Feature fusion based automatic aesthetics evaluation of robotic dance poses

Hua Peng a,h,c, Jing Li d,* Huosheng Hu c, Liping Zhao a, Sheng Feng a, Keli Hu a

a Department of Computer Science and Engineering, Shaoxing University, Shaoxing, China
b College of Information Science and Engineering, Jishou University, Jishou, China
c School of Computer Science and Electronic Engineering, University of Essex, United Kingdom
d Academy of Arts, Shaoxing University, Shaoxing, China

HIGHLIGHTS

• An approach to automatic aesthetics evaluation of robotic dance poses is proposed.
• Machine learning based on feature fusion is deployed to train aesthetics models.
• Inspired by human dance, the approach fuses information of visual and non-visual.
• Verified by experiments, the highest correct ratio of aesthetic evaluation 81.8%.
• The approach enables a robot to behave more autonomous and humanoid behaviours.

ARTICLE INFO

Article history:
Received 18 June 2018
Received in revised form 17 September 2018
Accepted 29 October 2018
Available online 9 November 2018

Keywords:
Robotic dance pose
Feature fusion
Machine learning
Automatic aesthetics estimation

ABSTRACT

Inspired by human dancers who make a comprehensive aesthetic judgement of their own dance poses by using both visual and non-visual information, this paper presents a novel feature fusion based approach to automatic aesthetics evaluation of robotic dance poses in order to improve the performance of robotic choreography creation. Four kinds of features are extracted, namely kinematic feature, region feature, contour feature, and spatial distribution feature of colour block. Based on different feature combinations, machine learning is deployed to train aesthetics models for the automatic judgement on robotic dance poses. The proposed approach has been implemented on a simulated robot environment, and experimental results are presented to verify its feasibility and good performance.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Robotic dance is an interesting research topic in the fields of artificial intelligence and human–robot interaction [1–3]. It can be classified into four categories: (i) cooperative human–robot dance; (ii) imitation of human dance motions; (iii) synchronization for music, and (iv) creation of robotic choreography. The creation of robotic choreography is mainly focused on improving the autonomous dancing ability of a robot so that it could understand all sorts of dance objects generated, including dance pose, dance motion, and robotic dance. Notably, the understanding of aesthetics is an important and essential aspect as good robotic choreography must be in accordance with human aesthetic. The category of robotic choreography creation has the aesthetic requirement [1].

As the fundamental part of robotic dance, dance pose is a static body shape and expresses emotion, character, feeling, meaning and theme [4]. Existing research results show that dance poses are always manifested in several different forms, such as stopping posture [5], key-pose [6], gesture [7], posture [8], etc. The essence of a dance pose, despite its diverse forms, is unity. The aesthetics cognition of a robot on its own dance pose will help it to create robotic dance and further improve its autonomous ability.

In the existing research, the aesthetic problem of dance robots has been explored from three levels: robotic dance pose [8–12], robotic dance motion [13,14], and robotic dance [15–18]. There are two kinds of aesthetic evaluation methods of robotic dance poses: human subjective aesthetics and machine learning based methods. However, the automatic aesthetic evaluation method of robotic dance poses, which draws lessons from the mature aesthetic experience of human beings, is still insufficient and rarely studied so far.

The quality of a dance pose directly affects the dance. To create a dance, human dancers always design and evaluate new dance poses by using a mirror. When human dancers present their dance poses before a mirror, they make their aesthetic judgements on their own dance poses by integrating multimodal information. More specifically, they observe the mirror images of their dance
poses by eyes (visual channel), and also perceive their body kines-
tate (non-visual channel) simultaneously. Inspired by this process, a
similar mechanism could make a humanoid robot achieve auto-
matic aesthetics evaluation on its own dance pose. However, the
following three questions remain:

1) How can a humanoid robot integrate multimodal information
(vision and non-vision) to comprehensively judge the aesthet-
ics of its own dance pose?

2) Which kind of feature combination can be used to describe
a robotic dance pose more completely and accurately?

3) Which kind of machine learning can achieve more accurate
results for the above aesthetic judgement?

To address these questions, this paper presents a new feature
fusion approach to the automatic machine aesthetics evaluation
of robotic dance poses. By processing information from visual and
non-visual channels, the humanoid robot can extract four kinds
of features (kinematic feature, region feature, contour feature,
and spatial distribution feature of colour block) respectively. After
feature selection, combination and fusion, machine learning is used
for training a machine aesthetics model of robotic dance poses for
the automatic judgement on robotic dance poses.

The rest of the paper is organized as follows: Section 2 outlines
some existing works related to this research. Then, the proposed
approach is detailed in Section 3, which consists of six parts: whole
framework, pre-processing, feature extraction, feature selection
and combination, feature fusion, and machine learning. Section 4
introduces the simulation process and presents the experimental
results, and Section 5 discusses our mechanism from six aspects.
Finally, a brief conclusion and future work are given in Section 6.

2. Related work

As an important survey of robotic dance, Peng et al. [1] classified
the area of robotic dance into four categories: cooperative human–
robot dance, imitation of human dance motions, synchronization
for music, and creation of robotic choreography. Among them, aes-
thetic requirements focus on the category of the creation of robotic
choreography. The machine aesthetics of robotic dance creation,
an important feature of good robotic choreography, must be in
accordance with human aesthetics [1]. Viewed from the abstract
level of aesthetic objects, there are three levels from low to high:
robotic dance pose, robotic dance motion, and robotic dance. Good
aesthetics in the lower level (robotic dance pose) help to improve
the aesthetics in the higher level (robotic dance). According to
different purposes, many researchers judged the aesthetics on
different aesthetic levels.

Using a human subjective evaluation of aesthetics as the basis
for interactive evolutionary computation, Vircikova et al. imple-
mented a multi–robot system, in which aesthetics were judged on
robotic dance poses and robotic dance motions, and then chore-
egraphed a dance of humanoid robots [8–10]. To find good robotic
dance poses that are in accordance with human aesthetics, we
presented an approach of semi-interactive evolutionary compu-
tation [11]. In this approach, an autonomous machine aesthetics
model of robotic dance poses is built by machine learning. Only the
kinematic (joint) feature is used to supervise learning. The highest
correct ratio for evaluating aesthetics, based on AD Tree, is 71.7%.

