisticEvaluation

Target	ML	Artistic Background			Scientific Background		
		MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
Story	LR	0.98	1.22	15.72	0.97	1.18	15.95
Story	DT	1.06	1.32	17.15	1.01	1.25	16.76
Story	RF	**0.96**	**1.19**	**15.41**	**0.96**	**1.16**	**15.72**
Story	GB	0.98	1.18	15.63	0.98	1.17	15.97
Rhythm	LR	**0.94**	**1.15**	**14.88**	0.89	1.14	13.47
Rhythm	DT	1.01	1.27	16.03	0.87	1.13	13.35
Rhythm	RF	0.95	1.14	15.04	0.86	1.11	13.12
Rhythm	GB	0.95	1.13	15.02	**0.84**	**1.09**	**12.88**
Movements	LR	0.96	1.14	16.01	0.93	1.16	14.96
Movements	DT	1.05	1.26	17.59	0.97	1.23	15.76
Movements	RF	0.96	1.13	15.93	**0.92**	**1.13**	**14.81**
Movements	GB	**0.94**	**1.10**	**15.75**	0.94	1.13	15.03
Public Impact	LR	1.05	1.21	17.28	0.98	1.19	15.66
Public Impact	DT	1.08	1.30	17.68	1.05	1.28	16.81
Public Impact	RF	1.02	1.18	16.80	**0.94**	**1.15**	**15.18**
Public Impact	GB	**1.02**	**1.17**	**16.80**	0.94	1.16	15.23
Space Use	LR	1.00	1.39	17.11	0.90	1.11	14.87
Space Use	DT	1.02	1.20	17.64	0.95	1.20	15.99
Space Use	RF	1.00	1.16	17.12	**0.88**	**1.09**	**14.70**
Space Use	GB	**0.99**	**1.16**	**16.97**	0.89	1.09	14.72
Human Char	LR	1.07	1.24	19.05	0.94	1.14	15.33
Human Char	DT	1.12	1.35	20.20	1.00	1.24	16.55
Human Char	RF	1.06	1.22	18.87	**0.90**	**1.10**	**14.76**
Human Char	GB	**1.05**	**1.21**	**18.77**	0.92	1.11	15.15
Human Reprod	LR	0.68	0.89	8.84	0.89	1.16	12.62
Human Reprod	DT	0.76	0.99	9.92	0.91	1.22	13.23
Human Reprod	RF	0.68	0.88	8.79	0.86	1.13	12.31
Human Reprod	GB	**0.67**	**0.87**	**8.67**	**0.85**	**1.13**	**12.18**

Table 2. Performance comparison between the different ML models.

The results highlight interesting trends. The time duration of a choreography seems to be an important feature to predict all the selected evaluation targets. The same holds for the number of different movements and the music bpm. The interesting trends are also to be considered for the different datasets. In Table 3, for the first target related to the coherence of the choreography story telling, the feature scores show differences between the two audience backgrounds: the number of movement repetitions and the speed of movements seem to be more relevant for the artistic audience, while for the scientific one these scores are related to the level of movement transitions, and the level of combination of the different body parts. The target related to the evaluation of choreography rhythm, shows the importance of the features related to the movement difficulty and the same position of the robot in the background space at the beginning and at the end of the performance; moreover the difference between the two datasets lies in the greater importance of legs movements (for A) and music genre rap (for S). Another interesting result

