International Journal of Social Robotics (2020) 12:5–20
https://doi.org/10.1007/s12369-019-00535-w

Multimodal Information Fusion for Automatic Aesthetics Evaluation of Robotic Dance Poses

Jing Li[1] · Hua Peng[2,3,4] · Huosheng Hu[4] · Zhiming Luo[5] · Chao Tang[6]

Accepted: 15 February 2019 / Published online: 22 February 2019
© Springer Nature B.V. 2019

Abstract
Aesthetic ability is an advanced cognitive function of human beings. Human dancers in front of mirrors estimate the aesthetics of their own dance poses by fusing multimodal information (visual and non-visual) to improve their dancing performances. Similarly, if a robot could perceive the aesthetics of its own dance poses, the robot could demonstrate more autonomous and humanoid behavior during robotic dance creation. Therefore, we propose a novel automatic approach to estimate the aesthetics of robotic dance poses by fusing multimodal information. From the visual channel, the shape features (including eccentricity, density, rectangularity, aspect ratio, Hu-moment Invariants, and complex coordinate based Fourier descriptors) are extracted from an image; from the non-visual channel, joint motion features are obtained from the internal kinestate of a robot. The above two categories of features are fused to portray completely a robotic dance pose. To automatically estimate the aesthetics of robotic dance poses, the following ten machine learning methods are deployed: Naive Bayes, Bayesian logistic regression, SVM, RBF network, ADTree, random forest, voted perceptron, KStar, DTNB, and bagging. Experimental results show the feasibility and good performance of the proposed mechanism, which was implemented in a simulated robot environment. The highest correct ratio of aesthetic evaluation is 81.6%, which comes from the ADTree, based on the above mixed features (joint + shape).

Keywords Automation · Machine aesthetics · Robotic dance pose · Feature fusion · Machine learning

1 Introduction

Robotic dance is an interesting research area and attracts many researchers to work on its development in terms of interaction, imitation, coordination and autonomy by using artificial intelligence and human–robot interaction technology [1–3]. As the fundamental part of robotic dance, dance pose is a static body shape and expresses emotion, character, feeling, meaning and theme [4]. In the existing research, dance pose presents several different forms, such as stopping posture [5], key-pose [4, 6], gesture [4, 7] and posture [8]. Despite diverse forms of robotic pose, its essence is unity and plays an important role in robotic dance.

Robotic dance has been classified into four categories in [2], namely cooperative human–robot dance, imitation of human dance motions, synchronization for music, and

✉ Hua Peng
6195340@qq.com

Jing Li
40296448@qq.com

Huosheng Hu
hhu@essex.ac.uk

Zhiming Luo
zmluo_xmu@qq.com

Chao Tang
tangchao@hfuu.edu.cn

[1] Academy of Arts, Shaoxing University, Shaoxing 312000, China

[2] Department of Computer Science and Engineering, Shaoxing University, Shaoxing 312000, China

[3] College of Information Science and Engineering, Jishou University, Jishou 416000, China

[4] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK

[5] Cognitive Science Department, Xiamen University, Xiamen 361005, China

[6] Department of Computer Science and Technology, Hefei University, Hefei 230601, China

creation of robotic choreography. However, only robotic choreography creation has aesthetic requirements in accordance with human aesthetics. Based on different research goals, some researchers have explored the aesthetic problem in robotic choreography creation. So far, the explored aesthetic problem involves three kinds of aesthetic objects: robotic dance pose [8–11], robotic dance motion [8–10, 12, 13], and robotic dance [11, 14–18].

As robotic dance is a sequence of dance poses, its aesthetics requirement (accordance with human aesthetics) should be decomposed naturally into these dance poses. Therefore, it is meaningful to explore the aesthetics of robotic dance poses. If a robot could perceive the aesthetics of its own dance poses, it would express more autonomous behavior in robotic choreography creation and promote human–robot interaction. Existing research involves only two aesthetic methods for robotic dance poses: human subjective aesthetics [8–10] and the machine learning based method [11]. For the former, although more accurate aesthetic evaluation results are obtained, human–robot interactions impose a heavy burden on people. For the latter, although people do not need to participate too much in human–robot interactions, it is difficult to build a suitable machine aesthetic model to achieve accurate aesthetic evaluation.

However, so far, the aesthetic method of the robotic dance pose, which draws lessons from the mature aesthetic experience of human beings, has been rarely studied.

A human dancer always actualizes his/her dance pose's aesthetic by integrating multimodal information. For instance, after presenting dance poses before a mirror, human dancers can clearly observe the mirror images of their dance poses and body kinestate. Combining the information, they can make a comprehensive aesthetic judgment on their own dance pose. Inspired by this, a humanoid robot should use the similar mechanism to achieve automatic aesthetics on its own dance pose. However, the following main questions remain: (1) How can a robot integrate multimodal information from two channels of vision and non-visual to make a comprehensive aesthetic judgment of its own dance poses? (2) How can a robot fuse multiple features to understand its own dance poses more completely? (3) Which method can achieve more accurate results on the above aesthetic judgment of robotic dance poses?

Inspired by the corresponding human aesthetics mechanism, to develop the autonomous and humanoid behavior of robots, we propose a new theory of automatic machine aesthetics for robotic dance poses based on multimodal fusion information. More concretely, to analyze robotic dance poses, an automatic image processing method is designed, which extracts useful shape features (including eccentricity, density, rectangularity, aspect ratio, Hu-moment Invariants, and complex coordinate based Fourier descriptors). To portray a robotic dance pose more completely, the shape features are combined with joint features to form mixed features. Then, ten machine learning methods are used to achieve the automatic aesthetic judgment for robotic dance poses.

The main contributions of this paper are as follows:

- From the perspective of self-aesthetic understanding of dance pose, this paper explores a feasible way to develop the robot's autonomous intelligence by imitating human dance behavior.
- By fusing multimodal information (visual and non-visual), this paper proposes a novel automatic approach to estimate the aesthetics of robotic dance poses. The approach improves the autonomy and cognitive ability of the robot to a certain extent.
- The mixed features (joint + shape), proposed by this paper, can characterize a robotic dance pose well. Moreover, based on the mixed features, ADTree has been verified as an effective machine learning method to achieve more accurate aesthetics evaluations of robotic dance poses.

The rest of the paper is organised as follows. Section 2 outlines the current works that are related to this research. A detailed explanation of the whole mechanism is presented in Sect. 3, including five parts: the whole framework, preprocessing, feature extraction, feature fusion, and machine learning. Section 4 describes the complete experimental process, and shows the experimental results via simulation. Our mechanism is further explained from four aspects in Sect. 5, based on the simulation experimental results. Finally, a brief conclusion and future work are presented in Sect. 6.

2 Related Work

As mentioned in the previous section, aesthetics requirements are mainly related to robotic choreography creation in which a humanoid robot is a carrier. The existing research in this area can be divided into three aspects: robotic dance pose, robotic dance motion and robotic dance, which are listed in the column of "Aesthetics Object" of Table 1. Vircikova and Sincak [8–10] constructed a multi-robot system and designed robotic choreography by using interactive evolutionary computation (IEC). They implemented human subjective aesthetic evaluation on robotic dance pose and robotic dance motion.

For seeking good robotic dance poses that are in accordance with human aesthetics, we present a theory of semi-interactive evolutionary computation (SIEC), which is a population-based searching algorithm [11]. Machine learning, an important stage and supervised learning process,

Table 1 Aesthetics in robotic choreography creation

Method of robotic choreography creation	Robot carrier	Dance	Aesthetics object	References
Interactive evolutionary computation	Nao robot	Unspecified	Robotic dance pose, robotic dance motion	[8–10]
Semi-interactive evolutionary computation	Nao robot	Chinese Tibetan Tap	Robotic dance pose, robotic dance	[11]
Traditional evolutionary computation	Bioloid humanoid robot	Unspecified	Robotic dance motion	[12]
Random generation	Humanoid robot developed by Nirvana Technology	Hip-hop	Robotic dance motion	[13]
Mapping rule	Lego NXT robot	Unspecified	Robotic dance	[14]
Hidden Markov model	Nao robot	Unspecified	Robotic dance	[15–17]
Hidden Markov model	Alpha1 Pro	Unspecified	Robotic dance	[18]

was embedded in SIEC and trained a robot to learn how to accomplish autonomously the aesthetic evaluation of dance poses, thereby giving the robot the ability to possess human aesthetics [11]. Moreover, according to the quality evaluation index and the three features of good robotic choreography, several robotic dances based on those good dance poses were evaluated by aesthetics, and the aesthetic results were acceptable.