Moreover, we presented a multimodal information fusion ap-
proach to estimate the aesthetics of the robot dance poses by fusing
visual pose images and corresponding joint motions [12]. Based on
machine learning, several machine aesthetic models were trained.
Among them, the highest correct ratio for evaluating aesthetics is
81.6%, which comes from AD Tree based on the mixed features
(shape feature + joint feature).

Eaton [13], by constructing an aesthetic fitness function for
robotic dance motions, presented a theory for the creation of
robotic choreography based on traditional evolutionary compu-
tation. The fitness function involves the sum of the values of all
the movements over all the joints multiplied by the time the
robot remains standing [13]. Moreover, Shinozaki et al. presented
a random generation method for the creation of robotic choreog-
raphy [14], in which human subjective aesthetic evaluations were
carried out on Hip-Hop robotic dance motions.

Oliveira et al. constructed a choreographic framework, in which
a Lego NXT robot performs its dance motions in response to the
inputs of multimodal events [15]. Their method is typical of a mapp-
ing rule for the creation of robotic choreography. The aesthetics
of the generated robotic dance were empirically evaluated with
questionnaires.

Manfre et al. [16] presented an automatic system for robotic
dance creation based on the Hidden Markov Model, which used
the Viterbi algorithm to search the optimal robotic dance choreog-
raphy according to the inputed music. Three professional dancers
evaluated the aesthetics of the generated robotic dances. The mean
value of their evaluation was 6.33 in a score range [1–10], where
10 is the best.

Furthermore, Manfre et al., showing that a Nao robot could learn
dance motions from a human demonstration, built a set of dance
motions to serve as the basis for the creation of robotic dance [17].
Using questionnaires, audiences evaluated the aesthetics of the
generated robotic dance. In a score range of [0–5], where 5 is the
best, the mean value for overall artistic value was 4.8.

Augello et al. presented a cognitive architecture (that integrated
Hidden Markov Models and Genetic Algorithms) embodied in
a humanoid robot, who, driven by the perception of music, created
and performed dances [18]. Improvisational robotic dances, based
on human–robot interaction, were explored and demonstrated by
organizing public live performances. By the use of questionnaires,
positive feedback from the audience was acquired.

Overall, in the existing literature, there are only the follow-
ing two methods for evaluating the aesthetics of robotic dance
poses: human subjective aesthetics [8–10] and the machine learn-
ing based method [11,12]. The paucity of automatic methods for
evaluating the aesthetics of robotic dance poses is holding back
the creation of autonomous robotic choreography, which devel-
ops the autonomous ability of a robot. Therefore, inspired by the
corresponding mechanism of human aesthetics, we propose a new
automatic machine approach, based on feature fusion, to evaluate
the aesthetics of robotic dance poses.

3. Automatic machine aesthetics of robotic dance pose

3.1. Whole framework

Inspired by the mature aesthetic experience of human beings
(See Section 1), a humanoid robot should be able to use a similar
mechanism to automatically evaluate the aesthetics of its own
dance poses. Placed before a mirror, a humanoid robot uses its
visual cameras (“eyes”) to observe its own presented dance poses.
It uses its embedded sensors to feel its own internal kinestate
(motor parameters). Combining the two, a humanoid robot fully
understands its presented dance poses and comprehensively eval-
uates their aesthetics. Thus, regarding the aesthetics of a dance
pose, a humanoid robot exhibits greater humanoid behaviour.
Thus, to evaluate the aesthetics of robotic dance poses, we propose
an automatic machine approach based on feature fusion. Fig. 1
shows the whole framework of the approach.

At first, kinestate data and visual images are pre-processed on
the two different channels. Then kinematic features and visual
features (region, contour, spatial distribution of colour block) are
extracted from the kinestate data of the robot and the visual image,
respectively. The existence of these different kinds of features
results in a diverse feature selection and combination, and the different feature combinations result in the different aesthetic effects of the robotic dance poses.

After suitable feature selection and combination, several specific features are fused and combined with the human evaluation of the aesthetics (provided by human dance experts) to train a machine aesthetics model (supervised learning). Finally, the trained machine aesthetics model is used for automatic machine aesthetics on robotic dance poses.

3.2. Pre-processing

Pre-processing is carried out from two aspects: kinematic parameter pre-processing and image pre-processing. In this paper, the pre-processing is extended based on that of [12], and the spatial distribution image of colour blocks is acquired supplementally as a new product of image pre-processing.

3.2.1. Kinematic parameter pre-processing

When it presents a robotic dance pose, a humanoid robot acquires all the kinematic status information from its embedded sensors. Among all the kinematic statuses, the joint motor status is an important part that effectively depicts the body pose of a robot. Therefore, kinematic parameter pre-processing focuses on joint kinematic data. For each joint, \( J_i \), of the whole body of a robot, a joint motor status, \( V_i \), is acquired. Thus, when presented, a robotic dance pose is expressed by \( V_1, V_2, \ldots, V_N \). (Note: Our proposed approach uses the dance formalization of a humanoid robot (HRDF) [11] as a base.)

3.2.2. Image pre-processing

Image pre-processing includes three phases: automatic target location, target segmentation, and shape extraction. The prerequisite for image pre-processing is that a humanoid robot has unique colour blocks on the important parts of its body (such as head, shoulder, hand, foot, leg, etc.). The colour blocks differ from the embodied environment of the robot [12].

