Received: 8 December 2020 | Revised: 10 March 2021 | Accepted: 13 March 2021

DOI: 10.1002/cav.1996

RESEARCH ARTICLE

WILEY

EmoDescriptor: A hybrid feature for emotional classification in dance movements

Junxuan Bai[1,2] | **Rong Dai**[1] | **Ju Dai**[2] | **Junjun Pan**[1,2]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

[2]Peng Cheng Laboratory, Shenzhen, China

Correspondence
Junjun Pan, State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China.
Email: pan_junjun@buaa.edu.cn

Funding information
Baidu academic collaboration program; Beijing Natural Science Foundation Haidian Primitive Innovation Joint Fund, Grant/Award Number: L182016; China Postdoctoral Science Foundation, Grant/Award Number: 2020M682827; National Key R&D Program of China, Grant/Award Number: 2018YFC0115102; National Natural Science Foundation of China, Grant/Award Numbers: 61872020, U20A20195; Shenzhen Research Institute of Big Data; Global Visiting Fellowship of Bournemouth University

Abstract

Similar to language and music, dance performances provide an effective way to express human emotions. With the abundance of the motion capture data, content-based motion retrieval and classification have been fiercely investigated. Although researchers attempt to interpret body language in terms of human emotions, the progress is limited by the scarce 3D motion database annotated with emotion labels. This article proposes a hybrid feature for emotional classification in dance performances. The hybrid feature is composed of an explicit feature and a deep feature. The explicit feature is calculated based on the Laban movement analysis, which considers the body, effort, shape, and space properties. The deep feature is obtained from latent representation through a 1D convolutional autoencoder. Eventually, we present an elaborate feature fusion network to attain the hybrid feature that is almost linearly separable. The abundant experiments demonstrate that our hybrid feature is superior to the separate features for the emotional classification in dance performances.

KEYWORDS

dance performances, emotional classification, feature fusion, hybrid feature

1 | INTRODUCTION

Similar to language and music, dance performances provide an effective way to express human emotions, especially in dance dramas. Nowadays, motion capture (MOCAP) systems are often utilized to obtain 3D human motion in animation and game industry. With the abundance of the MOCAP data, the content-based motion retrieval and classification have been fiercely investigated.[1-6] Although researchers attempt to interpret body language in terms of human emotions, the progress is limited by the scarce 3D motion database annotated with emotion labels.[7] Compared with the motion content, emotions are more difficult to be recognized from the body movements.

Nowadays, human emotions play an increasingly fundamental role in all fields in computer society. In the field of computer vision, researchers investigated emotional classification based on videos or images.[8,9] Moreover, emotional classification for speech or audio has always been a research hotspot[10-12] in natural language processing. Furthermore, sensor-based methods are proposed to classify human emotions.[13] Compared with the above fields, researchers in computer graphics pay less attention to human emotions. Although some scholars investigated the style transfer for motion data,[14-17] the proposed models can hardly be applied in emotional classification for complex motion such as dance performance. In fact, human emotion is becoming continuously important in computer graphics, for example, crowd simulation[18] and VR Rehabilitation.[19]

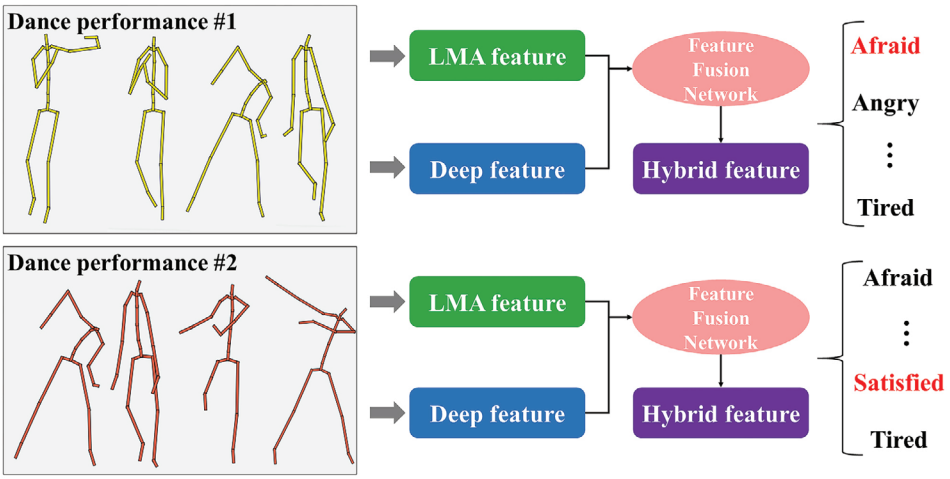


FIGURE 1 Method overview. The LMA feature and the deep feature are extracted separately, and then we use the proposed feature fusion network to obtain the hybrid feature. The hybrid feature is capable of classifying the emotion of dance performance correctly

To distinguish the emotion conveyed in human movement, a description for the human body needs to be defined first. To extract the emotions in videos or images, either a composition of human body parts or a kinematic chain is selected to model the human body.[7] Researchers in computer graphics prefer the kinematic chain model to deal with MOCAP data and then define descriptors to recognize the emotion. Recently, Aristidou et al.[20] quantified the Laban movement analysis (LMA), a technique that measures the dynamic properties for dance performances, and proposed 86 specific variables calculated based on motion data. Since the dance movements can be interpreted based on these variables, they are considered an explicit feature for emotion. In deep learning applications, Gram matrix has been utilized for image style transfer[21] and motion style transfer.[22] Our intuition is to use the Gram matrix as a deep feature, but it is not practicable to utilize this feature directly due to the high dimension of the matrix, which is expensive in computation time.

In this article, we propose a hybrid feature for emotional classification in dance performances. The hybrid feature contains the abovementioned features, that is, the explicit feature and the deep feature. Then we design an elaborate feature fusion network (FFN) to attain the hybrid feature. The overview of our approach is illustrated in Figure 1. The experimental results demonstrate that our proposed method is able to enhance classification accuracy greatly without too much computation cost.

The technical contributions of the article are summarized as follows:

- First, we propose a hybrid feature for the emotional classification in dance performances, which greatly enhances the classification accuracy without too much computation cost.
- Second, we design an elaborate neural network for the feature fusion, which describes the two types of features well. Moreover, the novel hybrid feature is almost linearly separable for the classification problem.
- Third, a large number of experiments are conducted to demonstrate the effectiveness of the new feature. The result proves that our innovative hybrid feature can classify dance emotion correctly.

2 | RELATED WORK

In this section, we review the techniques closely related to our approach, including human motion descriptors, emotion-related applications, and Laban movement analysis (LMA).