Eaton proposed an synthesis approach to create robotic dance choreography based on traditional evolutionary computation (TEC), and built an aesthetic fitness function on robotic dance motions [12]. The fitness function involved the sum of all movement values over all of the joints multiplied by the time that the robot remained standing [12]. It assessed the quality of dance movements, which referred to "performance competence evaluation measure" [19]. Furthermore, Shinozaki et al. [13] designed a robot dance system for investigating the role of robots in entertainment. The system used each Hip-Hop robotic motion as a dance unit, and several dance units were concatenated randomly to form dance choreography. Then, human subjective aesthetic evaluations were conducted on robotic dance motion, and the evaluation items included dynamic, exciting, wonder and smooth, etc.

Oliveira et al. [14] constructed a choreography framework, in which a Lego NXT robot could perform its dance motions in response to the inputs of multimodal events. Moreover, an empiric evaluation was made on robotic dance, and each evaluator was required to fulfill a Likert scaled questionnaire to achieve aesthetic evaluation. The evaluation indexes included: the robot's musical-synchrony, its variety of movements, its human characterization, and the flexibility of the user control over the system, etc. [14].

Manfrè et al. [15] proposed an automatic system for robotic dance creation based on hidden Markov model (HMM). They choose suitable robotic dance motions to be a robotic dance according to the perceived musical rhythm. By calculating the loudness per beat of the inputted music signal, a sequence of music classes was generated and associated and regarded as the HMM's observed sequence. Moreover, each robotic dance motion was regarded as a hidden state of HMM, and the Viterbi algorithm was introduced to find the optimal sequence of robotic dance motions according to the sequence of music classes. Finally, the created robotic dances were evaluated by three professional dancers, and the aesthetic impact of the whole sequence of robotic dance motions had a mean value of 6.33 in the score range [1–10] (10 best).

By integrating this automatic system into the cognitive architecture of a humanoid robot dancer, Augello et al. [16] explored the live performances based on human–robot interaction, among which the creative dance motions were generated to form an improvisational robotic dance. After each performance, the spectators were asked to fill a questionnaire to evaluate the performance. The aesthetic evaluation indexes included four aspects: originality of the choreography, naturalness of the robot–dancers interaction, timing and movements of the robot, evaluation of overall performance [16].

In the same way, the automatic system in [15] was integrated into a computational creativity framework, aiming to drive robotic dance creation [17]. More specifically, Manfrè et al. [17] presented a method of demonstration learning that a Nao robot could learn dance motions by human demonstration, and then the set of dance motions was built as the basis of robotic dance creation. Furthermore, the aesthetic evaluation of robotic dance, given by the audiences, involved three aspects: timing and movements, dance naturalness, and overall artistic value. The aesthetic evaluation results demonstrated that the robotic dance performance was depended on the set of dance motions learned from human demonstration.

Qin et al. [18] proposed a humanoid robot dance system driven by musical structures and emotions. In their system,

phrases were regarded as the basic structural unit of music and dance, and a piece of music was converted into an emotion sequence by the emotion recognition algorithm they designed. Based on this emotion sequence, a hidden Markov model (HMM) was used for searching a matching dance phrase sequence from a predesigned action library. Additionally, a chance method was adopted as a choreography guide. Ten dance students and ten non-dance students (aesthetic evaluators) were invited to evaluate, by using questionnaires, the creation results of the robot dance system. All twenty concluded the robot did a good job dancing to the music [18].

Furthermore, based on feature perception on visual images, Tutsoy et al. constructed rule-based classifiers to recognize facial characteristics [20] and facial emotion [21]. The perceived visual features (facial distance measurements/facial muscle movements) were evaluated with physiognomy science, and the evaluation results showed that the rule-based classifiers performed well. Thus, a machine or a humanoid robot is given the aesthetic cognitive ability to understand human faces. In addition, to imitate human daily behaviors, Gongor et al. [22, 23] presented a sit-to-stand (STS) motion algorithm for humanoid robots. Based on the calculations on kinematic parameters (joint angle states), the algorithm had driven a Nao humanoid robot to achieve autonomous human-like motions.

3 Automatic Machine Aesthetics of Robotic Dance Pose

This section describes the mechanism of automatic machine aesthetics of robotic dance poses based on multimodal information fusion, which contains five parts: the whole framework, pre-processing, feature extraction, feature fusion, and machine learning.

3.1 The Whole Framework

Aesthetic ability is an advanced cognitive function of human beings. For the human aesthetic mechanism of dance pose, a mirror acts as an important tool to help human dancers to observe the visual effect on their dance poses. Moreover, human dancers can perceive simultaneously the movement status of their body parts. By combining these two kinds of information, human dancers could make a comprehensive aesthetic judgment on their own dance poses.

Similarly, a humanoid robot could use such a mechanism to achieve automatic estimation of aesthetics on its own dance poses. More specifically, a humanoid robot uses its "eyes" (visual cameras) to observe its own dance poses in a mirror, and feels its internal kinestate (motor parameters) using its embedded sensors. By combining these two kinds of information, the humanoid robot could make a comprehensive aesthetic judgment on its own dance poses. Therefore, this paper proposes a mechanism of automatic machine aesthetics of robotic dance poses based on multimodal information fusion. Figure 1 shows the whole framework of the mechanism.

When its dance pose is presented before a mirror, the humanoid robot can read its own joint motor status from its embedded sensors (encoders, accelerometers and Gyros), and then extract the corresponding joint features. Meanwhile, it can capture the mirror images of its own dance poses by its cameras, and pre-processed these images in three stages (automatic target location, target segmentation, and shape extraction). Then six kinds of shape features (eccentricity, density, rectangularity, aspect ratio, Hu-moment Invariants, and complex coordinate based Fourier descriptors) are extracted.

Thus, each robotic dance pose is described collectively by the joint feature and shape features (mixed features). A human dance expert will give his/her aesthetic evaluation on

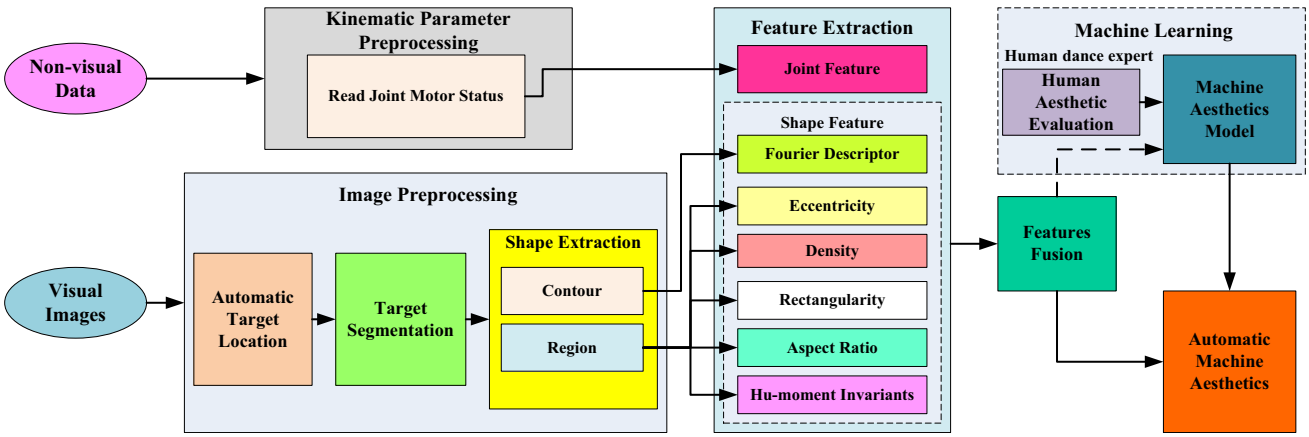

Fig. 1 The proposed framework

a robotic dance pose observed. Both the aesthetic evaluation (label) and the fused features (instance) form an example of the robotic dance pose. When enough samples are acquired, the phase of machine learning will start and a machine aesthetics model of robotic dance pose is trained. Finally, the trained machine aesthetics model is used for automatic aesthetic judgment on a new robotic dance pose when the humanoid robot presents it before a mirror.