Feature	(%) feature coeff for Story Telling		(%) feature coeff for Rhythm		(%) feature coeff for Mov Technique	
	A	S	A	S	A	S
time	**15.03**	**15.71**	**8.69**	**29.75**	**15.66**	**16.28**
n_mov	**8.87**	**8.59**	**7.43**	**7.78**	**6.23**	**7.38**
difficulty	3.91	3.70	**11.64**	**4.11**	**4.70**	2.75
speech	2.12	2.30	0.98	0.52	2.78	1.83
acrobatic	1.67	3.52	2.67	2.20	1.91	2.00
mov_rep	**4.53**	3.74	3.05	2.79	3.12	2.86
mov_trans	3.78	**4.45**	1.97	1.89	**4.17**	**5.68**
human_mov	3.23	3.31	3.81	1.10	3.96	2.66
balance	3.71	3.54	1.20	1.86	3.24	**4.78**
speed	**4.86**	3.55	1.57	2.57	**4.56**	2.88
body_comb	3.91	**4.15**	1.96	3.37	3.61	3.38
bpm	**11.21**	**12.91**	**15.63**	**8.76**	**12.08**	**12.18**
equal_place	2.64	2.06	**10.04**	**16.67**	2.89	2.24
head_mov	2.51	3.22	2.68	0.53	**4.12**	3.60
arms_mov	3.08	0.87	3.43	2.39	2.36	3.15
hands_mov	3.91	2.40	1.99	2.44	**4.28**	**4.58**
legs_mov	2.46	2.50	**6.91**	1.72	3.17	2.10
feet_mov	2.54	2.61	1.03	0.36	1.27	2.63
constraints	1.69	1.81	1.82	0.11	1.83	2.98
planning	2.02	2.49	1.13	0.11	2.84	2.64
search	1.95	0.75	1.26	0.31	1.00	1.56
electronic	0.84	3.31	2.22	0.31	1.69	1.43
folk	2.91	1.88	2.71	1.78	2.13	2.45
indie	1.15	1.06	0.09	0.84	1.48	1.70
pop	2.72	1.71	0.20	0.31	1.16	1.96
rap	0.93	2.77	0.96	**4.05**	2.28	1.26
rock	1.71	0.98	2.80	0.44	2.38	1.93

Table 3. Feature importance comparison of GBs over the audiences.

is represented by the different feature importance for the third target related to the evaluation of movements technique. In this case, we can notice that the movements difficulty, the movements transitions, and the head movements are more important for the artistic audience, while the scientific audience is more focused on the balance of the robot. In Table 4 are analysed the remaining targets, and we can notice that we have some differences between the two audiences for all the targets. For the target related to the public involvement, we can see that A is more focused on the feature related to the presence of human movements, to the body combinations level, and to the head movements; while S is more focused on the movements repetitions and transitions, and hands movements and AI planning technique used to compute the choreography. For the target related to the use of the space, we can see a more focus on movements repetitions and transitions for the scientific audience. An interesting insight is proposed by the analysis of the human characterization target: the artistic audience is more focused on movements speed and body combinations, while the scientific audience considers more the presence of human movements, balance and head movements. Finally, for the target related to the human reproducibility, we can observe that A is more focused on the presence of acrobatic

Feature	(%) feature coef Public Impact A	(%) feature coef Public Impact S	(%) feature coef Space Use A	(%) feature coef Space Use S	(%) feature coef Human Char A	(%) feature coef Human Char S	(%) feature coef Human Repro A	(%) feature coef Human Repro S
time	**28.29**	**16.28**	**11.48**	**15.32**	**15.84**	**19.14**	**13.45**	**15.90**
n_mov	**8.81**	**8.45**	**9.30**	**8.08**	**8.85**	**7.83**	**6.92**	**10.28**
difficulty	2.22	3.53	3.88	3.65	3.65	2.13	**4.92**	**5.58**
speech	0.68	1.13	2.41	3.48	2.14	1.92	3.58	1.96
acrobatic	0.62	1.88	2.12	2.09	2.16	1.30	**6.71**	3.98
mov_rep	1.59	**4.29**	3.68	**4.55**	**5.04**	2.35	1.61	**5.19**
mov_trans	2.20	**9.24**	1.97	**5.89**	**4.39**	**7.08**	3.64	**4.09**
human_mov	**4.99**	3.33	3.91	3.28	3.14	**4.40**	**4.40**	**4.00**
balance	2.21	3.69	3.88	3.54	3.24	**4.09**	3.69	**4.57**
speed	2.91	2.86	3.78	3.41	**4.65**	2.57	2.40	3.37
body_comb	**5.76**	2.60	**4.89**	**4.11**	**4.34**	**4.00**	3.43	2.30
bpm	**13.34**	**11.30**	**15.63**	**8.94**	**13.92**	**10.45**	**10.40**	**10.80**
equal_place	2.33	2.68	2.62	2.48	1.37	1.27	0.65	1.93
head_mov	**4.04**	2.54	**5.96**	**4.17**	3.59	**4.10**	2.49	**4.03**
arms_mov	1.34	1.51	2.63	2.36	2.19	1.93	1.31	1.02
hands_mov	1.85	**4.41**	3.40	**4.03**	1.85	3.72	**8.04**	3.29
legs_mov	0.49	1.53	2.27	2.04	3.10	1.74	1.58	1.18
feet_mov	0.42	2.01	2.94	2.06	2.41	1.39	0.93	2.64
constraints	0.78	0.79	1.56	2.23	2.31	3.52	**4.99**	1.92
planning	3.88	**4.00**	1.61	2.07	1.27	3.51	1.68	1.41
search	1.04	1.07	1.92	1.86	1.57	1.11	1.45	1.29
electronic	1.37	0.88	1.25	1.90	1.39	1.20	0.52	1.25
folk	1.39	1.95	2.19	2.07	1.27	1.57	1.07	1.36
indie	1.11	1.95	1.64	1.13	0.87	1.26	0.65	1.45
pop	1.12	2.39	2.29	2.19	2.06	2.96	0.81	1.68
rap	3.14	2.59	1.43	1.83	1.77	1.91	**7.77**	1.08
rock	1.99	1.04	2.40	1.12	1.51	2.71	0.76	2.31