When a humanoid robot observes its own dance pose in a mirror, the robot’s visual cameras capture the mirror image of the robotic dance pose. Thus, the original image is acquired. Based on the spatial distribution information of the colour blocks, in the automatic target location phase, a relatively precise position of the robot is found in the original image, and a suitable rectangle (a bounding box) is formed to enclose the robot. Notably, the implementation of automatic target location, in this paper, uses the corresponding algorithm of [12]. More concretely, the specific colour (the colour of the above colour blocks) is filtered from the original image, and the other colours are set as background colour (black colour); then, according to the spatial distribution of the specific colour on the original image, an approximate minimum enclosing rectangle (AMER) is identified to contain all the pixels that have the specific colour; finally, a humanoid robot, which has unique colour blocks on the important parts of its body, can be located in the original image, and the above AMER is marked as a suitable rectangle (a bounding box).

The target segmentation phase uses the GrabCut algorithm [19] to extract the sub-image of a robotic dance pose from the original image. As an interactive foreground extraction method, the GrabCut algorithm must know where the foreground is. Consequently, the above rectangle (bounding box), which is determined in the automatic target location phase and encloses the robot, is regarded as the input of the GrabCut algorithm.

Based on the sub-image of a robotic dance pose, which is the output from the target segmentation, the shape extraction phase extracts three types of shape images: region, contour, and spatial distribution of colour blocks. Notably, the implementation of extraction algorithm for the region and contour images, in this paper, uses the corresponding algorithm, see Fig. 5 in [12]. More concretely, by using morphological digital image processing technology, the sub-image of a robotic dance pose is converted to a black-and-white image firstly, and then a series of operations (fill holes, corrode, dilate) are successively implemented on the black-and-white image, and finally the region and contour images are extracted respectively.

Furthermore, by filtering the specific colour (the colour of the above colour blocks) from the sub-image of a robotic dance pose, and setting the other colours as background colour (black colour), the spatial distribution image of the colour blocks can be extracted; plainly, the extraction algorithm is similar to the early processing algorithm for the automatic target location [12] (see paragraph 2, Section 3.2.2), and the difference lies in the different source objects (the original image, and the sub-image of a robotic dance pose).

Fig. 2 shows the following sample images generated in the image pre-processing stage: original image, image as a result of automatic target location, image as a result of target segmentation, region image, contour image, and the spatial distribution image of colour blocks.

3.3. Feature extraction

Based on the results of the pre-processing stage, the feature extraction stage focuses on the following four features: kinematic, region, contour, and spatial distribution of colour block. (Note: Each feature is specified as one or more representative features in this paper. Details are described in the following subsections.)

![Fig. 1. The proposed framework.](image-url)
3.3.1 Kinematic feature

As mentioned in the kinematic parameter pre-processing section (Section 3.2.1), joint kinematic status (focused on in our approach) is an important part that effectively depicts the body pose of a robot. Therefore, each of the joint motors of a biped humanoid robot is abstracted as a Joint Kinematic Feature (JKF).

If a biped humanoid robot has \( N \) joint motors on its whole body, there are \( N \) joint kinematic features \( \{ JKF_1, JKF_2, \ldots, JKF_N \} \). When a humanoid robot presents a dance pose, the status \( (V_1, V_2, \ldots, V_N) \) for each of the \( N \) joint motors is acquired and regarded as an original instance of \( N \) joint kinematic features \( \{ JKF_1, JKF_2, \ldots, JKF_N \} \).

3.3.2 Region feature

Based on the region image, region features are extracted from the basic region shape features, which reflect the holistic characteristics of the regional appearance. They include four parts, which are defined as follows \([12]\):

1. **Eccentricity Feature (EF)**: the eccentricity of the ellipse that has the same second-order central moments with the region of the robotic dance pose;
2. **Density Feature (DF)**: the ratio of the square of the regional perimeter to the regional area;
3. **Rectangularity Feature (RF)**: the area ratio of the robot ontology region to its Minimum Enclosing Rectangle (MER);
4. **Aspect Ratio Feature (ARF)**: the ratio of the width of the MER to the height of the MER.

3.3.3 Contour feature

Among the various typical methods for recognition of 2-D shapes, Fourier descriptors are verified to have the best performance \([20]\). Therefore, we use centroid distance based Fourier descriptors as a contour feature, which characterizes the contour of a robotic dance pose.

Based on the region image, the coordinate of the centroid of the whole region of a robotic dance pose can be calculated, assuming it is \( P_{RC} (X_{RC}, Y_{RC}) \). After the corresponding contour image is acquired, equidistance sampling is adopted for the contour of a robotic dance pose. Suppose there are \( T \) sampling points on the contour and each is expressed as \( P_{FD-i} (x_{FD-i}, y_{FD-i}) \) \( (i \in [1, T]) \). Centroid distance, defined as the distance between a sampling point of contour, \( P_{FD-i} \), and the centroid, \( P_{RC} \), is expressed as follows:

\[
\text{CentroidDistance}(P_{FD-i}, P_{RC}) = \sqrt{(x_{FD-i} - X_{RC})^2 + (y_{FD-i} - Y_{RC})^2}, \quad (i \in [1, T]);
\]

(1)

The sampling points (blue points) and the regional centroid (red point) are shown in Fig. 3, where a yellow line indicates an example of the centroid distance. Thus, the centroid distances of \( T \) sampling points can be calculated, assuming they form a contour shape expression function \( R \), as follows:

\[
R(i) = \text{CentroidDistance}(P_{FD-i}, P_{RC}), \quad (i \in [1, T]);
\]

(2)

Then, the discrete \( R(i) \) is operated on with a Fourier transformation:

\[
b(v) = \frac{1}{T} \sum_{i=0}^{T-1} R(i) \exp(-j2\pi vi/T), \quad v = 0, 1, 2, \ldots, T-1;
\]

(3)

These discrete Fourier coefficients (Fourier descriptors), once acquired, must be normalized further. The normalized Fourier descriptors, based on centroid distance, are defined as follows:

\[
\text{CDFD}(u) = \frac{\|b(u)\|}{\|b(0)\|}, \quad u = 1, 2, \ldots, T-1;
\]

(4)
The diagram of centroid distance (the red point: the regional centroid, the blue points: the sampling points on the contour, the yellow line: an example of centroid distance). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Considering that the high-frequency components of the normalized Fourier descriptors always correspond to the details of the contour and their low-frequency components always correspond to the whole shape of the contour, we selected some low-frequency components of the normalized Fourier descriptors (CDFD$_1$, CDFD$_2$, ..., CDFD$_k$) ($k = [T/5]$) as the contour shape feature of a robotic dance pose.