3.2 Pre-processing

As shown in Fig. 1, the pre-processing stage includes two processes: kinematic parameter pre-processing and image pre-processing, which come from two different information channels. Notably, the proposed mechanism in this paper uses the dance formalization of humanoid robot (HRDF) [11] as a base. The humanoid robot has unique colour blocks on its important parts of body (such as head, shoulder, hand, foot, leg, etc.), and the colour blocks differ from the robot's embodied environment. Comparing with image pre-processing, kinematic parameter pre-processing is simpler and more convenient. In the kinematic parameter pre-processing process, the humanoid robot reads a joint motor status (V_i) on each joint (J_i) of the whole body, and the presented robotic dance pose is expressed as a vector (V_1, V_2, …, V_S), which is the original kinematic parameter data of the robotic dance pose.

Image pre-processing is divided into three phases: (1) automatic target location; (2) target segmentation; (3) shape extraction. The automatic target location phase is to locate the robot position in the original image captured and determinate a suitable rectangle to enclosure the robot. In the phase of target segmentation, the GrabCut algorithm [24] is used for extracting the sub-image of robotic dance pose from the original image. As the GrabCut algorithm is an interactive foreground extraction method, it requires users to be involved, i.e. informing the foreground by drawing a rectangle on the original image interactively. The shape extraction phase is to extract region and contour on the above sub-image of robotic dance pose, which is output from target segmentation. The more details of these three phases are described in the following subsections.

3.2.1 Automatic Target Location

To automatically locate the robot position in the original image, we build a target location method based on the colour block information of a humanoid robot. The prerequisite of our method requires that a humanoid robot has unique colour blocks on the important parts of its body and the colour blocks differ from the robot's embodied environment. A Nao humanoid robot is used for describing the method. Notably, the humanoid robot always captures its own mirror image of dance poses by its onboard cameras, so the captured original image is the RGB image.

After the original image is acquired, our method firstly finds out the pixels with specific colour, by setting all the pixels without specific colour to be black (background colour). Thus, the specific colour is regarded as foreground colour. Subsequently, the processed image is corroded to eliminate noise, and then dilated to eliminate very small or narrow pixels. Finally, the processed image contains the robot position information, which is described by the position of the foreground colour in the image. To provide the foreground object (the sub image of robotic dance pose) for the stage of target segmentation (the GrabCut algorithm), the robot position must be labelled by a rectangle.

According to the position of foreground colour in the processed image, an approximate minimum enclosing rectangle (AMER) is identified as the input of the GrabCut algorithm in target segmentation. AMER adds a positive bias on the width and height of minimum enclosing rectangle (MER) respectively, aiming to make the robot fall into this range more accurately. In addition, there are double bias increment in the width direction, and a bias increment in the height direction, aiming to eliminating the shadow influence from robot, shown in formulas (1) and (2). The bias can be defined by computing based on the height ratio of MER to original image as shown in formula (3).

$$\text{Width (AMER)} = \text{Width (MER)} + 2 * \text{Bias} \tag{1}$$

$$\text{Height (AMER)} = \text{Height (MER)} + \text{Bias} \tag{2}$$

$$\text{Bias} = \left[\omega * \frac{\text{Height (MER)}}{\text{Height (original image)}}\right] \tag{3}$$

where ω is a constant adjustment parameter. Notably, ω, a positive value, is a whole number multiple of ten pixels. Moreover, ω must be adjusted according to the ductility of the presented dance poses, thereby making the robot fall entirely into AMER. On the premise that the other parameters remain unchanged, the larger ω, the larger AMER, and vice versa. Figure 2 shows a sequence of automatic target location process and Fig. 3 shows the corresponding algorithm. Essentially, the algorithm is a colour threshold method that effectively utilizes the unique colour information of foreground objects in an image.

3.2.2 Target Segmentation

Target segmentation in our approach aims to separate robot ontology from a RGB original image of robotic dance pose, in which the GrabCut algorithm is adopted. GrabCut is an interactive foreground extraction algorithm using iterated graph cuts [24] and requires user to mark a rectangle around the object on the original image. Thus, the outer part of the

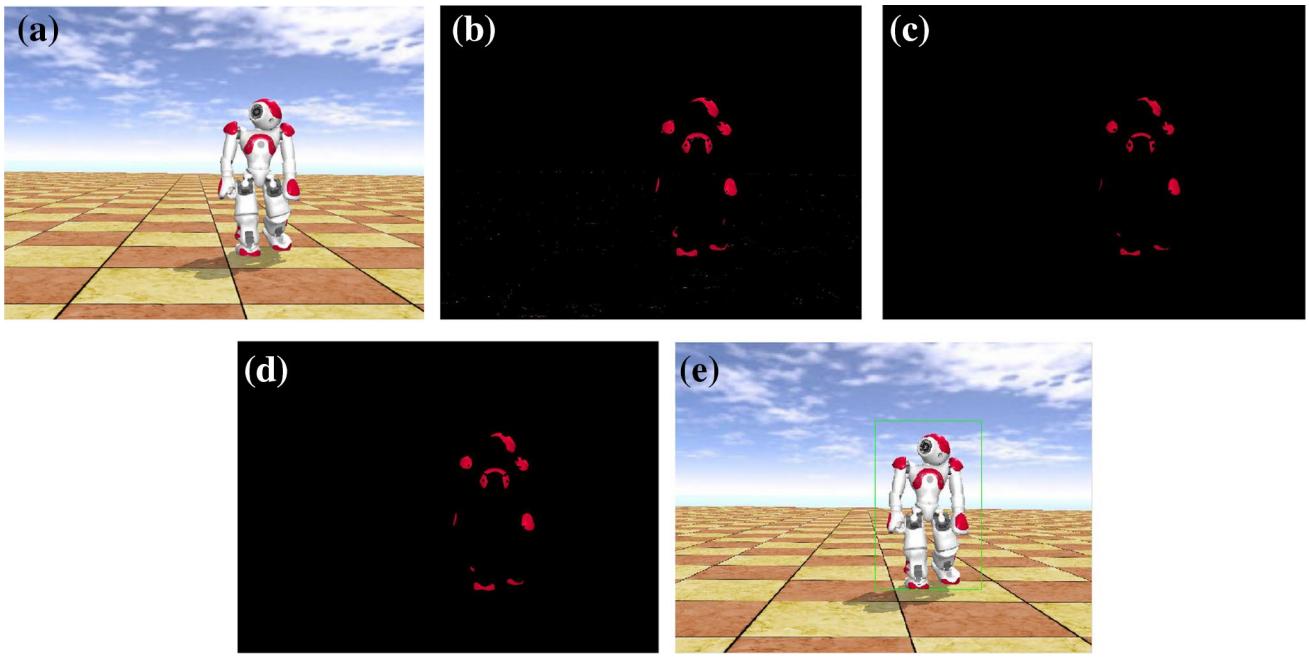

Fig. 2 The procedure of automatic target location. **a** Original image, **b** filtered image, **c** corroded image, **d** dilated image, **e** final image with target location rectangle (green rectangle). (Color figure online)

Fig. 3 The algorithm of automatic target location

1. Read original image;
2. Acquire each dimension data of the original image [M,N,C] (Line Num, Column Num, Channel Num);
3. **For** m=1:M
4. **For** n=1:N
5. traverse each pixel position [m,n], if its colour is not the specific colour then its colour is set to be background colour (black);
6. corrode the current image;
7. identify the approximate minimum enclosing rectangle (AMER) that contains robot;
8. compute a repeated corrosion parameter that is defined by the area ratio of AMER to original image, aiming to check if there exists the discrete error foreground pixels;
9. **If** (the repeated corrosion parameter<= the pre-set threshold value)
10. corrode the current image again;
11. dilate the current image;
12. identify the approximate minimum enclosing rectangle (AMER) that contains robot;
13. draw the AMER on the original image.

rectangle is defined as background, and the inner part is a combination of the object (foreground) and some background. Subsequently, the probability distribution models of foreground and background are built and then optimized for segmentation by minimizing energy function in several iterations until the target is finally separated from background.

In the phase of automatic target location, the humanoid robot is located by an automatically marked rectangle, and the rectangle is the input of the target segmentation phase. In other words, the GrabCut algorithm uses the rectangle as input, and an automatic image segmentation is then processed for separating robot ontology from a RGB original image of robotic dance pose.

Notably, in the GrabCut algorithm, an energy function, E, is defined so that its minimum corresponds to a good segmentation. The method of iterative energy minimization, which is used, guarantees convergence to at least a local minimum of **E**. When **E** converges, a set of parameter

values, α_n, on opacity are determined, which are then used for the best foreground segmentation.