2.1 | Human motion descriptors

Human motion can be considered as time series data containing consecutive 3D human poses. Early methods generally defined motion models in observation space to extract the high-level properties. Tian et al.[1] proposed a semantic feature for motion data and retrieved desired motion using the feature. In their approach, they extracted keyframes and constructed a Gaussian mixture model (GMM) based on the keyframes, and then a semantic feature is obtained using the GMM probabilities. Cimen et al.[23] defined descriptors to categorize the affective state based on posture, dynamic, and

frequency. Some researchers attempt to define the descriptors using computer vision techniques. Laraba et al.[3] projected motion sequences to RGB images for action recognition and evaluated the accuracy using traditional classifiers. With the development of deep neural networks, researchers investigated the deep representation for human motion. Holden et al.[24] employed a 1D convolutional autoencoder to construct a motion manifold. Aristidou et al.[25] employed deep neural networks to obtain a descriptive representation for motion data. Recently, Chen et al.[6] presented a hybrid feature for action recognition composed of CNN-based feature and LSTM-based feature. Overall, there is a trend to represent human motion using deep learning techniques.

2.2 | Emotion-related applications

Emotion is expressed in a variety of ways and plays an increasingly fundamental role in human–computer interactions. The accurate recognition of emotion helps computers better understanding human thoughts, and it will benefit the design of the natural user interface. At the application level, VR/AR research on human emotion is expanding. Researchers surveyed how visualization style affects users' emotional responses,[26] how to inspect users' emotions in HMD,[27] and how to measure users' trust in virtual environments.[28] On the other hand, it is more practical to recognize human emotion from body language, especially when one is wearing VR glasses. Compared with facial expression or speech, recognizing emotion from body movement is difficult due to its diversity, which exists in the individual and the cultural differences.[29] Noroozi et al.[7] reviewed the literature on emotional body gesture recognition and summarized the advanced deep learning techniques for video data. As the author mentioned, complex representations for emotional recognition from the human movement are scarce because there is a lack of large-scale labeled databases. Particularly, this method[30] is helpful to construct a 3D motion database with emotional labels.

2.3 | Laban movement analysis

Laban movement analysis (LMA) originated from Rudolf Laban and has developed into a qualitative theory for human movement in dance, theater, pedestrian movement, and other nonverbal behavior.[31] It considers four aspects: body, effort, shape, and space. Aristidou et al.[20] quantified LMA and proposed 86 variables to describe motion sequences' properties, and then they utilized these features for emotional classification in dance movement. Then the LMA features were expanded to 121 and were used for motion control.[32] Senecal et al.[33] proposed a motion classifying method using neural network to map motions onto Russell circumplex model, forming an emotion trajectory. Although there exist similar definitions for motion data, the LMA features are more comprehensive. In addition to the aforementioned features, there are similar definitions for LMA.[34,35]

3 | PROPOSED METHOD

In our approach, the first step is to calculate the LMA features and the deep features separately, and then the feature fusion is achieved using our proposed fusion neural network.

3.1 | Data preparation

Recording human motion in different emotions is difficult, and it is subjective for emotion labeling. In our work, we select the dance performance database constructed by the University of Cyprus.[36] It contains contemporary dance performances in various emotions, which are performed by seven professional dancers. Compared with annotating the dance emotion after capturing, the database constructors used another way to label the emotion. To obtain dance performance in a specific emotion, performers are required to dance under a song in the desired emotion. Therefore, the whole movement is considered to be under the desired emotion. The performance is captured using an optical MOCAP system, and the missing data due to noise has been handled. The retargeting work has been done for all the movements. Each one lasts 90–120 s with 30 frames per second. There are 123 clips of dance performance in the database, and we select 109 clips in 12 emotions in the experiments, including afraid, angry, annoyed, bored, excited, happy, miserable, pleased, relaxed, sad, satisfied, and tired.

Instead of analyzing the emotion based on the entire dance performance, the emotional features are computed on 2-s subclips. There are 60 frames in each subclip with 30 frames overlapped between adjacent clips. After dividing the entire clip into small ones, we obtain about 5500 subclips. The initial skeleton has 54 joints, including 30 joints for the hand fingers. We merge the hand fingers and reduce the initial skeleton joints to 26 and keep three values (x, y, and z positions) for each joint. The final input is represented as $\mathbf{X} \in \mathbb{R}^{n \times d}$, where $n = 60$ and $d = 78$ (n is the number of frames in the subclip, and d is the total degree of freedom for the skeleton, i.e., $d = 26 \times 3$).

3.2 | LMA feature

For the explicit feature, we employ the 121 LMA features[32] to estimate the dance emotion, which consist of four components: *Body*, *Effort*, *Shape*, and *Space*. Compared with the deep representations, the LMA features are more interpretable than those. In Table 1, we list all the features utilized in our approach. The *Body* components describe how the body parts move, the *Effort* components express the dynamic properties of the movement, the *Shape* components analyze the changes of the entire body shape, and the *Space* components depict the relationship between the body and the environment. After the features are calculated, standardization is performed on the initial features. We denote the explicit feature as $\mathbf{F}_E$. Although it is possible to remove some redundant variables, we still use the 121 variables in our method. The reason is discussed in Section 4.3.

3.3 | Deep feature

To calculate the deep feature, a motion manifold is constructed using a 1D convolutional autoencoder.[24] Since the motion sequences are constructed by connected poses, the 1D convolution is able to extract the dynamic properties in a short time. After the latent variables are generated using the autoencoder, we construct the deep features based on the latent variables. The autoencoder is illustrated in Figure 2. Compared with the 2D convolution in image processing, the 1D convolution is more suitable for handling time series data.

The autoencoder consists of two operations: an encoder Φ and a decoder $\Phi^\dagger$. The input of the encoder is the joint positions $\mathbf{X}$, and the output of the encoder is the latent variable $\mathbf{H} = \Phi(\mathbf{X}) \in \mathbb{R}^{\frac{n}{2} \times m}$, where $m = 256$ in our approach. The decoder takes the latent variable $\mathbf{H}$ as input and outputs the position $\tilde{\mathbf{X}} = \Phi^\dagger(\mathbf{H}) \in \mathbb{R}^{n \times d}$. The encoder and decoder are formulated as follows,

$$\Phi(\mathbf{X}) = ReLU(\Psi(\mathbf{X} \otimes \mathbf{W}_0 + \mathbf{b}_0)), \tag{1}$$

$$\Phi^\dagger(\mathbf{H}) = (\Psi^\dagger(\mathbf{H}) - \tilde{\mathbf{b}}_0) \otimes \tilde{\mathbf{W}}_0, \tag{2}$$

where Ψ and $\Psi^\dagger$ are the max-pooling layer and the unpooling layer, and $\otimes$ denotes a convolution operation. The loss function with respect to the network parameters $\theta = \{\mathbf{W}_0, \mathbf{b}_0, \tilde{\mathbf{W}}_0, \tilde{\mathbf{b}}_0\}$ is defined as follows,

$$L(\mathbf{X}, \theta) = \|\mathbf{X} - \Phi^\dagger(\Phi(\mathbf{X}))\|_2^2 + \alpha\|\theta\|_1. \tag{3}$$

The first term minimizes the reconstruction error, and the second term is a regularizer to avoid the overfitting problem. α controls the influence of the regularizer.