Table 4. Feature importance comparison of GBs over the audiences.

movements, hands movements and AI constraints technique; while S is more focused on movements repetitions and transitions, balance and head movements.

7. Discussion & Conclusion

Due to the emerging need of exploring the relationships between AI and evaluation of robotic dance performance, in this work we make a first step to address the complex task of defining a criterion for objective evaluation in robotic dance performance. We take into account human evaluation of AI artistic outputs, and we use AI techniques to discover relations between dance choreography features and audience evaluation.

The results highlights interesting trends: the time duration of a choreography seems to be an important feature to predict all the selected evaluation targets; the same holds for the number of different movements and the music bpm. The interesting trends are also to be considered for the different audiences: for example, for the target related to the public involvement, we can see that the artistic audience is more focused on the feature related to the presence of human movements, to the body combinations level, and to the head movements; while the scientific audience is more focused on the movements repetitions and transitions, and hands movements and AI planning technique used to compute the choreography.

Related to this emerging trends, with this work, we suggest new directions for future interdisciplinary research. Firstly, we plan to analyze a further step in the proposed competition, by providing feature importance results to students of the next year as guidelines to implement the robotic choreographies. The idea is to provide this information to half of the classroom and then compare the different choreographies created with and without this information. We also plan to involve different class of users to enrich the audience and the study on the different artistic evaluation provided. Moreover, we have observed that the didactic strategy based on competitions has successfully improved students' motivation and collaboration during the course. Secondly, in the field of interpretable machine learning, we can try to explain the relationship between the input features and the final value of the predicted target. The idea is to understand if a positive increase in the feature importance is also related to a positive increase in the predicted evaluation target. In this sense, we can be able to provide more useful indications to both human choreographers and AI algorithms to create robotic choreographies. We plan to analyze, by exploiting the interdisciplinary context of the performing robot group, if this feature importance can be also applied to human performers. Finally, these ML models can be embedded in an optimization problem, which provides the optimal feature values for a specific choreography in input, given a set of user-specified constraints (e.g., bounding the total time, or the number of movements, or the music bpm,...) [27]. The advantage is that we can train the ML models only once and reuse them in the optimization model on different data instances and different user-defined constraints, as they are posted on the backbone of the optimization model when needed. In this perspective, this work also provides a useful starting benchmark.

Acknowledgements

This work has been partially supported by European ICT-48-2020 Project TAILOR (g.a. 952215). We thank the performing robots group [13] and the students Mattia Marcanti, Davide Lanzoni and Nicolò Bari for the support to the construction of the artistic evaluation questionnaire. We would like to thank all the students who have participated to the competition.

References

[1] Manfrè A, Infantino I, Vella F, Gaglio S. An automatic system for humanoid dance creation. Biologically Inspired Cognitive Architectures. 2016;15:1-9.

[2] Ramos OE, Mansard N, Stasse O, Benazeth C, Hak S, Saab L. Dancing humanoid robots: Systematic use of osid to compute dynamically consistent movements following a motion capture pattern. IEEE Robotics & Automation Magazine. 2015;22(4):16-26.

[3] Shinozaki K, Iwatani A, Nakatsu R. Concept and construction of a dance robot system. In: Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts; 2007. p. 161-4.