Specifically, the GrabCut algorithm uses two Gaussian mixture models (GMM)—one for the foreground and another for the background—as well as the Gibbs energy function defined as follows [24]:

$$\mathbf{E}\left(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z}\right) = U\left(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z}\right) + V\left(\underline{\alpha}, \mathbf{z}\right) \tag{4}$$

where $\underline{\alpha}$ refers to the unknown opacity variables; $\mathbf{k}$ refers to the GMM component variables; $\mathbf{z}$ refers to the given image data. The data term, U, evaluates the fit of the opacity distribution, $\underline{\alpha}$, to the data, $\mathbf{z}$, given the Gaussian mixture models, $\underline{\theta}$, and is defined as follows [24]:

$$U\left(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z}\right) = \sum_{n} D\left(\alpha_n, k_n, \underline{\theta}, z_n\right) \tag{5}$$

The expansion of the term, D, (up to a constant) is defined as follows [24]:

$$\begin{aligned} D\left(\alpha_n, k_n, \underline{\theta}, z_n\right) = &-\log \pi\left(\alpha_n, k_n\right) + \frac{1}{2}\log det \sum\left(\alpha_n, k_n\right) \\ &+ \frac{1}{2}\left[z_n - \mu\left(\alpha_n, k_n\right)\right]^{\mathrm{T}} \sum\left(\alpha_n, k_n\right)^{-1}\left[z_n - \mu\left(\alpha_n, k_n\right)\right] \end{aligned} \tag{6}$$

Moreover, the parameters of the model are defined as follows:

$$\underline{\theta} = \left\{\pi(\alpha, k), \mu(\alpha, k), \sum(\alpha, k), \alpha = 0, 1, k = 1 \ldots K\right\} \tag{7}$$

where π refers to the weights; μ refers to the means; Σ refers to the covariances of the 2K Gaussian components for the background and foreground distributions [24]. Additionally, the smoothness term, V, is defined as follows:

$$V\left(\underline{\alpha}, \mathbf{z}\right) = \gamma \sum_{(m,n)\in\mathbf{C}} \left[\alpha_n \neq \alpha_m\right] \exp - \beta\left\|z_m - z_n\right\|^2 \tag{8}$$

where $[\psi]$ denotes the indicator function taking values 0,1 for a predicate ψ; $\mathbf{C}$ is the set of pairs of neighboring pixels; β is a constant that ensures the exponential term switches appropriately between high and low contrast; γ is another constant that takes a value of 50 [24].

Based on colour data modeling, the GrabCut algorithm achieves foreground segmentation in still images by iterative energy minimization. Notably, the GrabCut algorithm is applied directly to target segmentation in our approach and has not been optimized for that specific task.

3.2.3 Shape Extraction

Shape is an important visual content of image, and it is one of key information needed by human visual system to recognize objects. Moreover, it is the stable information of objects, and does not change with the surrounding environment's variation. Therefore, shape provides a feasible way to make machine understand a robotic dance pose.

In the stage of image pre-processing, the shape extraction phase follows the target segmentation phase, and the segmentation result is regarded as the input of the shape extraction phase. Meanwhile, shape extraction is processed from two aspects: region and contour, and they are regarded as the basis of the further shape feature extraction (detailed in Sect. 3.3). Figure 4d, e show the results of shape extraction, and Fig. 5 shows the corresponding algorithm. Essentially, the algorithm is designed based on morphological digital image processing technology.

3.3 Feature Extraction

In general, feature extraction refers to convert the primitive features to be a group of physical or statistical features. In our mechanism, feature extraction is built on the results of pre-processing, aiming to acquire the suitable features to describe a robotic dance pose. For a robotic dance pose, the result of kinematic parameter pre-processing is a kinematic parameter data vector (V_1, V_2, …, V_S); and the results of image pre-processing are region shape image and contour shape image. Therefore, our feature extraction focuses on three aspects: joint feature, region shape feature, and contour shape feature.

Notably, each aspect mentioned above selects its own representative features (for details, see the following subsections). All the visual features (including region shape, and contour shape) are considered as a whole. These feature extraction methods, which are just directly applied in our approach, have not been modified for the specific task in this paper.

3.3.1 Joint Feature

Joint feature, a good description of kinematic properties, is always used for portraying a dance pose [8–11]. In general, a humanoid robot has many joint motors in its whole body. Each joint motor can move in a particular direction, and the humanoid robot presents a dance pose by simultaneously actualizing all joint motors.

Thus, in our mechanism, a joint motor of a robot is translated into a joint feature, which describes a specific motor ability. Moreover, joint motor status is regarded as a value of the corresponding joint feature. When a humanoid robot has S joint motors, there are S joint features ($\{JF_1, JF_2, \ldots, JF_S\}$). As the result of kinematic parameter pre-processing, a kinematic parameter data vector, (V_1, V_2, …, V_S), is regarded as an original instance of the joint feature sequence, (JF_1, JF_2, …, JF_S), and should be further normalized.

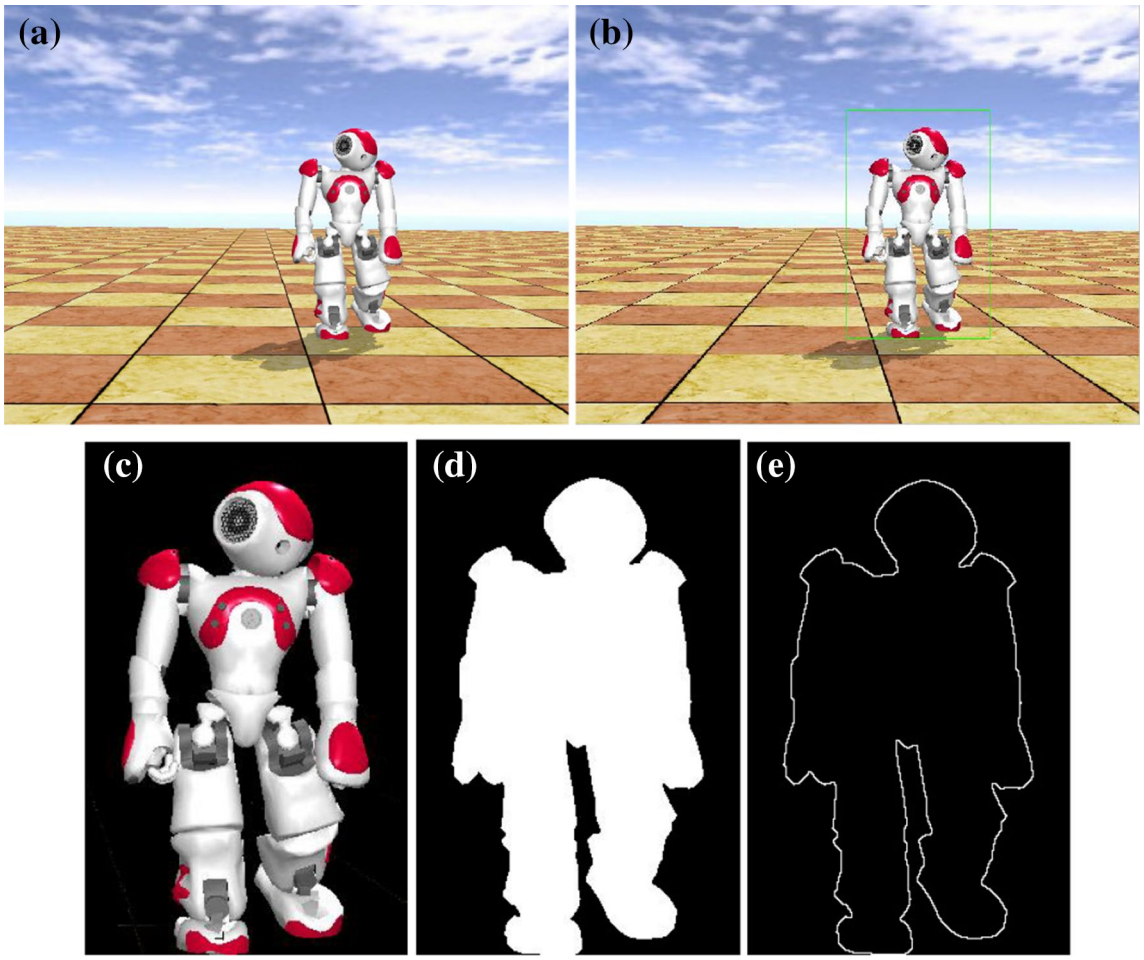

Fig. 4 The procedure of image preprocessing: **a** original image, **b** the result of automatic target location, **c** the result of target segmentation, **d** the result of shape extraction (region), **e** the result of shape extraction (contour). (Color figure online)

Fig. 5 The algorithm of shape extraction

1. Read the result of target segmentation (the sub image of robot ontology) IR_1;
2. Convert the RGB color image IR_1 to the single gray image IR_2;
3. Binarize the gray image IR_2 to be the black-and-white image IR_3;
4. Use the eight-connected breed filling algorithm to fill holes in IR_3, and get the result image IR_4;
5. Corrode the image IR_4, and get the result image IR_5;
6. Dilatethe image IR_5, and get the result image IR_6(the region of shape extraction);
7. Extract contour base on the image IR_6, and get the result image IR_7(the contour of shape extraction).