Then we calculate the Gram matrix $\mathbf{G}$ as follows,

$$\mathbf{G} = Gram(\mathbf{H}) = \sum_i \mathbf{H}_i \mathbf{H}_i^T, \tag{4}$$

where i is the index on the temporal axis. We demonstrate the Gram matrices of two dance performances in Figure 3. Since the dimension of the matrix is 65,536, that is, 256×256, we perform principal component analysis (PCA) to reduce the dimensions, greatly decreasing the computation time for classification. Finally, we use 133 components (90%) to represent the deep feature. The reduced deep feature is denoted as $\mathbf{F}_D$ that will be utilized in the following procedures.

TABLE 1 The 121 LMA features utilized in our approach

LMA features		Max	Min	SD	Mean
Body	Left foot–hip distance	f_1	f_2	f_3	f_4
	Right foot–hip distance	f_5	f_6	f_7	f_8
	Left-hand–shoulder distance	f_9	f_{10}	f_{11}	f_{12}
	Right-hand–shoulder distance	f_{13}	f_{14}	f_{15}	f_{16}
	Hands distance	f_{17}	f_{18}	f_{19}	f_{20}
	Left-hand–head distance	f_{21}	f_{22}	f_{23}	f_{24}
	Right-hand–head distance	f_{25}	f_{26}	f_{27}	f_{28}
	Left-hand–hip distance	f_{29}	f_{30}	f_{31}	f_{32}
	Right-hand–hip distance	f_{33}	f_{34}	f_{35}	f_{36}
	Hip-ground distance	f_{37}	f_{38}	f_{39}	f_{40}
	Hip-ground minus feet-hip	f_{41}	f_{42}	f_{43}	f_{44}
	Feet distance	f_{45}	f_{46}	f_{47}	f_{48}
	Left-hand and chest	f_{49}	f_{50}	f_{51}	f_{52}
	Right-hand and chest	f_{53}	f_{54}	f_{55}	f_{56}
Effort	Deceleration peaks				f_{57}
	Pelvis velocity	f_{58}		f_{59}	f_{60}
	Left-hand velocity	f_{61}		f_{62}	f_{63}
	Right-hand velocity	f_{64}		f_{65}	f_{66}
	Left foot velocity	f_{67}		f_{68}	f_{69}
	Right foot velocity	f_{70}		f_{71}	f_{72}
	Pelvis acceleration	f_{73}		f_{74}	
	Left-hand acceleration	f_{75}		f_{76}	
	Right-hand acceleration	f_{77}		f_{78}	
	Left foot acceleration	f_{79}		f_{80}	
	Right foot acceleration	f_{81}		f_{82}	
	Jerk	f_{83}		f_{84}	
Shape	Volume (five joints)	f_{85}	f_{86}	f_{87}	f_{88}
	Volume (all joints)	f_{89}	f_{90}	f_{91}	f_{92}
	Torso height	f_{93}	f_{94}	f_{95}	f_{96}
	Hands level				f_{97}, f_{98}, f_{99}
	Volume (upper body)	f_{100}	f_{101}	f_{102}	f_{103}
	Volume (lower body)	f_{104}	f_{105}	f_{106}	f_{107}
	Volume (right side)	f_{108}	f_{109}	f_{110}	f_{111}
	Volume (left side)	f_{112}	f_{113}	f_{114}	f_{115}
Space	Total distance				f_{116}
	Area per second	f_{117}	f_{118}	f_{119}	f_{120}
	Total volume				f_{121}

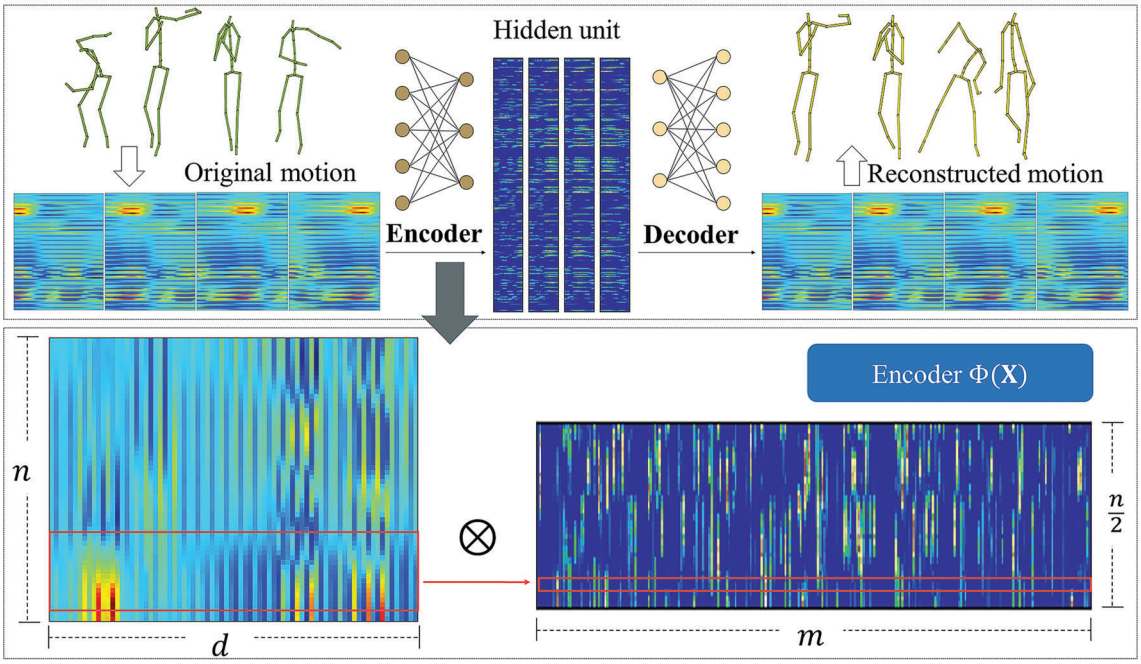


FIGURE 2 The illustration of the 1D convolutional autoencoder. The upper part is the autoencoder that includes an encoder and a decoder. The encoder takes the original motion as input and outputs the hidden unit, and the decoder takes the hidden unit as input and reconstructs the human motion. The lower part demonstrates the convolution operation of the encoder. As depicted in the figure, the hidden unit contains the properties for motion data in a period of time