3.3.4. Spatial distribution feature of colour block
When a robotic dance pose is presented, all the joint motors are activated and changed simultaneously. Thus, there may be multiple variations (spatial position, rotation, occlusion, etc.) of the body parts of the robot (especially on the upper body). Also, these variations of the robotic body parts are bound to affect the changes (such as the quantity of the colour blocks, size, position, distance, direction, etc.) in the spatial distribution of the corresponding colour blocks. Moreover, at a fixed angle view, similar robotic dance poses should have similar spatial distributions of the colour blocks, and vice versa. Therefore, an analysis of the spatial distribution of the colour blocks will help to understand the spatial characteristics (e.g., ductility) of a robotic dance pose from its static image. Thus, we regard the spatial distribution of the colour blocks as a spatial feature.

The spatial distribution feature of colour block includes two features: basic spatial distribution and spatial relationship. The basic spatial distribution feature (SDF$_{bas}$) used in this paper is the quantity of colour blocks. Moreover, inspired by the spatial relationship theory of point clusters in Geosciences [21], based on the spatial distribution of colour blocks, we construct a computational method of spatial relationships for spatial point clusters. The centroid of each colour block is calculated and regarded as a spatial point. Thus, all the centroids of the colour blocks form a spatial point cluster. Then, the method computes three kinds of spatial relationship features: topology, direction, and distance.

**Topology relationship feature.** The topology relationship feature, based on the neighbourhood of spatial points, reflects the relative adjacent relationship of spatial points in a 2D space. To express the neighbourhood of spatial points, there are many forms: Voronoi neighbour (which we used), K nearest neighbour, fixed distance neighbour, etc. More specifically, according to a spatial point cluster, a Voronoi diagram is constructed. Then the number of neighbours of each cell acquired from the Voronoi diagram is calculated.

Thus, the topology relationship feature (SDF$_{top}$) is defined as the ratio of the total number of neighbours in the Voronoi diagram to the total number of spatial point clusters. Details of the algorithm are given in Algorithm 1. The extraction diagram of the topology relationship feature is shown in Fig. 4.

**Algorithm 1** The extraction algorithm of topology relationship feature.

1. Read the spatial distribution image of colour block CB$_i$;
2. Convert the RGB colour image CB$_i$ to the single grey image CB$_i$;
3. Binarize the grey image CB$_i$ to the black-and-white image CB$_i$;
4. Corrode the image CB$_i$, and get the result image CB$_i$;
5. Dilate the image CB$_i$, and get the result image CB$_i$;
6. Calculate all the centroids of colour blocks based on CB$_i$;
7. Construct the corresponding Voronoi diagram based on the above centroids;
8. Calculate the Voronoi diagram’s total neighbour number;
9. Compute the topology relationship feature based on its definition.

**Direction and distance relationship features.** In addition to all the centroids of the colour blocks, the extraction of the other two relationship features (direction and distance) also must use the centroid of the whole region of a robotic dance pose, which can be calculated from the region image. Assuming there are M colour blocks acquired from the further processed (dilated) image (See Algorithm 1), all the centroid coordinates of the colour blocks are expressed by $P_i(x_i, y_i)~[i \in [1,M]]$ in an image coordinate system (XOY). Thus, the mean centre of all the centroid coordinates of the colour blocks $P_{MC}(X_{MC}, Y_{MC})$ are computed as follows:

$$P_{MC}(X_{MC}, Y_{MC}) = \left( \frac{\sum_{i=1}^{M} x_i}{M}, \frac{\sum_{i=1}^{M} y_i}{M} \right).$$ (5)

In the same image coordinate system, (XOY), assume the centroid coordinate of the whole region of a robotic dance pose is $P_{RC}(X_{RC}, Y_{RC})$.

The direction relationship feature, (SDF$_{dir}$), which reflects the deviative degree of direction between $P_{MC}$ and $P_{RC}$, is defined as the angle, $\theta$, between the vector, $(P_{RC}, P_{MC})$, and the positive X axis, (See Formula (6)).

$$\theta = \arctg\left( \frac{Y_{MC} - Y_{RC}}{X_{MC} - X_{RC}} \right);$$ (6)

The distance relationship feature, (SDF$_{dis}$), which reflects the deviative degree of distance between $P_{MC}$ and $P_{RC}$, is defined as the Euclidean distance between $P_{MC}$ and $P_{RC}$ (See Formula (7)).

$$EuclideanDistance(P_{MC}, P_{RC}) = \sqrt{(X_{MC} - X_{RC})^2 + (Y_{MC} - Y_{RC})^2};$$ (7)

Fig. 5 shows a diagram of relationship features of direction and distance.

3.4. Feature selection and combination

In the feature extraction stage, four kinds of features (kinematic, region, contour, and spatial distribution of colour block) are extracted respectively. From a certain aspect, each kind of feature portrays the intrinsic essence of a robotic dance pose (See Section 5.1). These different kinds of features provide for diversity in the feature selections and combinations, and the different feature combinations may provide for different aesthetic effects in the robotic dance poses.

More specifically, the above four kinds of features provide for a total of fifteen feature combinations. However, because there
Fig. 4. The extraction diagram of topology relationship feature: (a) the spatial distribution image of colour block; (b) the single grey image; (c) the black-and-white image; (d) the corroded image; (e) the dilated image; (f) the centroid marked image; (g) the generated Voronoi diagram; (h) the overlapped image that the Voronoi diagram (subfigure (g)) is superimposed on the dilated image (subfigure (e)).