3.3.2 Region Shape Feature

The region shape is viewed as a whole in the region shape image, and all the pixels within the region shape are utilized effectively to describe shape information. In this way, the region shape is affected slightly by noise and shape changes. To describe effectively the region shape of a robotic dance pose, five types of region shape features are extracted respectively, namely eccentricity (EC), density (DE), rectangularity (RE), aspect ratio (AR), and Hu-moment Invariants (HuIM). Among them, the first four (EC, DE, RE, and AR) belong to simple geometric features, and the last one (HuIM) belongs to a statistical feature that is described by nonlinear combinations of geometric moments. They are defined as follows:

(1) Eccentricity: the eccentricity of ellipse which has the same second-order central moments with the region of the robotic dance pose;
(2) Density: the ratio of the square of the regional perimeter to the regional area;
(3) Rectangularity: the area ratio of the robot ontology region to its minimum enclosing rectangle (MER);
(4) Aspect ratio: the ratio of the MER's width to the MER's height;
(5) Hu-moment Invariants: there are seven invariant moment combinations, and their definitions is following:

$$\varphi_1 = \tau_{20} + \tau_{02} \tag{9}$$

$$\varphi_2 = \left(\tau_{20} - \tau_{02}\right)^2 + 4\tau_{11}^2 \tag{10}$$

$$\varphi_3 = \left(\tau_{30} - 3\tau_{12}\right)^2 + \left(\tau_{03} - 3\tau_{21}\right)^2 \tag{11}$$

$$\varphi_4 = \left(\tau_{30} + \tau_{12}\right)^2 + \left(\tau_{03} + \tau_{21}\right)^2 \tag{12}$$

$$\begin{aligned}\varphi_5 = {} & \left(\tau_{30} - 3\tau_{12}\right)\left(\tau_{30} + \tau_{12}\right) * \left[\left(\tau_{30} + \tau_{12}\right)^2 - 3\left(\tau_{21} + \tau_{03}\right)^2\right] \\ & + \left(\tau_{03} - 3\tau_{21}\right)\left(\tau_{03} + \tau_{21}\right)\left[\left(\tau_{03} + \tau_{21}\right)^2 - 3\left(\tau_{30} + \tau_{12}\right)^2\right]\end{aligned} \tag{13}$$

$$\begin{aligned}\varphi_6 = {} & \left(\tau_{20} - \tau_{02}\right) * \left[\left(\tau_{30} + \tau_{12}\right)^2 - \left(\tau_{03} + \tau_{21}\right)^2\right] \\ & + 4\tau_{11}\left(\tau_{30} + \tau_{12}\right)\left(\tau_{03} + \tau_{21}\right)\end{aligned} \tag{14}$$

$$\begin{aligned}\varphi_7 = {} & \left(3\tau_{21} - \tau_{03}\right)\left(\tau_{30} + \tau_{12}\right)\left[\left(\tau_{30} + \tau_{12}\right)^2 - 3\left(\tau_{03} + \tau_{21}\right)^2\right] \\ & + \left(\tau_{30} - 3\tau_{12}\right)\left(\tau_{03} + \tau_{21}\right)\left[\left(\tau_{03} + \tau_{21}\right)^2 - 3\left(\tau_{30} + \tau_{12}\right)^2\right]\end{aligned} \tag{15}$$

where τ_{jk} is the normalized $(j+k)$-order central moment:

$$\tau_{jk} = \frac{M_{jk}}{\left(M_{00}\right)^r}, \quad r = \left[\frac{j+k}{2} + 1\right] \tag{16}$$

and M_{jk} is the $(j+k)$-order central moment based on the region shape $f(x, y)$.

In the above five types of region shape features, eccentricity and aspect ratio reflect the broadness characteristics of the region, and density reflects the compactness characteristics of the region, and rectangularity reflects the fullness characteristics of the object to its minimum enclosing rectangle, and Hu-moment Invariants reflect the distribution characteristics of image grayscale.

Finally, a group of region shape features (EC, DE, RE, AR, HuIM_1, HuIM_2, …, HuIM_p) ($p \leq 7$) can be extracted from a region shape image of robotic dance pose. Moreover, the data of region shape features, acquired from the region shape image, should be further normalized.

3.3.3 Contour Shape Feature

The contour shape refers to a set of pixels that constitute the boundary of a region. By characterizing the geometrical distribution of a regional boundary, the contour shape feature can be described with some kind of descriptor. Fourier descriptors are a classical shape description method in transform-domain, and they are the Fourier transform coefficients of object shape boundary curve. That Fourier descriptors based on the coordinate sequence of object contours perform best among the various typical methods for 2-D shape recognition has been verified in the literature [25].

Therefore, complex coordinate based Fourier descriptors extracted from the contour shape image are regarded as the contour shape features of robotic dance pose. The abscissa for the contour shape image is taken as the real axis, and its ordinate is taken as the imaginary axis. Thus, a point on the X–Y plane corresponds to a complex coordinate. Starting from any point of the closed boundary on the X–Y plane, a one-dimensional complex sequence of points is obtained by traversing the boundary in a counter-clockwise direction. The one-dimensional complex sequence of points is shown as follows:

$$g(t) = x(t) + iy(t), \quad t = 0, 1, 2, \ldots, N-1, \quad i = -\sqrt{-1} \tag{17}$$

where N is the total number of sampled boundary pixel points.

The discrete Fourier coefficients of one dimensional sequence are defined as follows:

$$f(u) = \frac{1}{N}\sum_{t=0}^{N-1} g(t)\exp\left(\frac{-j2\pi ut}{N}\right), \quad u = 0,\ 1,\ 2,\ \ldots,\ N-1. \tag{18}$$

These discrete Fourier coefficients are Fourier descriptors, and then need to be further normalized. As f(0) describes the geometric centre position of the region that is surrounded by the contour boundary, f(0) is excluded from normalization and the rest $N-1$ Fourier coefficients are normalized. The normalized Fourier descriptors are defined as follows:

$$CCFD(v) = \frac{\|f(v)\|}{\|f(1)\|}, \quad v = 1, 2, \ldots, N-1. \tag{19}$$

The normalized Fourier descriptors have the invariance characteristics of rotation, translation, scale, and the starting

position. Moreover, the low-frequency components of the normalized Fourier descriptors always describe the contour and outperformed their high-frequency components. Therefore, some low-frequency components of the normalized Fourier descriptors ($CCFD_1$, $CCFD_2$, …, $CCFD_q$) ($q \leq [N/4]$) should be selected as the contour shape feature of robotic dance pose.

3.4 Feature Fusion

The purpose of feature fusion is to integrate several features to describe or portray an object completely. In this paper, to portray a robotic dance pose more completely, joint and shape features, which are extracted from two information channels (vision and non-vision), are fused. We believe the joint feature portrays the kinematic properties of a dance pose; shape features portray the overall silhouette and peripheral form; fusion of the above two simultaneously describes a robotic dance pose from the perspectives of movement and appearance. More concretely, in our mechanism, joint and shape features are fused into a mixed feature (joint + shape features). Specifically, the mixed feature is expressed by (JF_1, JF_2, …, JF_S, EC, DE, RE, AR, $HuIM_1$, $HuIM_2$, …, $HuIM_p$, $CCFD_1$, $CCFD_2$, …, $CCFD_q$) ($p \leq 7$, $q \leq [N/4]$).

3.5 Machine Learning

After feature fusion is processed, the stage of machine learning starts. Its task is to train a machine aesthetics model, aiming at making machine possess human aesthetic ability and implementing autonomous aesthetic judgment on robotic dance pose. By feature extraction and fusion, each robotic dance pose is expressed as an instance of the mixed feature. When a sufficient number of robotic dance poses are processed, the corresponding data set is produced.

To make machine possess human aesthetic ability on robotic dance pose, the supervised learning is necessary. Thus, human dance experts are invited to give their aesthetic evaluation (good/bad) on the all robotic dance poses after their observations. Viewing from machine learning, the example of each robotic dance pose is constituted of two parts: an instance on the mixed feature, and the corresponding aesthetic label (good/bad). Therefore, an example data set of robotic dance poses can be built to form a basis for further training a machine aesthetics model.