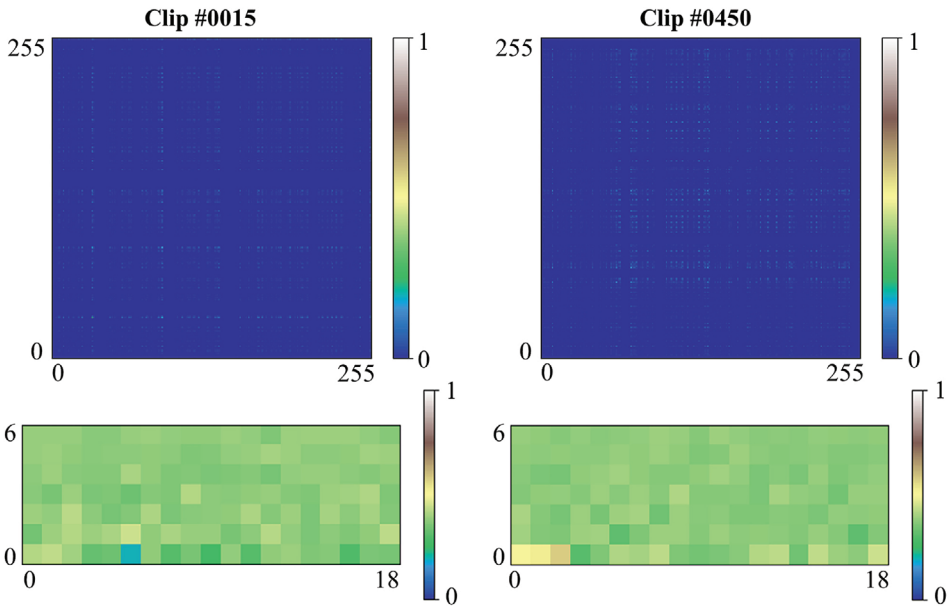


FIGURE 3 Visualization for the Gram matrix and the reduced ones of two dance clips. The upper ones are the original Gram matrix, and the lower ones are the reduced ones. We reshape the vector (133 dimensions) into a 7×19 matrix. The difference of the reduced ones is much more obvious than the original Gram matrix

3.4 | Feature fusion network

Our feature fusion network generates a novel hybrid feature that absorbs the advantages of the explicit feature $\mathbf{F}_E$ and the deep feature $\mathbf{F}_D$. This hybrid feature can precisely express the emotion of dance performance and enhance the accuracy of classification. The structure of the FFN is illustrated in Figure 4.

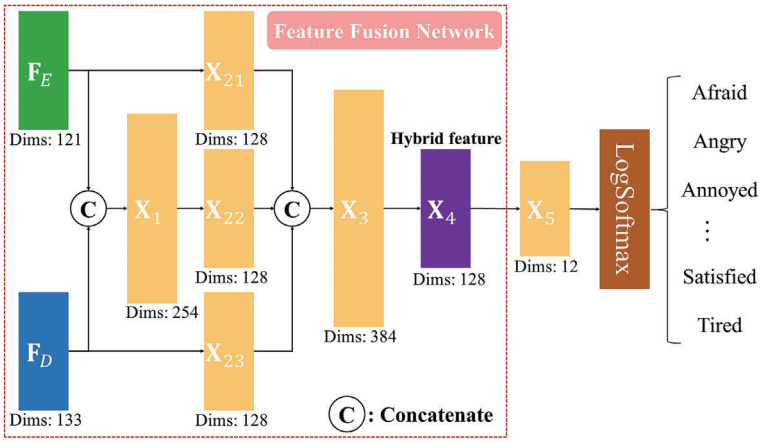


FIGURE 4 The structure of the feature fusion network (FFN). The inputs of FFN are the explicit feature $\mathbf{F}_E$ and the deep feature $\mathbf{F}_D$. In FFN, we only use fully connected layers. The hybrid feature $\mathbf{X}_4$ comes from three aspects, $\mathbf{X}_{21}$, $\mathbf{X}_{22}$, and $\mathbf{X}_{23}$. $\mathbf{X}_{21}$ and $\mathbf{X}_{23}$ contain the properties of $\mathbf{F}_E$ and $\mathbf{F}_D$, and $\mathbf{X}_{22}$ holds the property of the concatenated variable $\mathbf{X}_1$ that couples the explicit feature and the deep feature. The dimensionality reduction from $\mathbf{X}_3$ to $\mathbf{X}_4$ will generate a compact feature for classification. To train FFN, we need to obtain a 12-dimension variable $\mathbf{X}_5$, and we use the log softmax function to estimate the probabilities for the emotions

The inputs of FFN are the explicit feature $\mathbf{F}_E$ and the deep feature $\mathbf{F}_D$. Our network is constructed using the fully connected layer Υ. Given an input $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, the fully connected layer Υ is defined as follows,

$$\Upsilon(\mathbf{x}) = \sum_{i}^{N} w_i x_i + b, \tag{5}$$

where b is a bias term. The entire calculation procedures can be written as follows,

$$\mathbf{X}_1 = Concat(\mathbf{F}_E, \mathbf{F}_D), \tag{6}$$

$$\mathbf{X}_{21} = ReLU(\Upsilon(\mathbf{F}_E)), \tag{7}$$

$$\mathbf{X}_{22} = ReLU(\Upsilon(\mathbf{X}_1)), \tag{8}$$

$$\mathbf{X}_{23} = ReLU(\Upsilon(\mathbf{F}_D)), \tag{9}$$

$$\mathbf{X}_3 = Concat(\mathbf{X}_{21}, \mathbf{X}_{22}, \mathbf{X}_{23}), \tag{10}$$

$$\mathbf{X}_4 = ReLU(\Upsilon(\mathbf{X}_3)). \tag{11}$$

Finally, $\mathbf{X}_4$ is selected as our new feature for the emotional classification, which is denoted as $\mathbf{F}_H$. The hybrid feature $\mathbf{F}_H$ owns strong classification ability. On the one hand, our FFN inherits the properties of the explicit feature $\mathbf{F}_E$ and the deep feature $\mathbf{F}_D$, on the other hand, the network holds the properties of the concatenated feature $\mathbf{X}_1$. As a result, the concatenated variable $\mathbf{X}_3$ and the 128-dimension variable $\mathbf{X}_4$ are superior to the separate features $\mathbf{F}_E$ and $\mathbf{F}_D$, and $\mathbf{X}_1$. Since we wish to obtain a discriminative descriptor in low dimension, we take $\mathbf{X}_4$ as the hybrid feature.

To train the network, we add a fully connected layer to $\mathbf{X}_4$ to obtain a 12-dimension variable $\mathbf{X}_5 = ReLU(\Upsilon(\mathbf{X}_4))$, which represents the 12 emotions. After that, a log softmax layer is appended to the last for the classification as demonstrated in Figure 4. The log softmax layer is written as follows,

$$y_i = LogSoftmax(x_i) = log\left(\frac{exp(x_i)}{\sum_j exp(x_j)}\right). \tag{12}$$

In Equation (12), y_i represents the probability that the item belongs to class i. We employ the negative log-likelihood loss as the loss function, which is defined as follows,

$$Loss = \sum_{i} t_i log y_i, \tag{13}$$

where t_i represents target. We use the fully connected layers, the ReLU layers, and the log softmax layer in the entire network. The dimensions of the input vectors and the output vectors of each layer are described in Figure 4. Specifically,

the batch size is set as 20, and the number of the epoch is 100 in the training process. Also, we use 70% data for training and 30% data for evaluation.

4 | EXPERIMENTAL RESULTS

The training for the autoencoder, the training for FFN, and the emotional classification are executed on a desktop with Intel(R) Core(TM) i7-8700K CPU and NVIDIA GeForce GTX 1080 Ti GPU. The training for the autoencoder is operated on GPU, and the training for FFN and the classification are carried out on CPU. *PyTorch* is utilized to construct the autoencoder and FFN, and *sklearn* is employed for the classification.