Fig. 5. The diagram of relationship features of direction and distance (the red point: \( P_{RC} \), the green point: \( P_{MC} \), the angle \( \theta \): the direction relationship feature, the yellow line: the distance relationship feature). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

may be conflicts among the different features (See Section 5.3), the more features included in a feature combination does not mean the better. Therefore, to form a feature combination, some suitable features must be picked out. However, a good feature combination cannot easily be found directly; it can be determined only by experiments. Thus, with the expectation to find some good feature combinations, we tested the above fifteen feature combinations.

3.5. Feature fusion

In the feature fusion stage, for each feature combination, the included features form a mixed feature. The mixed feature, which portrays the intrinsic essence of a robotic dance pose from many different aspects (See Section 5.1), must describe the robotic dance pose more comprehensively.

In this paper, the fullest mixed feature is expressed by \( \{JKF_1, JKF_2, \ldots, JKF_N, EF, DF, RF, ARF, CDFD_1, CDFD_2, \ldots, CDFD_k, SDF_{bas}, SDF_{top}, SDF_{dir}, SDF_{dis}\} \), \( k < \lfloor T/5 \rfloor \). However, whether the fullest mixed feature results in the best aesthetic effects must be tested experimentally. Likewise, the best mixed feature must also be determined experimentally. Furthermore, the feature fusion stage is regarded as the basis for the machine learning stage, and the former prepares a sample space for the latter.

3.6. Machine learning

The task in the machine learning stage, based on a fixed feature determined in the feature fusion stage, is to use some specific method of machine learning to train a machine aesthetics model. The trained machine aesthetics model, which possesses the aesthetic ability of a human dance expert, autonomously evaluates the aesthetics of a robotic dance pose.

Taking into consideration that the training of the machine aesthetics model is a supervised learning procedure, human dance experts are invited to evaluate all the robotic dance poses they observe and are asked to provide a label space for machine learning. Each robotic dance pose is expressed as an instance of a mixed feature, and it is labelled with an aesthetic evaluation (good/bad). So each robotic dance pose is described completely as a sample by the above instance and label. Thus, after an example set of robotic dance poses is prepared, a machine aesthetics model can begin to be trained.

Although there are many machine learning methods to choose for training the machine aesthetics model, it is unclear which
kind of machine learning method is more suitable and effective in artistic cognition aesthetics [11]. Thus, in our work, we chose several mainstream machine learning methods to (1) verify their machine aesthetic effects and (2) find out the best machine learning method for a suitable fixed feature. When an effective machine aesthetics model is realized, a humanoid robot can use it to automatically evaluate the aesthetics of its own dance poses. Thus, the autonomous creation of robotic choreography can be implemented further.

4. Experiments

The proposed approach was tested on a simulated NAO robot. The simulated environment that was used is a Webots 7.4.1 simulator. The machine learning platform that was used is Weka 3.6.

When presenting a dance pose in a Webots simulator, a simulated NAO robot perceives its joint motor data (internal kinestat) directly, and the picture shown in the “Simulation View” area of the Webots simulator is regarded as the original “mirror” image of the presented robotic dance pose that the robot itself observes in the “mirror”.

Our experiments used 500 randomly generated robotic dance poses of Chinese Tibetan Tap. These randomly generated dance poses have two constraints: (1) they combine the innovativeness and preservation of dance characteristics of human [11]; (2) the dance pose, which causes a robot to fall, is excluded. Among the two constraints, the former ensures that the robotic dance poses of Chinese Tibetan Tap are generated; and the latter shows that the dance pose, which causes a robot to fall, has no sense of beauty, and is meaningless to continue to be an experimental object. Furthermore, this paper uses the random generation method of [11,22], which contains these steps as follows: (1) the whole body of a humanoid robot must be divided into several limbs \((l_1, l_2, ..., l_k)\), according to the characteristics of Chinese Tibetan Tap; (2) a set of dance elements for each limb is built by imitation and imagination; (3) by assembling several dance elements that come from all of limbs (the whole body), a dance pose is generated randomly.

For supervised learning, a Chinese folk dance expert labelled each robotic dance pose with an aesthetic category (good/bad). A NAO robot has 26 joints in its whole body. Considering that hands always maintain a naturally relaxed state when human dancers perform Chinese Tibetan Tap, the two hand joints (\(\{\text{LHand, RHand}\}\)) of our NAO robot were kept fixed [12]. We used the remaining 24 joints of our NAO robot to characterize the kinematic features of a robotic dance pose, resulting in 24 joint kinematic features, \((JKF_1, JKF_2, ..., JKF_{24})\). Based on the contour image, the number of sampling points was set at 800 (\(T = 800\)). Thirty low-frequency components of normalized Fourier descriptors, based on centroid distance \((CDFD_1, CDFD_2, ..., CDFD_30)\), were used as the contour features of the robotic dance poses.

Overall, in a robotic dance pose, there are four kinds of features: kinematic \((JKF_1, JKF_2, ..., JKF_{24})\); region \((EF, DF, RF, ARF)\); contour \((CDFD_1, CDFD_2, ..., CDFD_{30})\); and spatial distribution of colour blocks \((SDF_{bas}, SDF_{typ}, SDF_{dir}, SDF_{dis})\). The four kinds of features have fifteen feature combinations. Based on each feature combination, eight kinds of mainstream machine learning methods were used for machine aesthetics, and ten cross-validation methods were used for evaluation. The detailed results for machine learning are shown in Table 1.

To simplify the feature category expressions in Table 1, kinematic features \((JKF_1, JKF_2, ..., JKF_{24})\); region features \((EF, DF, RF, ARF)\); contour features \((CDFD_1, CDFD_2, ..., CDFD_{30})\); and spatial distribution features of colour block \((SDF_{bas}, SDF_{typ}, SDF_{dir}, SDF_{dis})\) are expressed as \(F_1, F_2, F_3, \) and \(F_4\), respectively.

Moreover, the highest correct ratio of a machine learning method, based on the different feature combinations, is marked with a red background in Table 1. As seen from the final results for machine learning, based on the feature combination \(F_1 + F_2 + F_4\) (kinematic + region + spatial distribution of colour block), the highest correct ratio for aesthetic evaluation is 81.8%.