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]. Therefore, it is necessary to implement mainstream machine learning methods to compare their machine aesthetic effects, and find a more suitable and effective machine learning method among them. After a machine aesthetics model is built, a humanoid robot automatically evaluates the aesthetics by perceiving and observing its own new dance poses so that the further autonomous creation of robotic choreography is possible.

4 Experiments

As one of the most stylistic folk dances in China, Chinese Tibetan Tap has abundant variations on upper-body movements and relatively little variation on lower-body movements, as well as the most common form of standing body shape in the dance. Therefore, we have chosen Chinese Tibetan Tap as robotic dance form in our experiment. As one of the most popular humanoid robots nowadays, a Nao robot is selected as dance carrier in our experiment.

The simulated experimental environment includes four kinds of software: Webots7.4.1 simulator, Matlab R2014a, Dev-C++ 5.11, and Weka 3.6. After perceiving its joint motor data (internal kinestate), a simulated Nao robot displays a dance pose in the "Simulation View" area of Webots simulator. The joint motor data is considered as the perceived information source of the Nao robot. The pictures shown in "Simulation View" are treated as the visual information source in which the robot observes its own dance pose in the "mirror". Notably, the following underlying assumption exists in our simulation experiments: A robot always observes its own dance poses from a mirror, ignoring some limitations in real scenes (e.g. when the mirror is not placed in front of the robot, or the mirror does not appear in the range of vision of the robot, the robot cannot observe its own dance poses from the mirror). All the examples of original images (e.g. Fig. 2a) and the experimental images were acquired based on the above assumption.

Moreover, the shape features are extracted by image processing programs in Matlab. In Dev-C++, robotic dance poses are generated randomly, and data file formats are transformed. Furthermore, Weka is used for machine learning based on the extracted feature data (single feature or mixed features).

Five hundred robotic dance poses of Chinese Tibetan Tap were generated randomly based on the dance formalization of humanoid robot (HRDF) and three dance element sets [11]. For supervised learning, a Chinese folk dance expert was invited to label aesthetic categories (good/bad) on the 500 robotic dance poses. Considering that hands always maintain a naturally relaxed state when human dancers perform Chinese Tibetan Tap, the two hand joints ({LHand, RHand}) of our Nao robot were kept a fixed appearance without change. Thus, the remaining 24 joints of our Nao

robot were regarded as the joint features to describe a robotic dance pose ($S=24$). Moreover, the joint features of each dance pose were extracted after its joint data were acquired and normalized.

In the visual image pre-processing, the whole procedure was automatic. Figure 6 shows an automatic image processing GUI based on single captured image of robotic dance pose. Some parameters were set as follows: K = 6, Beta = 0.3 (GrabCut Algorithm); 7 Hu-moment Invariants ($p=7$) were all taken as one of region shape features of image; the total number of sampled boundary pixel points in each contour shape image is 800 ($N=800$); and 30 low-frequency components of complex coordinate based Fourier descriptors ($q=30$) were taken as contour shape features of image. Noted worthily, shape features of robotic dance pose consist of the following parts: eccentricity, density, rectangularity, aspect ratio, Hu-moment Invariants, and complex coordinate based Fourier descriptors, expressed by (EC, DE, RE, AR, HuIM_1, HuIM_2, …, HuIM_7, CCFD_1, CCFD_2, …, CCFD_{30}).

Generally, the mixed features of robotic dance pose are expressed by (JF_1, JF_2, …, JF_{24}, EC, DE, RE, AR, HuIM_1, HuIM_2, …, HuIM_7, CCFD_1, CCFD_2, …, CCFD_{30}). After normalizing the joint features and shape features, ten machine learning methods were used for performing automatic machine aesthetics of robotic dance poses, and ten-fold cross-validation methods were used for evaluation. Notably, all the machine learning methods used in our experiments came from the platform of Weka 3.6 and were not optimized according to our experimental tasks. For comparison, the machine learning methods were applied on three different feature combinations: joint feature, shape feature, and mixed features (joint feature + shape feature). The detailed machine learning results are shown in Table 2. Viewing from the final aesthetic evaluation results, the highest correct evaluation ratio is 81.6%, which comes from ADTree based on the mixed features.

Additionally, the following fact is exhibited in Table 2: The aesthetic evaluation results for the mixed features are close to those for the joint features. Determining the extent of the difference between both of the above is statistically meaningful; thus, a one-way analysis of variance was used for the statistically significant test. In the analysis, the null hypothesis is that there is no difference between them, and the significance level takes the value of 0.05 ($\delta=0.05$). The result of the analysis shows that the significance probability is 4.84573E−07 ($\zeta=4.84573\text{E}{-}07$). Thus, the significance probability is less than the significance level ($\zeta<\delta$), and then the null hypothesis is refused. Consequently, a significant difference exists between the aesthetic evaluation results for the mixed features and those for the joint feature.

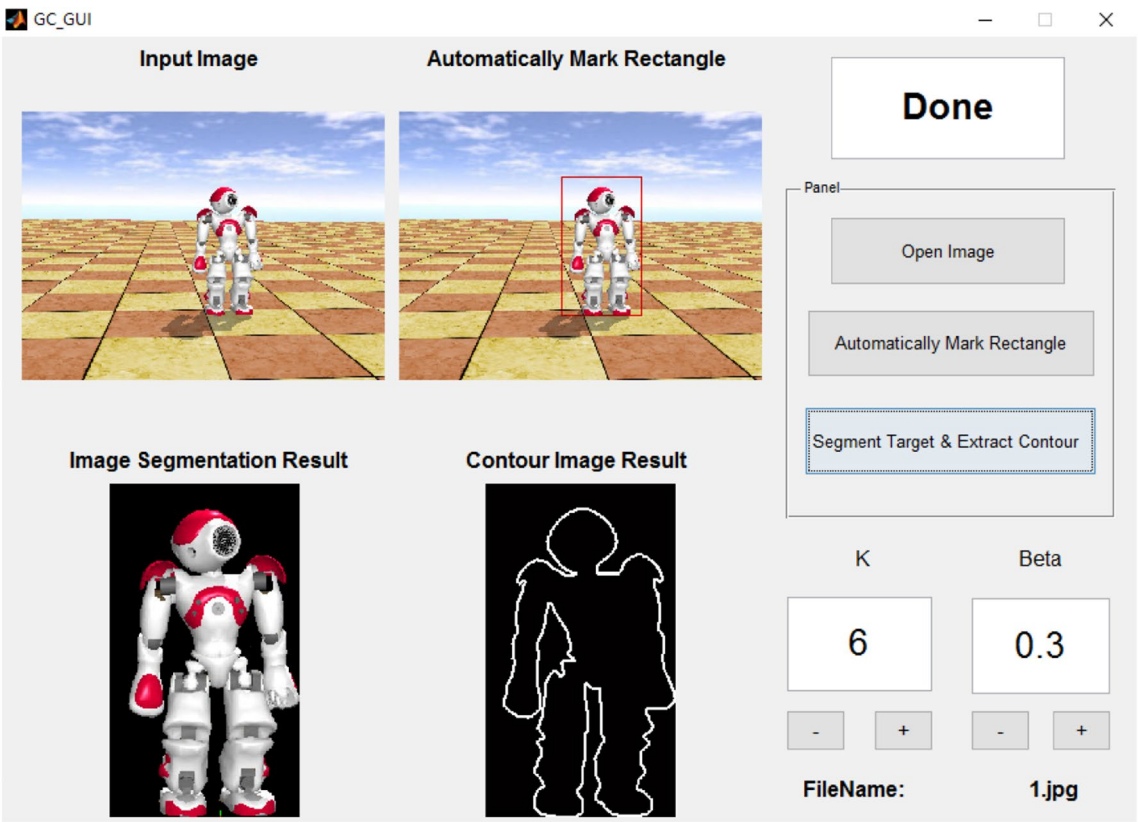

Fig. 6 Automatic image processing GUI of robotic dance pose

Table 2 The effect comparison on different machine learning methods based on different feature combination

Machine learning method	Only joint feature (correct ratio) (%)	Only shape feature (correct ratio) (%)	Mixed features (correct ratio) (%)
NaiveBayes	73	67.6	74
BayesianLogisticRegression	80.2	75.4	80
SVM	73.8	73.8	73.8
RBFNetwork	76	72.8	77.2
ADTree	80.8	73.6	81.6
RandomForest	80.2	73	79.6
VotedPerceptron	77.4	73.8	78.2
KStar	75.4	69.8	76.4
DTNB	80.2	71	79.2
Bagging	80.4	76	79.8
Average correct ratio	77.7	72.7	78
Highest correct ratio	80.8	76	81.6

5 Discussion

5.1 Feature Selection

In the above experiments, there are three feature combinations: joint feature, shape feature, and mixed features (joint feature + shape feature). The different feature combinations have different effects on machine aesthetics. As can be seen from the results, the correct ratio of mixed features is highest, joint feature is in the middle, and shape feature is lowest. Joint feature comes from the self-perception of internal kinestate of robot, and reflects the essential characteristics of robotic dance pose well. Therefore, even if only joint feature is used, the machine aesthetic effect of robotic dance pose is acceptable. In the above experiment, the average correct ratio of joint feature is 77.7%, and the highest correct ratio of joint feature is 80.8%.