We utilize the support vector classifier in the experiments. The linear kernel (Linear), the polynomial kernel (Polynomial), the radial basis function kernel (RBF), and the sigmoid kernel (Sigmoid) are employed in the classifier. In the classification, we use the one-vs-one decision function. To measure the performance of the feature, we calculate the accuracy and the time consumption of the classifications.

In the training process of the autoencoder, all the motion data are utilized. For FFN, 70% data are used for training and 30% data are used for evaluation. The aim is to obtain a good classification result using the deep neural network, and then we can use the intermediate variable as a hybrid feature. For the evaluation of the support vector classifier, we also employ 70% data for training and 30% data for testing.

4.1 | Classification for separate features

The classification results for the explicit feature $\mathbf{F}_E$ are documented in Table 2. Different kernel functions (Kernel), degree of the polynomial kernel function (Degree), and gamma (γ) are tested. To evaluate the features, the accuracy

TABLE 2 The classification results for the explicit feature, the Gram matrix, and the hybrid feature

Descriptor			$\mathbf{F}_E$			**G**			$\mathbf{F}_H$		
Kernel	**Degree**	γ	**Acc.**	T_{tra}/s	T_{pre}/s	**Acc.**	T_{tra}/s	T_{pre}/s	**Acc.**	T_{tra}/s	T_{pre}/s
Linear			51.33%	4.15	0.71	80.49%	834.77	403.56	95.00%	0.28	0.20
Polynomial	2	1.0	73.27%	1.88	0.71	**83.04%**	**1466.99**	**423.14**	94.56%	0.26	0.20
		0.1	74.56%	1.88	0.69	81.77%	1453.75	435.93	94.80%	0.25	0.19
		0.01	60.44%	2.05	0.77	82.44%	1510.49	417.25	**95.86%**	**0.29**	**0.25**
	3	1.0	**81.70%**	**2.10**	**0.67**	79.73%	1613.72	409.87	93.70%	0.23	0.18
		0.1	80.42%	2.07	0.69	81.40%	1595.81	418.53	93.58%	0.26	0.19
		0.01	63.81%	2.06	0.73	79.84%	1557.58	414.33	94.15%	0.27	0.20
	4	1.0	77.22%	2.67	0.69	74.77%	1770.44	415.90	94.12%	0.26	0.17
		0.1	77.54%	2.75	0.7	73.15%	1900.39	455.79	93.93%	0.24	0.17
		0.01	59.97%	2.39	0.75	73.98%	1810.57	414.44	93.10%	0.27	0.18
RBF		1.0	15.75%	3.77	0.96	11.36%	1978.41	474.38	11.77%	4.29	1.03
		0.1	58.70%	4.10	0.95	10.77%	2027.81	468.22	72.71%	4.14	0.95
		0.01	68.64%	2.32	0.89	12.05%	1987.86	483.23	95.75%	0.86	0.53
		0.001	37.61%	2.51	0.89	16.93%	2053.41	462.19	95.82%	0.87	0.72
Sigmoid		1.0	10.58%	1.93	0.95	11.03%	1038.41	447.71	11.09%	3.34	0.86
		0.1	10.91%	2.28	1.01	9.64%	1090.83	445.63	10.95%	3.57	0.94
		0.01	21.70%	1.85	0.86	8.60%	1115.81	452.83	21.13%	2.05	1.03
		0.001	30.53%	2.70	0.88	10.74%	1140.42	453.18	94.83%	1.06	0.77

(Acc.), the training time (T_{tra}), and the predicting time (T_{pre}) are calculated. The highest accuracy reaches 81.70% using a third-order polynomial kernel function when the coefficient is 1.0. Increasing the degree of the polynomial function will not improve the accuracy. The training time and the predicting time also need to be considered since they are meaningful for evaluating the efficiency of the feature. It needs $T_{tra} = 2.10$ s and $T_{pre} = 0.67$ s to train the model and predict the classification.

Also, In Table 2, we list the classification result for the Gram matrix **G**. The best accuracy reaches 83.04% using a second-order polynomial kernel function. Although the accuracy is increased slightly, it takes much more time to train the classifier ($T_{tra} = 1466.99$ s) and needs more time to predict the result ($T_{tra} = 423.14$ s) for the high dimensions of the deep features.

Then we employ the hybrid feature $\mathbf{F}_H$ for the emotional classification. The accuracy and time computations are greatly enhanced. The best accuracy reaches 95.85% with $T_{tra} = 0.29$ s and $T_{pre} = 0.25$ s using a second-order polynomial kernel. Compared with the separate features, the hybrid feature performs better in emotional classification.

4.2 | Classification for hybrid feature

To prove the utility of FFN, we compare the accuracy of a simple concatenated feature $\mathbf{X}_1$, that is, a vector that contains $\mathbf{F}_E$ and $\mathbf{F}_D$. We record the classification results for $\mathbf{F}_D$ and $\mathbf{X}_1$ in Table 3. $\mathbf{F}_D$ is obtained from the Gram matrix **G** using PCA. Although the dimension is reduced, the accuracy is improved (87.50%) since redundant information is removed. Moreover, the training time and predicting time are decreased drastically, $T_{tra} = 3.08$ s and $T_{pre} = 0.80$ s.

Besides, the result of the concatenated feature $\mathbf{X}_1$ is no better than the separate features. The best accuracy only achieves 77.77%. It can be seen that this simple combination cannot stand for the emotion of dance performance. However,

TABLE 3 The classification results for the deep feature, the concatenated feature, and the hybrid feature

Descriptor			$\mathbf{F}_D$			$\mathbf{X}_1$			$\mathbf{F}_H$		
Kernel	**Degree**	γ	**Acc.**	T_{tra}/s	T_{pre}/s	**Acc.**	T_{tra}/s	T_{pre}/s	**Acc.**	T_{tra}/s	T_{pre}/s
Linear			57.53%	3.90	0.68	57.50%	37.73	1.53	95.00%	0.28	0.20
Polynomial	2	1.0	85.18%	3.01	0.80	75.44%	4.07	1.74	94.56%	0.26	0.20
		0.1	**87.50%**	**3.08**	**0.80**	76.04%	4.42	1.79	94.80%	0.25	0.19
		0.01	76.85%	2.71	0.83	74.08%	4.28	1.87	**95.86%**	**0.29**	**0.25**
	3	1.0	87.23%	3.54	0.76	76.96%	4.67	1.72	93.70%	0.23	0.18
		0.1	86.31%	3.62	0.76	**77.77%**	**4.20**	**1.52**	93.58%	0.26	0.19
		0.01	64.90%	3.28	0.82	76.19%	4.26	1.54	94.15%	0.27	0.20
	4	1.0	77.71%	3.76	0.76	72.72%	5.84	1.63	94.12%	0.26	0.17
		0.1	78.55%	3.68	0.77	72.34%	5.81	1.53	93.93%	0.24	0.17
		0.01	49.22%	3.23	0.82	72.89%	5.88	1.52	93.10%	0.27	0.18
RBF		1.0	12.78%	4.41	1.03	11.46%	8.98	2.22	11.77%	4.29	1.03
		0.1	32.39%	4.23	1.08	10.20%	8.69	2.32	72.71%	4.14	0.95
		0.01	77.50%	3.75	1.01	11.04%	8.77	2.22	95.75%	0.86	0.53
		0.001	46.53%	3.29	1.04	22.31%	7.83	1.95	95.82%	0.87	0.72
Sigmoid		1.0	9.41%	2.83	1.18	10.99%	3.91	1.86	11.09%	3.34	0.86
		0.1	10.11%	2.68	1.01	10.35%	3.89	1.87	10.95%	3.57	0.94
		0.01	31.78%	2.47	1.07	10.18%	4.09	1.87	21.13%	2.05	1.03
		0.001	39.81%	3.61	1.06	9.76%	4.43	2.05	94.83%	1.06	0.77