[13] https://site.unibo.it/performingrobots/en

[4] Krasnow D, Chatfield SJ. Development of the "performance competence evaluation measure": assessing qualitative aspects of dance performance. Journal of Dance Medicine & Science. 2009;13(4):101-7.

[5] Oliveira JL, Reis LP, Faria BM. Empiric Evaluation of Robot Dancing Framework based. TELKOMNIKA. 2012;10(8):1701-8.

[6] Sagasti F. Information technology and the arts: the evolution of computer choreography during the last half century. Dance Chronicle. 2019;42(1):1-52.

[7] Joshi M, Chakrabarty S. An extensive review of computational dance automation techniques and applications. Proceedings of the Royal Society A. 2021;477(2251):20210071.

[8] Shinozaki K, Iwatani A, Nakatsu R. Construction and evaluation of a robot dance system. In: Entertainment Computing Symposium. Springer; 2008. p. 83-94.

[9] Grunberg D, Ellenberg R, Kim IH, Oh JH, Oh PY, Kim YE. Development of an autonomous dancing robot. International Journal of Hybrid Information Technology. 2010;3(2):33-43.

[10] Angulo C, Comas J, Pardo D. Aibo jukeBox–A robot dance interactive experience. In: International Work-Conference on Artificial Neural Networks. Springer; 2011. p. 605-12.

[11] Chatfield S, Byrnes W. Correlational analysis of aesthetic competency, skill acquisition and physiologic capabilities of modern dancers. In: 5th Hong Kong International Dance Conference Papers; 1990. p. 79-100.

[12] Aucouturier JJ, Ikeuchi K, Hirukawa H, Nakaoka S, Shiratori T, Kudoh S, et al. Cheek to chip: Dancing robots and AI's future. IEEE Intelligent Systems. 2008;23(2):74-84.

[13] Gemeinboeck P, Saunders R. Movement matters: How a robot becomes body. In: Proceedings of the 4th international conference on movement Computing; 2017. p. 1-8.

[14] Gemeinboeck P. Dancing with the nonhuman. Thinking in the world London: Bloomsbury Academic. 2019.

[15] Gemeinboeck P. The aesthetics of encounter: a relational-performative design approach to human-robot interaction. Frontiers in Robotics and AI. 2021;7:217.

[16] Likert R. A technique for the measurement of attitudes. Archives of psychology. 1932.

[17] Manfrè A, Infantino I, Augello A, Pilato G, Vella F. Learning by demonstration for a dancing robot within a computational creativity framework. In: 2017 First IEEE International Conference on Robotic Computing (IRC). IEEE; 2017. p. 434-9.

[18] Peng H, Li J, Hu H, Zhao L, Feng S, Hu K. Feature fusion based automatic aesthetics evaluation of robotic dance poses. Robotics and Autonomous Systems. 2019;111:99-109.

[19] Saffiotti A, Fogel P, Knudsen P, de Miranda L, Thörn O. On human-AI collaboration in artistic performance. In: First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostella, Spain, September 4, 2020. CEUR-WS; 2020. p. 38-43.

[20] Taherdoost H. What is the best response scale for survey and questionnaire design; review of different lengths of rating scale/attitude scale/Likert scale. Hamed Taherdoost. 2019:1-10.

[21] Wicke P, Veale T. Creative Action at a Distance: A Conceptual Framework for Embodied Performance With Robotic Actors. Frontiers in Robotics and AI. 2021;8:115.

[22] Cuan C, Berl E, LaViers A. Measuring human perceptions of expressivity in natural and artificial systems through the live performance piece Time to compile. Paladyn, Journal of Behavioral Robotics. 2019;10(1):364-79.

[23] Pignatiello GA, Martin RJ, Hickman Jr RL. Decision fatigue: A conceptual analysis. Journal of health psychology. 2020;25(1):123-35.

[24] Augello A, Cipolla E, Infantino I, Manfre A, Pilato G, Vella F. Creative robot dance with variational encoder. arXiv preprint arXiv:170701489. 2017.

[25] Jeon M. Robotic arts: Current practices, potentials, and implications. Multimodal Technologies and Interaction. 2017;1(2):5.

[26] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.

[27] De Filippo A, Borghesi A, Boscarino A, Milano M. HADA: an automated tool for hardware dimensioning of AI applications. Knowledge-Based Systems. 2022.