Shape feature comes from the appearance impression of robotic dance pose via the robotic visual channel, which is the same as human beings. Although having the lowest correct ratio in the machine aesthetic effect among three features, shape feature has 72.7% on the average correct ratio and 76% on the highest correct ratio. It is useful for automatic machine aesthetics of robotic dance pose so that the robot could understand the beauty of its dance pose. Moreover, the poor machine aesthetic effect based on shape feature is caused by the following reasons:

(1) Humanoid robot shows its dance pose in 3-dimensional space. However, the image of robotic dance pose captured by robotic cameras is in a 2D space and 1D dimensional spatial data (depth data) is lost, which may result in the loss of shape feature.
(2) The adopted combination of shape features (eccentricity, density, rectangularity, aspect ratio, Hu-moment Invariants, and complex coordinate based Fourier descriptors) is insufficient to describe a robotic dance pose. The more powerful shape feature descriptors are required.
(3) There is the shadow in the captured image of robotic dance pose. GrabCut algorithm may not be able to segment the robot ontology and shadow. Therefore, the extracted shape of robotic dance pose may have certain distortion.

To improve the machine aesthetic effect of robotic dance pose based on shape feature, the following three measures can be considered:

- By using the depth image sensor for capturing the image of a real robot in a mirror, the RGB image and depth image of robotic dance pose can be acquired simultaneously. Therefore, the missing 1D spatial data (depth data) could be obtained.
- By combining other shape features (such as wavelet descriptor, scale space, Zernike moments, autoregressive, etc.), the more powerful or suitable shape feature descriptors could be found.
- By improving GrabCut algorithm, the robot ontology and shadow in the image of robotic dance pose could be correctly segmented.

Mixed features come from two information channels (vision and non-visual sensors) of the humanoid robot via multimodal information fusion. They describe the robotic dance pose and outperformed over single source feature (joint feature/shape feature). The machine aesthetic effect of robotic dance pose based on mixed features is the best among the three feature combinations. It has 78% on the average correct ratio and 81.6% on the highest correct ratio.

Compare to joint feature, mixed features bring the average correct ratio increment of 0.3%, and the highest correct ratio increment of 0.8%. It is foreseeable that the machine aesthetic effect of robotic dance pose based on mixed features (joint feature + shape feature) will improve further, if a more powerful shape descriptor can be extracted.

5.2 ADTree

Alternating Decision Tree (ADTree) is a boosting-based decision tree algorithm for classification and has a wide range of applications. ADTree consists of an alternation of decision nodes, which contain a single number, to specify a prediction condition and prediction nodes. An instance classified by an ADTree follows all paths for which all decision nodes are true, summing any prediction nodes that are traversed [26].

As can be seen from the experiments conducted above, ADTree demonstrates a better aesthetic effect among all machine learning methods listed in Table 2. It has gained the highest correct ratio (80.8%) on joint feature and the highest correct ratio (81.6%) on mixed features. Although it has not gained the highest correct ratio on shape feature, its correct ratio (73.6%) has exceeded the average correct ratio on shape feature (72.7%). Predictably, if the bottleneck of aesthetic performance based on shape feature is overcome, ADTree may achieve the highest correct ratio on shape feature.

Furthermore, as a concrete method of machine learning stage in semi-interactive evolutionary computation (SIEC), ADTree is used for machine aesthetics of robotic dance pose, and has gained the highest correct ratio among three machine learning methods (SVM, RBF network, and ADTree) [11]. Therefore, ADTree is an effective machine learning method for estimating robotic dance pose aesthetics. To further improve the correct ratio of robotic dance pose aesthetics obtained from ADTree, several aspects (such as information gain, Gini index, pruning, etc.) should be considered.

5.3 Multimodal Information Fusion

From the perspective of human ethology, by fusing multimodal information, human beings always exhibit a variety of daily behaviors (e.g. speech, walking, eating, sports, etc.) in their embodied environments. For humans, as the result of natural evolution, these actions happen in a conscious or unconscious way [27]. For example, when a person wipes a desk, he watches the desk with his eyes, and his hands simultaneously execute the wiping motion. Thus, visual and kinesthetic information work hand-in-hand for the task of wiping. As another kind of human daily behavior, the evaluation of aesthetics by humans of their own dance poses remains a procedure of multimodal information fusion. It should be noted that the imitation of human behavior is an effective way to develop artificial intelligence. Therefore, with this as inspiration, we propose a corresponding approach to make a robot imitate human behavior.

From the perspective of cognitive neuroscience, humans always perform perception tasks more precisely and effectively when multiple sense information (e.g. vision, audition, etc.) is provided. Although the information provided by each sense is distinct, the resulting representation of the surrounding world is not one of disjointed sensations, but of a unified multisensory experience [28]. Moreover, viewed from the cellular level, some cells in specific regions of the human brain respond to stimuli that emanate from multiple sensory information. For example, many cells in the superior colliculus fuse the information emanating from different sensory channels. In a phenomenon called multisensory integration [29], the cells then integrate this information and make an appropriate response [28]. Also inspired by this, we propose a corresponding approach to make a robot imitate the cognitive style that occurs in the human brain.

As seen from our experimental results (Table 2), multimodal information fusion brings about the improvement of the correct ratio of the aesthetics evaluation of robotic dance poses. Compared with the correct ratio of aesthetics evaluation brought about by a single information channel (visual or non-visual), multimodal information fusion results in the highest average correct ratio (78%) and the highest correct ratio (81.6%).

Moreover, compared with the correct ratio of aesthetics evaluation brought about by non-visual information channels (joint features), multimodal information fusion (mixed features) brings about only the average correct ratio increment of 0.3%, and the highest correct ratio increment of 0.8%. This phenomenon shows that, from the aspect of kinematic properties, a joint feature is a good feature for portraying a robotic dance pose. Although shape features bring about a limited promotion for the correct ratio of aesthetics evaluation on mixed features, shape features remain effective for multimodal information fusion. As to the reason why the performance improvement caused by multimodal information fusion is not obvious, we believe feature conflicts exist in the mixed features. However, at present, feature conflict is still an open problem in the aesthetics evaluation of robotic choreography, which we will explore in the future. Meanwhile, we believe that some more suitable mixed features, having fewer feature conflicts, exist for better portraying robotic dance poses. We will search for those features in the future.

Table 3 The comparison between the state-of-the-art approaches and our approach

	The approach in [8–10]	The approach in [11]	Our approach
Information channel	Non-visual	Non-visual	Non-visual and visual
Number of channels	Single	Single	Dual
Multimodal information fusion	No	No	Yes
Feature type involved	Kinematic	Kinematic	Kinematic; shape (region & contour)
Specific feature	Joint feature	Joint feature	Joint feature; region shape features (including eccentricity, density, rectangularity, aspect ratio, Hu-moment Invariants); contour shape feature (including complex coordinate based Fourier descriptors)
Feature fusion	No	No	Yes
Aesthetic manner	Human subjective aesthetics	Machine learning based method	Machine learning based method
Machine learning method involved	N/A	SVM, RBF network, ADTree	Naive Bayes, Bayesian logistic regression, SVM, RBF network, ADTree, random forest, voted perceptron, KStar, DTNB, bagging
Highest correct ratio	N/A	71.6667%	81.6%
Best feature combination	Joint feature	Joint feature	Joint feature + shape features
Best machine learning method	N/A	ADTree	ADTree

5.4 Comparison with the State-of-the-Art Approaches

There is a paucity of literature regarding the aesthetics of robotic dance poses, and what does exist focuses on the following two methods: human subjective aesthetics [8–10] and the machine learning based method [11]. Although more accurate aesthetic evaluation results are obtained for the former, extensive human–robot interaction creates a heavy burden for people. For the latter, although people need not extensively participate in human–robot interaction, it is difficult to build a suitable machine aesthetic model that will yield accurate aesthetic evaluation results.