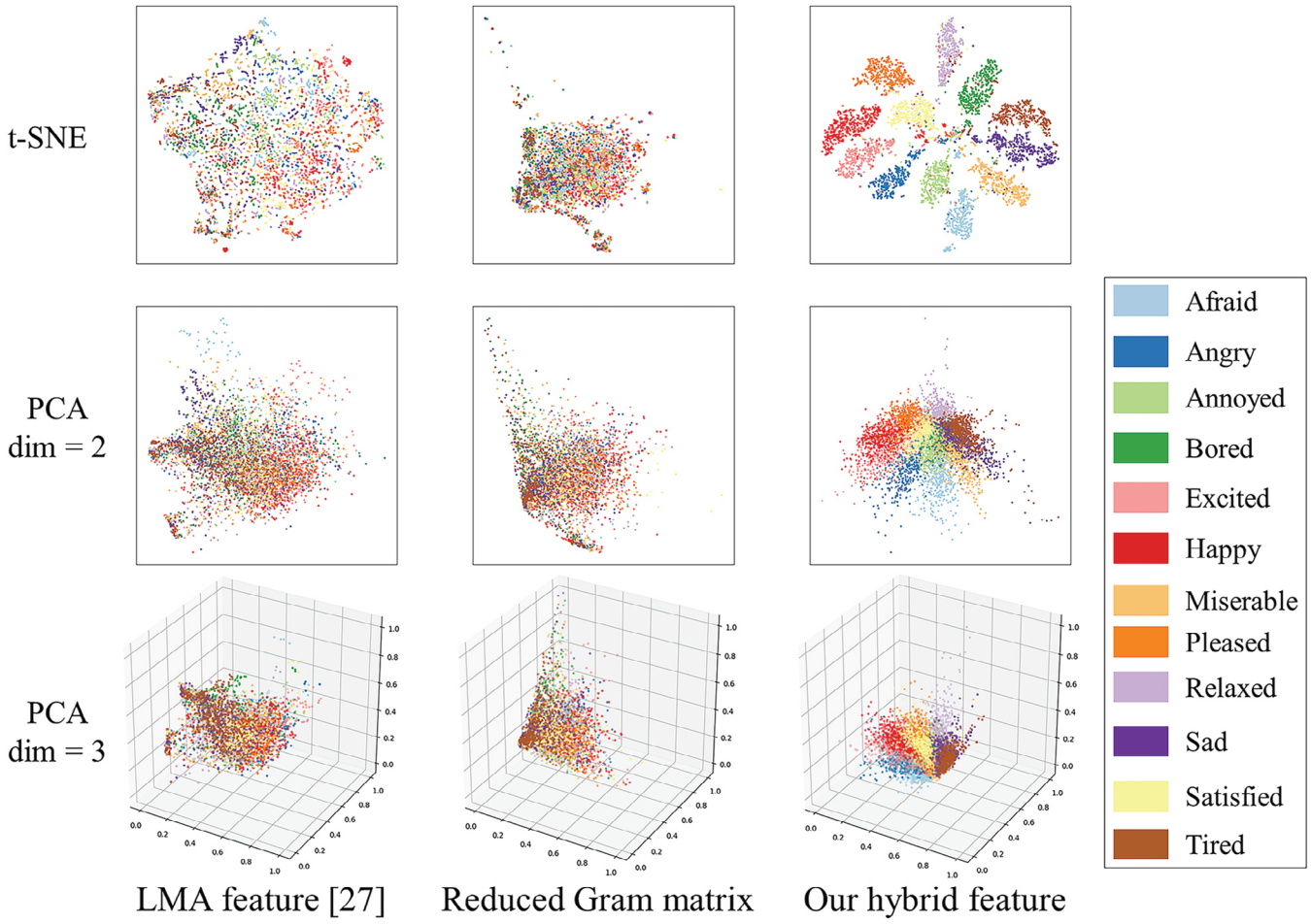


FIGURE 5 The comparison between our hybrid feature and the LMA feature,[32] and the reduced Gram matrix. In the first row, we utilize a nonlinear dimensionality reduction on these features to obtain a 2D representation. In the second and the third rows, we employ linear dimensionality reduction on these features. Clearly, our hybrid feature is almost linearly separable

the accuracy reaches 95.00% for the hybrid feature $\mathbf{F}_H$ even with a linear kernel, proving that FFN generates a powerful feature for representing the emotion of dance performance.

In Figure 5, we visualize the LMA feature[32] (i.e., the explicit feature $\mathbf{F}_E$), the reduced Gram matrix (i.e., the deep feature $\mathbf{F}_D$), and our hybrid feature $\mathbf{F}_H$. We employ linear dimensionality reduction (PCA) and nonlinear dimensionality reduction (t-SNE) on these features. It can be observed that the hybrid feature $\mathbf{F}_H$ is almost linearly separable.

4.3 | Discussion

In this section, we conduct a series of experiments to demonstrate the effectiveness of the new feature. The accuracy of the classification is significantly enhanced, and the time consumption is decreased. It verifies our hybrid feature is superior to the LMA feature[32] and the Gram matrix.[22] Moreover, our hybrid feature is almost linearly separable, so that it is suitable for measuring the emotion of human movement.

The goal of our neural network is to obtain a better representation for the emotional classification problem. Since the final task in FFN is a linear classification, the rest of the network tries to learn a representation to this classifier.[37] We also attempt to find a subset of the LMA variables to represent $\mathbf{F}_E$. We take the accuracy of SVM classifier (Polynomial, Degree $= 3$, $\gamma = 1.0$) as the standard. In the first experiment, we compute the classification accuracy after removing one variable, and then we delete all the variables whose accuracy is beyond 81.70%. However, the accuracy is decreased to

79.12%. The results are documented in Table A1. In the second experiment, we employ the backward elimination[38] to find a subset. The backward elimination is a greedy algorithm, which removes the worst feature at one step until the result cannot improve. In the experiment, we rank the classification accuracy after removing one variable, and then we delete the variables according to the accuracy. The results are listed in Table A2. The accuracy is slightly enhanced by 0.56% after removing four variables $\{f_{31}, f_{11}, f_{106}, f_{105}\}$. However, it is difficult to explain why these variables have a negative impact based on the data directly. Moreover, the increase of the accuracy is not obvious . As a result, we still use the 121 LMA variables to represent $\mathbf{F}_E$.