5. Discussion

5.1. Feature semantics

From the proposed framework (shown in Fig. 1), information flows from visual and non-visual channels. Then, four kinds of features (kinematic, region, contour, and spatial distribution of colour block) are extracted. From a certain aspect, each kind of feature portrays the intrinsic essence of a robotic dance pose. Feature combination provides a more complete description for a robotic dance pose. Each kind of feature and its semantics are described in detail within Table 2.

5.2. Feature selection and combination

As seen in Table 1, among the aesthetic effects based on a single feature \((F_1, F_2, F_3, \) and \(F_4)\), the kinematic feature, \(F_1\), with the Machine Learning Method, AD Tree, has the highest correct ratio for aesthetic evaluation (80.8%). The reason may be the kinematic feature reflects the self-perception of the internal kinestat of the robot, which, from the perspective of ontology, portray the essential characteristics of a robotic dance pose.

Also, the weak aesthetic effects based on a single visual feature \((Table 1: F_2, F_3, \) and \(F_4)\) may be caused by the following reasons: (1) A single visual feature is insufficient to describe a robotic dance pose; (2) A more suitable visual feature has not been found. To solve these problems, the following measures should be considered: (1) Use a combination of several visual features to describe a robotic dance pose more completely; (2) To find a more suitable visual feature, expand the category of visual features (e.g. depth feature) or change the specific forms of a visual feature (e.g. use a wavelet descriptor as a specific implement of a contour feature).

Our experiments confirmed the aesthetic effects based on the following combinations of multiple visual features: \(F_2 + F_3, F_2 + F_4; F_3 + F_4; F_2 + F_3 + F_4.\) Among them, the feature combination, \(F_2 + F_4\) (contour + spatial distribution of colour block), with the Machine Learning Method, Random Forest, has the highest correct ratio for aesthetic evaluation (75.8%). Although the correct ratio (75.8%) could be greater, it is better than the correct ratios based on the single visual features \((F_2, F_3, \) and \(F_4)\), which illustrates that a suitable combination of several visual features improves the aesthetic evaluation of a robotic dance pose.

Generally, in our experiments, the feature combination, \(F_1 + F_2 + F_4\) (kinematic + region + spatial distribution of colour block), with the Machine Learning Method, AD Tree, has the highest correct ratio for aesthetic evaluation (81.8%). The ratio of 81.8% is still better than the 80.8% for the aesthetic evaluation based on the single kinematic feature, \(F_1\), with the AD Tree learning method, which also illustrates that a suitable feature combination improves the aesthetic evaluation of a robotic dance pose. Among the eight methods of machine learning, four methods (Bayesian Logistic Regression, AD Tree, Random Forest, and Bagging) with the feature combination \(F_1 + F_2 + F_4\) (rather than the other feature combinations) achieved, for aesthetic evaluation, the highest correct ratios of 80.6%, 81.8%, 80.4%, and 81.6%, respectively. Therefore, as shown in our experiments, the feature combination, \(F_1 + F_2 + F_4\), is the best to completely and accurately describe a robotic dance pose.

For the feature combination, \(F_1 + F_2 + F_4\) (kinematic + region + spatial distribution of colour block), the kinematic features are
Table 1
The effect comparison on different machine learning methods based on different feature combination.

<table>
<thead>
<tr>
<th>Feature combination</th>
<th>Machine learning method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive Bayes</td>
</tr>
<tr>
<td>$F_1$</td>
<td>73%</td>
</tr>
<tr>
<td>$F_2$</td>
<td>74.2%</td>
</tr>
<tr>
<td>$F_3$</td>
<td>67.4%</td>
</tr>
<tr>
<td>$F_4$</td>
<td>73.8%</td>
</tr>
<tr>
<td>$F_1 + F_2$</td>
<td>74%</td>
</tr>
<tr>
<td>$F_1 + F_3$</td>
<td>73.8%</td>
</tr>
<tr>
<td>$F_1 + F_4$</td>
<td>73.4%</td>
</tr>
<tr>
<td>$F_2 + F_3$</td>
<td>71.8%</td>
</tr>
<tr>
<td>$F_2 + F_4$</td>
<td>74.4%</td>
</tr>
<tr>
<td>$F_3 + F_4$</td>
<td>69.4%</td>
</tr>
<tr>
<td>$F_1 + F_2 + F_3$</td>
<td>76.4%</td>
</tr>
<tr>
<td>$F_1 + F_2 + F_4$</td>
<td>74%</td>
</tr>
<tr>
<td>$F_1 + F_3 + F_4$</td>
<td>74.6%</td>
</tr>
<tr>
<td>$F_2 + F_3 + F_4$</td>
<td>71.4%</td>
</tr>
<tr>
<td>$F_1 + F_2 + F_3 + F_4$</td>
<td>76.8%</td>
</tr>
</tbody>
</table>

Table 2
Feature semantics.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinematic feature</td>
<td>The internal motion skeleton of a dance pose;</td>
</tr>
<tr>
<td>Region feature</td>
<td>The overall silhouette of a dance pose;</td>
</tr>
<tr>
<td>Contour feature</td>
<td>The overall peripheral shape of a dance pose;</td>
</tr>
<tr>
<td>Spatial distribution feature of colour block</td>
<td>The spatial variation and ductibility of a dance pose, &amp; the motorial tendency of a dance pose.</td>
</tr>
</tbody>
</table>

from the non-visual channel. The region and spatial distribution features of colour blocks are from the visual channel. Thus, the experimental results also illustrate that, like for human dancers, information fusion on the visual and non-visual channels improves the machine aesthetics effect of robotic dance poses.

However, the correct ratio of 81.8% (reported above) is not high enough. Compared with the value of 80.8% for kinematic features ($F_1$), the value for the highest correct ratio for the best feature combination ($F_1 + F_2 + F_3$), at 81.8%, is only incrementally higher by 1%. The root cause for this phenomenon is that the effect visual features have on the correct rate of machine aesthetics is limited.