To reduce the human burden and develop artificial intelligence, the machine learning based method for aesthetics evaluation is advocated. In general, our proposed approach belongs to the machine learning based method of aesthetics evaluation. Meanwhile, of note is the following fact: It is important to determine how to make a machine aesthetic model possess human aesthetic ability, which is still an open problem. We believe that good feature combinations and good machine learning methods, collectively, will help solve this problem.

A comparison between the state-of-the-art approaches and our approach is shown in Table 3. Different from what is presented in the existing literature, we explore the automatic aesthetic evaluation of robotic dance poses from the perspective of multimodal information fusion, which involves two channels, non-visual and visual. By fusing joint and shape features, we used mixed features to more completely portray a robotic dance pose. A good result (81.6%) on the mixed features is achieved with automatic aesthetic evaluation. Moreover, as an effective machine learning method for estimating robotic dance pose aesthetics, ADTree has been verified. Thus, the three main unsolved questions, mentioned in Sect. 1, have been answered well.

6 Conclusion

By using image processing and machine learning technologies, this paper presented an automatic machine aesthetics mechanism based on mixed features of robotic dance pose. The simulated experimental results show that the humanoid robot can integrate sensing data from two channels, implement multimodal information fusion, and evaluate the aesthetics of its own dance pose. Thus, the robot could conduct the autonomous dance activity as a human does. Moreover, it is proved that the shape features is useful to evaluate aesthetic feeling of robotic dance pose, and the mixed features (joint feature and shape features) can bring higher accuracy than single source feature (joint features or shape features) in the automatic machine aesthetics of robotic dance pose. Meanwhile, ADTree is also verified as a suitable and effective machine learning method of robotic dance pose aesthetics.

In the future, our work will be focused on three aspects: (1) to implement the proposed mechanism on a real Nao robot that is placed before a mirror, so that it could complete aesthetic evaluation of its own dance pose autonomously;

(2) to find more useful mixed features to describe a robotic dance pose; (3) to build an automatic aesthetic evaluation of robotic dance motion, based on the proposed mechanism.

Funding This work was supported by National Natural Science Foundation of China (Grant Nos. 61662025, 61806172), and the Research Foundation of Philosophy and Social Science of Hunan Province (Grant No. 16YBX042), the Research Foundation of Education Bureau of Hunan Province (Grant No. 16C1311), and the Startup Project of Doctor Scientific Research of Shaoxing University (Grant No. 20185003).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aucouturier JJ (2008) Cheek to chip: dancing robots and AI's future. Intell Syst 23(2):74–84
2. Peng H, Zhou C, Hu H, Chao F, Li J (2015) Robotic dance in social robotics—a taxonomy. IEEE Trans Hum-Mach Syst 45(3):281–293
3. Or J (2009) Towards the development of emotional dancing humanoid robots. Int J Soc Robot 1(4):367–382
4. Jeon M (2017) Robotic arts: current practices, potentials, and implications. Multimodal Technol Interact 1(2):5
5. Shiratori T, Ikeuchi K (2008) Synthesis of dance performance based on analyses of human motion and music. Inf Media Technol 3(4):834–847
6. Santiago CB, Oliveira JL, Reis LP, Sousa A (2011) Autonomous robot dancing synchronized to musical rhythmic stimuli. In: 2011 6th Iberian conference on information systems and technologies (CISTI 2011), pp 1–6
7. Meng Q, Tholley I, Chung PWH (2014) Robots learn to dance through interaction with humans. Neural Comput Appl 24(1):117–124
8. Vircikova M, Sincak P (2010) Dance choreography design of humanoid robots using interactive evolutionary computation. In: 3rd workshop for young researchers on human-friendly robotics (HFR 2010)
9. Vircikova M, Sincak P (2010) Artificial intelligence in humanoid systems, FEI TU of Kosice
10. Vircikova M, Sincak P (2011) Discovering art in robotic motion: from imitation to innovation via interactive evolution. In: Kim T, Adeli H, Robles RJ, Balitanas M (eds) International conference on ubiquitous computing and multimedia applications (UCMA), vol 150. Springer, Heidelberg, pp 183–190
11. Peng H, Hu H, Chao F, Zhou C, Li J (2016) Autonomous robotic choreography creation via semi-interactive evolutionary computation. Int J Soc Robot 8(5):649–661
12. Eaton M (2013) An approach to the synthesis of humanoid robot dance using non-interactive evolutionary techniques. In: 2013 IEEE international conference on systems, man, and cybernetics (SMC), pp 3305–3309
13. Shinozaki K, Iwatani A, Nakatsu R (2008) Construction and evaluation of a robot dance system. In: New frontiers for entertainment computing, Milano, Italy, vol 279. Springer, New York, pp 83–94
14. Oliveira JL, Reis LP, Faria BM (2012) An empiric evaluation of a real-time robot dancing framework based on multi-modal events. TELKOMNIKA Indones J Electr Eng 10(8):1917–1928
15. Manfrè A, Infantino I, Vella F, Gaglio S (2016) An automatic system for humanoid dance creation. Biol Inspired Cogn Archit 15:1–9
16. Augello A, Infantino I, Manfrè A, Pilato G, Vella F, Chella A (2016) Creation and cognition for humanoid live dancing. Robot Auton Syst 86:128–137
17. Manfré A, Infantino I, Augello A, Pilato G, Vella F (2017) Learning by demonstration for a dancing robot within a computational creativity framework. In: Proceedings—2017 1st IEEE international conference on robotic computing, IRC 2017, pp 434–439
18. Qin R, Zhou C, Zhu H, Shi M, Chao F, Li N (2018) A music-driven dance system of humanoid robots. Int J Humanoid Robot 15(5):1850023
19. Krasnow D, Chatfield SJ (2009) Development of the 'performance competence evaluation measure' assessing qualitative aspects of dance performance. J Dance Med Sci 13(4):101–107
20. Tutsoy O, Gongor F (2017) Analysis of facial characteristics. In: International conference on technology, engineering and science (IConTES), pp 262–272
21. Tutsoy O, Gongor F, Barkana DE, Kose H (2017) An emotion analysis algorithm and implementation to NAO humanoid robot. In: international conference on technology, engineering and science (IConTES), pp 316–330
22. Gongor F, Tutsoy O, Barkana DE, Colak S (2017) Sit-to-stand motion analysis for NAO humanoid robot. In: International conference on innovation trends in multidisciplinary academic research (ITMAR)
23. Gongor F, Tutsoy O, Colak S (2017) Development and implementation of a sit-to-stand motion algorithm for humanoid robots. J Adv Technol Eng Res 3(6):245–256
24. Rother C, Kolmogorov V, Blake A (2004) 'GrabCut': interactive foreground extraction using iterated graph cuts. ACM Trans Graph 23(3):309–314
25. Kauppinen H, Seppanen T, Pietikainen M (1995) An experimental comparison of autoregressive and Fourier-based descriptors in 2-D shape classification. IEEE Trans Pattern Anal Mach Intell 17(2):201–207
26. Freund Y, Mason L (1999) The alternating decision tree learning algorithm. In: Proceeding of the sixteenth international conference on machine learning, pp 124–133
27. Muehlenbein MP (2010) Human evolutionary biology. Cambridge University Press, Cambridge
28. Gazzaniga MS, Ivry RB, Mangun GR (2013) Cognitive neuroscience: the biology of the mind, 4th edn. W. W. Norton & Company, New York
29. Holmes NP, Spence C (2005) Multisensory integration: space, time and superadditivity. Curr Biol 15(18):R762–R764

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jing Li received her M.S. Degree from Yunnan University, Kunming, China in 2009. She is currently an associate professor in the Academy of Arts, Shaoxing University, China. Her research activity is focused on Chinese national dances, and robotic dance.

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, China, and also a lecturer in the College of Information Science and Engineering, Jishou University,

China. He is currently working as an academic visitor at the University of Essex. His research interests include brain-like intelligent systems, human–robot interaction, networked service robots, and machine learning.

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

Zhiming Luo received the B.E. degree from Xiamen University, China, in 2011, and Ph.D. degree from Xiamen University, China, in 2017. He is currently working as an Postdoc Researcher at Xiamen University. His research interests include computer vision, machine learning and medical image analysis.

Chao Tang received his Ph.D. degree from Xiamen University, China in 2014. Since then he is a lecturer at the Department of Computer Science and Technology at Hefei University. His research and project work focus on safety systems in transportation, pattern recognition and computer vision.