Nevertheless, our work is not without limitation. First, since most motion databases are not annotated with emotional labels, the hybrid feature is only evaluated on dance performances. Our experiment is constrained by the inadequate 3D motion data set. As mentioned in Reference 7, the quantity of labeled data is scarce, and there is no agreement among experts on the definition of primary emotion states. Compared with 3D MOCAP data, it is easier to collect data via videos. Then it is possible to transform the video data to 3D motion data. While the data qualities are not satisfactory, estimated from 2D videos,[39-41] constructing 3D motion database from 2D videos is promising. Second, we do not fully utilize the interpretability of the LMA features though the hybrid feature is beneficial for improving the emotional classification. It is meaningful to establish a connection between the explicit feature and the deep feature.

5 | CONCLUSION

In this article, we proposed a hybrid feature for the emotional classification of dance performances. Our hybrid feature was composed of two types of features, that is, the explicit feature and the deep feature. Instead of concatenating these features directly, we designed the feature fusion network to obtain the new feature, which was almost linearly separable for the classification. The experimental results demonstrated that our hybrid feature could enhance the classification accuracy and reduce the computation time.

In the future, we plan to construct a large-scale MOCAP database with common behaviors in different emotions and justify the proposed feature. The abundance of motion data with emotional annotations will expand the application scenarios for our technique. The new hybrid feature might be suitable for common behaviors, which can be utilized for mental health monitoring.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Andreas Aristidou for his kind discussion. This work was supported in part by National Key R&D Program of China (NO.2018YFC0115102), National Natural Science Foundation of China (No.61872020, U20A20195), Beijing Natural Science Foundation Haidian Primitive Innovation Joint Fund (L182016), Shenzhen Research Institute of Big Data, Shenzhen, 518000, China Postdoctoral Science Foundation (2020M682827), Baidu academic collaboration program, and Global Visiting Fellowship of Bournemouth University. Motion capture data used in this work were obtained from http://dancedb.cs.ucy.ac.cy, the Dance Motion Capture Database of the University of Cyprus.

ORCID

Junxuan Bai https://orcid.org/0000-0002-7941-0584

REFERENCES

1. Tian Q, Feng Y, Xiao J, Zhuang Y, Yang X, Zhang J. A semantic feature for human motion retrieval. Comput Animat Virt W. 2013;24(3-4):399–407.
2. Valcik J, Sedmidubský J, Zezula P. Assessing similarity models for human-motion retrieval applications. Comput Animat Virt W. 2016;27(5):484–500.
3. Laraba S, Brahimi M, Tilmanne J, Dutoit T. 3D skeleton-based action recognition by representing motion capture sequences as 2D-RGB images. Comput Animat Virt W. 2017;28(3-4):e1782. https://onlinelibrary.wiley.com/action/showCitFormats?doi=10.1002%2Fcav.1782.
4. Lv N, Jiang Z, Huang Y, Meng X, Gopi M, Peng J. Generic content-based retrieval of marker-based motion capture data. IEEE Trans Vis Comput Graph. 2018;24(6):1969–82.
5. Men Q, Leung H. Retrieval of spatial-temporal motion topics from 3D skeleton data. Vis Comput. 2019;35(6-8):973–84.
6. Chen Z, Pan J, Yang X, Qin H. Hybrid features for skeleton-based action recognition based on network fusion. Comput Animat Virt W. 2020;31(4-5).

7. Noroozi F, Kaminska D, Corneanu C, Sapinski T, Escalera S, Anbarjafari G. Survey on emotional body gesture recognition. IEEE Trans Affect Comput. 2018;1–1. https://ieeexplore.ieee.org/document/8493586.
8. Zhang H, Xu M. Recognition of emotions in user-generated videos with kernelized features. IEEE Trans Multimed. 2018;20(10): 2824–35.
9. Tian Y, Cheng J, Li Y, Wang S. Secondary information aware facial expression recognition. IEEE Signal Process Lett. 2019;26(12): 1753–7.
10. Zão L, Cavalcante D, Coelho R. Time-frequency feature and AMS-GMM mask for acoustic emotion classification. IEEE Signal Process Lett. 2014;21(5):620–4.
11. Chen M, He X, Yang J, Zhang H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process Lett. 2018;25(10):1440–4.
12. Shepstone SE, Tan Z, Jensen SH. Audio-based granularity-adapted emotion classification. IEEE Trans Affect Comput. 2018;9(2): 176–90.
13. Wen W, Qiu Y, Liu G, Cheng N, Huang X. Construction and cross-correlation analysis of the affective physiological response database. Sci China Inf Sci. 2010;53(9):1774–84.
14. Xia S, Wang C, Chai J, Hodgins JK. Realtime style transfer for unlabeled heterogeneous human motion. ACM Trans Graph. 2015;34(4):119:1–119:10.
15. Yümer ME, Mitra NJ. Spectral style transfer for human motion between independent actions. ACM Trans Graph. 2016;35(4):137:1–8.
16. Smith HJ, Cao C, Neff M, Wang Y. Efficient neural networks for real-time motion style transfer. Proc ACM Comput Graph Interact Tech. 2019;2(2):13:1–13:17.
17. Aberman K, Weng Y, Lischinski D, Cohen-Or D, Chen B. Unpaired motion style transfer from video to animation. ACM Trans Graph. 2020;39(4):64.
18. Xue J, Yin H, Lv P, Xu M, Li Y. Crowd queuing simulation with an improved emotional contagion model. Sci China Inf Sci. 2019;62(4):44101:1–3.
19. Bortone I, Leonardis D, Mastronicola N, Crecchi A, Bonfiglio L, Procopio C, et al. Wearable haptics and immersive virtual reality rehabilitation training in children with neuromotor impairments. IEEE Trans Neural Syst Rehabil Eng. 2018;26(7):1469–78.
20. Aristidou A, Charalambous P, Chrysanthou Y. Emotion analysis and classification: understanding the performers' emotions using the LMA entities. Comput Graph Forum. 2015;34(6):262–76.
21. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV: IEEE Computer Society June 27-30, 2016. p. 2414–23.
22. Holden D, Habibie I, Kusajima I, Komura T. Fast neural style transfer for motion data. IEEE Comput Graph Appl. 2017;37(4):42–9.
23. Cimen G, Ilhan H, Capin TK, Gürçay H. Classification of human motion based on affective state descriptors. Comput Animat Virt W. 2013;24(3-4):355–63.
24. Holden D, Saito J, Komura T. A deep learning framework for character motion synthesis and editing. ACM Trans Graph. 2016;35(4):138:1–138:11.
25. Aristidou A, Cohen-Or D, Hodgins JK, Chrysanthou Y, Shamir A. Deep motifs and motion signatures. ACM Trans Graph. 2018;37(6):187:1–187:13.
26. Volonte M, Babu SV, Chaturvedi H, Newsome ND, Ebrahimi E, Roy T, et al. Effects of virtual human appearance fidelity on emotion contagion in affective inter-personal simulations. IEEE Trans Vis Comput Graph. 2016;22(4):1326–35.
27. Yong H, Lee J, Choi J. Emotion recognition in gamers wearing head-mounted display. Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2019, Osaka, Japan: IEEE, March 23-27, 2019. p. 1251–2.
28. Gupta K, Hajika R, Pai YS, Duenser A, Lochner M, Billinghurst M. Measuring human trust in a virtual assistant using physiological sensing in virtual reality. Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2020, Atlanta, GA: IEEE; 2020. p. 756–65.
29. Kleinsmith A, Bianchi-Berthouze N. Affective body expression perception and recognition: a survey. IEEE Trans Affect Comput. 2013;4(1):15–33.
30. Kleinsmith A, Bianchi-Berthouze N, Steed A. Automatic recognition of non-acted affective postures. IEEE Trans Syst Man Cybern Part B. 2011;41(4):1027–38.
31. Laban/Bartenieff + somatic studies international. movement analysis; 2020. https://labaninternational.org/scope-of-practice/movement-analysis/. Accessed 26 Nov 2020.
32. Aristidou A, Zeng Q, Stavrakis E, Yin K, Cohen-Or D, Chrysanthou Y, et al. Emotion control of unstructured dance movements. Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation, Los Angeles, CA: Eurographics Association / ACM. July 28-30, 2017. p. 9:1–9:10.
33. Senecal S, Cuel L, Aristidou A, Magnenat-Thalmann N. Continuous body emotion recognition system during theater performances. Comput Animat Virt W. 2016;27(3-4):311–20.
34. Chi DM, Costa M, Zhao L, Badler NI. The EMOTE model for effort and shape. In: Brown JR, Akeley K, editors. Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000 July 23-28. New Orleans, LA: ACM; 2000. p. 173–82.
35. Durupinar F, Kapadia M, Deutsch S, Neff M, Badler NI. PERFORM: perceptual approach for adding OCEAN personality to human motion using laban movement analysis. ACM Trans Graph. 2017;36(1):6:1–6:16.
36. Dance mocap database University of Cyprus; 2020. http://dancedb.eu/. Accessed 5 Apr 2020.