More specifically, the reasons are as follows: (1) The category of visual features that is used is inadequate (e.g. absence of depth feature) to describe a robotic dance pose; (2) The specific form of visual feature used (e.g. the centroid distance based Fourier descriptors are used as contour features) is unsuitable for describing a robotic dance pose; (3) There is a conflict among the features in the feature combination (See Section 5.3); (4) In the target segmentation phase, the existence of robotic shadow in the original image creates a certain distortion in the processing of the results of the GrabCut algorithm. The above problems can be solved by the following measures: (1) Expand new categories of visual features (e.g. depth feature) to describe a robotic dance pose; (2) Find new realization forms of visual features (e.g. use a wavelet descriptor as a newly specific contour feature) to describe a robotic dance pose more suitably; (3) Find more powerful or suitable feature combinations by eliminating the conflict among features in the feature combination; (4) By improving the GrabCut algorithm, separate the robot ontology and its shadow more accurately in the target segmentation phase.

5.3. Feature conflict

In this paper, the extraction of multiple features brings the diversity of feature combination. However, conflicts that may exist in a feature combination will reduce the effects of machine learning. From the experimental results, two kinds of feature conflicts existed: internal conflict in the same feature category and external conflict among feature categories. For the visual features (region, contour, and spatial distribution of colour block), the conflicts were internal. For example, the correct ratio for aesthetic evaluation from Bayesian Logistic Regression based on the single feature, $F_2$ (region), is 74.6%. Using the same method, the correct ratios, based on the feature combination, $F_2 + F_3$ (region + contour), and the feature combination, $F_2 + F_3$ (region + spatial distribution of colour block), are 73.8% and 74%, respectively.

In this work, because the kinematic and visual features are acquired from two different information channels, they are regarded as two different feature categories. Moreover, external conflict exists between them. For example, the correct ratio for aesthetic evaluation from Bagging based on the single feature $F_1$ (kinematic), is 80.4%. Using the same method, the correct ratios based on the feature combination $F_1 + F_2$ (kinematic + contour) and the feature combination $F_1 + F_2$ (kinematic + region + contour) are 76% and 79.6%, respectively.

Therefore, due to the existence of feature conflicts, the more features included in a feature combination does not mean the better. From our experiments, we found that, instead of the most full feature combination, $F_1 + F_2 + F_3 + F_4$ (kinematic + region + contour + spatial distribution of colour block), the best feature combination to describe a robotic dance pose is $F_1 + F_2 + F_4$ (kinematic + region + spatial distribution of colour block). Furthermore, our future work will include research on how to find suitable features to form a feature combination.

5.4. Comparison with the related methods

As mentioned in Section 2, the existing literatures involve the aesthetic evaluation of robotic dance poses, and the existing methods focus on two ways: human subjective aesthetics [8–10] and the machine learning based method [11, 12]. Table 3 compares the existing methods with the approach proposed in this paper. It can be observed that our approach is superior to the others. We believe this is achieved by a good feature combination and an appropriate machine learning method.
Meanwhile, some facts can also be found in Table 3: (1) in portraying a robotic dance pose, a feature combination is always better than single feature; (2) despite the complexity, the increase of feature types promotes the diversity of feature combination, and then brings possibility to improve the correct ratio of aesthetic evaluation; (3) the spatial distribution feature of colour block is better than the contour feature, so the former can replace the latter to achieve the better results (e.g. the best feature combination in our approach has not included the contour feature); by using the best machine learning method (ADTree) on the corresponding best feature combination (both our approach and [12]), our approach has achieved the better results than [12]; (4) for the automatic aesthetics evaluation of robotic dance poses, ADTree is a suitable method of machine learning, and we believe this is because ADTree is effective in solving the binary classification problem of aesthetics.

5.5. Influence on robotic choreography

From the dance formalization of a humanoid robot (HRDF) [11], dance poses are the basis of a robotic dance, and a robotic dance is regarded as a sequence of dance pose. Meanwhile, dance pose, as a static body shape, expresses emotion, character, feeling, meaning and theme [4]. Thus, good dance poses help to improve the quality of robotic dance, and then create good robotic choreography. By deconstructing the three requirements of good robotic choreography [1] in dance poses, good dance poses should meet the same three requirements simultaneously:

(1) preservation of the characteristics of human dance;
(2) innovativeness of the dance; and
(3) accordance with human aesthetics [1].

This paper focuses on the last requirement, and aims to make a robot possess the aesthetic ability of human dance experts on its own dance poses, which is achieved by machine learning. Notably, the proposed approach in this paper is inspired by the mature aesthetic experience of human beings, and imitates the human dance behaviour. Furthermore, the random generation method of robotic dance pose [11,22] ensures that the dance poses of Chinese Tibetan Tap, used in our experiments, combine the innovativeness and preservation of dance characteristics of human (See Section 4).

Additionally, considering that the generation procedure of robotic choreography, based on a sequence of dance pose, is dynamic, interpolation between the adjacent dance poses is necessary. Between dance poses, interpolation not only ensures the smoothness of transition, but also brings the variability of robotic limbs in space trajectory. Moreover, each interpolation algorithm brings the different spatial trajectories for movement, so it also brings the different aesthetic effect. Which kind of interpolation algorithm can bring better aesthetic effect needs further study. How to make a robot create good robotic choreography autonomously based on the good dance poses, picked out by the robot itself (with our approach), remains to be solved in the future.

5.6. Suitable approach range

In our research, we used a biped humanoid robot: the NAO robot regarded as a prototype robot. The prerequisite for the proposed approach is that, on the important parts of its body (such as head, shoulder, hand, foot, leg, etc.), a humanoid robot has unique colour blocks that differ from the embodied environment of the robot [12]. Based on the prerequisite, in our proposed approach the automatic target location was achieved and the spatial distribution image of the colour blocks was acquired. Then, the whole framework was implemented.

In the area of robotic dance, there are many kinds of biped humanoid robots that can be used as research carriers. However, not every biped humanoid robot has unique colour blocks on the important parts of its body. Biped humanoid robots that meet this requirement include: NAO robot [23], HRP-2 robot [24], Robonova robot [6,25,26], etc. Moreover, Biped humanoid robots that do not meet this requirement include: Hubo [25,26], ASIMO [27], QRIO [28], DARwIn-OP [29], etc. Biped humanoid robots that do not meet this requirement clearly cannot be used directly in our approach. Nevertheless, the problem can be solved by pasting colour stickers on the important body parts of these biped humanoid robots. Thus, more biped humanoid robots can be suitable for the proposed approach.

Furthermore, the prerequisite of the proposed approach requires that the colour blocks, which exist on the important body parts of a biped humanoid robot, differ from the embodied environment of the robot. If both are the same, or similar colours, the proposed approach may not work normally. This problem can be solved by changing the colours on the colour blocks on the important body parts of the robot. Specifically, after selecting a unique colour that differs from the embodied environment of the robot, colour stickers with the unique colour are pasted on the important body parts of the robot. Thus, the suitable range for the proposed approach is expanded further.

6. Conclusion

This paper presents a new feature fusion approach to automatic machine aesthetics evaluation of robotic dance poses. Its prerequisite is that a humanoid robot has unique colour blocks on its important parts of body, and the colour blocks differ from the robot’s embodied environment [12]. By processing information from visual and non-visual channels, four kinds of features (kinematic, region, contour, and spatial distribution of colour block) are extracted respectively. After the feature selection, combination and fusion, the machine learning method is used for training a machine aesthetics model of robotic dance poses, and the automatic aesthetic judgement on robotic dance poses is achieved.

The proposed approach has been validated via simulation by instantiating specific features under the four kinds of feature (kinematic, region, contour, and spatial distribution of colour block). The experimental results show that the best feature combination for describing a robotic dance pose is \( F_1 + F_2 + F_3 \) (kinematic + region + spatial distribution of colour block). Meanwhile, ADTree has achieved the highest correct ratio of aesthetic evaluation on robotic dance poses (81.8%) based on the best feature combination. Furthermore, a suitable feature combination always brings a better aesthetic effect on robotic dance poses than single feature.

Our future work will focus on four aspects: (1) to find a better specific feature respectively under each of the above four kinds of feature (e.g. use inertial features as a specific form of kinematic feature [30]), so that a robotic dance pose can be portrayed better from a certain aspect; (2) to find more powerful or suitable feature combinations to portray a robotic dance pose from a whole; (3) to find more effective methods of machine learning to improve the correct ratio of aesthetic evaluation on robotic dance poses; (4) to apply the proposed approach on a real NAO robot so that autonomous aesthetics evaluation can be conducted on its own dance pose.

Acknowledgements

The authors would like to thank the reviewers for their invaluable comments and suggestions, which greatly helped to improve the presentation of this paper.

This work was supported by National Natural Science Foundation of China (Grant No. 61662025, 61601200), and the Research Foundation of Philosophy and Social Science of Hunan Province,
Table 3  
The comparison between the existing methods and our approach.

<table>
<thead>
<tr>
<th>Feature type involved</th>
<th>Specific feature</th>
<th>Machine learning method involved</th>
<th>Highest correct ratio</th>
<th>Best feature combination</th>
<th>Best machine learning method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kinematic</td>
<td>Joint</td>
<td>N/A</td>
<td>N/A</td>
<td>kinematic + region +</td>
<td>AD tree</td>
</tr>
<tr>
<td>kinematic</td>
<td>Joint</td>
<td>Machine learning based method</td>
<td>71.6667%</td>
<td>contour</td>
<td></td>
</tr>
<tr>
<td>kinematic</td>
<td></td>
<td>SVM, RBF Network, AD Tree</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kinematic</td>
<td></td>
<td>Machine learning based method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Naive Bayes, Bayesian Logistic Regression, SVM, RBF Network, AD Tree, Random Forest, Voted Perceptron, KSstar, DTNB, Bagging 81.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Naive Bayes, Bayesian Logistic Regression, SVM, RBF Network, AD Tree, Random Forest, Voted Perceptron, Bagging 81.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References


Conflict of interest

The authors declare that they have no conflict of interest.
Hua Peng received his Ph.D. Degree from Xiamen University, China in 2016. He is now a lecturer in the Department of Computer Science and Engineering, Shaoxing University, Shaoxing, and also a lecturer in the College of Information Science and Engineering, Jishou University, China. His research interests includes brain-like intelligent systems, human–robot interaction, networked service robots, and machine learning.

Jing Li received her M.S. Degree from Yunnan University, Kunming, China. She is currently an associate professor in the College of Musicology, Shaoxing University, China. Her research activity is focused on Chinese national dances, and robotic dance.

Huosheng Hu received his Ph.D. degree from Oxford University in the UK. He is now a Professor in School of Computer Science & Electronic Engineering at the University of Essex. His research interests include behaviour-based robotics, human–robot interaction, embedded systems, mechatronics, learning algorithms, and networked service robots.

Liping Zhao received her Ph.D. degree in control science and engineering from Tongji University, Shanghai, China, in 2017. She is now with the Department of Computer Science and Engineering, Shaoxing University, Shaoxing, China. She was with the College of Mathematics, Physics and Information Engineering, Jiaxing University from 2009 and June 2017. Her current research interests include the areas of screen content coding.

Sheng Feng received his Ph.D. degree in pattern recognition and intelligent system from Northeastern University, Shenyang, China in 2017. He is currently a Lecturer with the Department of Computer Science and Engineering, Shaoxing University, Shaoxing, China. His research interests are in the area of intelligent robot, computer vision and wireless sensor networks.

Keli Hu received his Ph.D. degree from Chinese Academy of Sciences, Shanghai, China in 2014. He is currently a lecturer with the Department of Computer Science and Engineering, Shaoxing University, Zhejiang, China. His research interests include artificial intelligence, pattern recognition, computer vision and image processing.