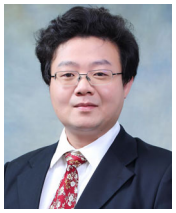
37. Goodfellow I, Bengio Y, Courville A. Representation learning. Deep Learning. London, UK: MIT Press; 2016. p. 524–54.
38. Han J, Kamber M, Pei J. Data preprocessing. Data Mining. 3rd ed. Boston, MA: Morgan Kaufmann; 2012. p. 83–124.
39. Martinez J, Hossain R, Romero J, Little JJ. A simple yet effective baseline for 3d human pose estimation. Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017. Venice, Italy: IEEE Computer Society. October 22-29, 2017. p. 2659–68.
40. Mehta D, Sridhar S, Sotnychenko O, Rhodin H, Shafiei M, Seidel H, et al. VNect: real-time 3D human pose estimation with a single RGB camera. ACM Trans Graph. 2017;36(4):44:1–44:14.
41. Yiannakides A, Aristidou A, Chrysanthou Y. Real-time 3D human pose and motion reconstruction from monocular RGB videos. Comput Animat Virt W. 2019;30(3-4):e1887. https://onlinelibrary.wiley.com/action/showCitFormats?doi=10.1002%2Fcav.1887.

AUTHOR BIOGRAPHIES

Junxuan Bai is currently a Ph.D. candidate in Computer Science, Beihang University. He studies in the State Key Laboratory of Virtual Reality Technology and Systems in China. He received his B.S. degree in Information and Computing Science from Dalian Maritime University in 2012, and M.S. degree in Computer Science from Beihang University in 2015. His research interests include computer animation, motion synthesis and editing, and deep learning techniques.

Rong Dai received her B.S. degree in computer science from Beihang University in 2020. Her research interests include human motion analysis and motion control in dance performance.

Ju Dai is currently a post doc in Peng Cheng Laboratory, Shenzhen, China. She received her B.S. degree and M.Sc. degree in Electronic Engineering, China University of Geosciences (CUG), Wuhan, China, in 2011 and 2014, respectively, and the Ph.D. degree in Signal Processing in Dalian University of Technology (DUT), Dalian, China, in 2020. Her research interests include person re-identification, saliency detection, human 3D pose estimation, and movement behavior analysis.



Junjun Pan is currently an associate professor in School of Computer Science and Engineering, Beihang University. He received both B.Sc. and M.Sc. degree in School of Computer Science, Northwestern Polytechnical University, China. In 2006, he studied in National Centre for Computer Animation (NCCA), Bournemouth University, UK as Ph.D. candidate with full scholarship. In 2010, he received the Ph.D. degree and worked in NCCA as Postdoctoral Research Fellow. From 2012 to 2013, he worked as a research associate in Center for Modeling, Simulation and Imaging in Medicine, Rensselaer Polytechnic Institute, USA. In November 2013, he was appointed as associate professor in School of Computer Science, Beihang University, China. His research interests include virtual surgery and computer animation.

How to cite this article: Bai J, Dai R, Dai J, Pan J. EmoDescriptor: A hybrid feature for emotional classification in dance movements. *Comput Anim Virtual Worlds*. 2021;e1996. https://doi.org/10.1002/cav.1996

APPENDIX A

Subset of the LMA variables

TABLE A1 The classification results after removing one feature

Removed f_i	Acc.	Removed f_i	Acc.	Removed f_i	Acc.	Removed f_i	Acc.
31	82.42%	11	82.33%	106	82.30%	105	82.28%
103	82.28%	118	82.09%	54	82.09%	37	82.08%
12	82.07%	76	82.04%	73	82.00%	13	81.99%
70	81.97%	26	81.97%	108	81.96%	116	81.95%
61	81.95%	18	81.91%	77	81.90%	82	81.87%
65	81.87%	29	81.87%	36	81.87%	90	81.85%
62	81.84%	64	81.84%	86	81.83%	98	81.82%
21	81.82%	79	81.82%	9	81.79%	88	81.78%
91	81.78%	6	81.77%	15	81.76%	87	81.73%
57	81.73%	99	81.72%	23	81.67%	40	81.65%
117	81.65%	66	81.63%	74	81.63%	95	81.62%
59	81.60%	56	81.60%	50	81.59%	38	81.58%
53	81.56%	4	81.56%	110	81.55%	30	81.54%
43	81.54%	27	81.54%	33	81.54%	39	81.51%
72	81.51%	84	81.50%	119	81.50%	71	81.49%
69	81.48%	121	81.48%	75	81.48%	24	81.47%
1	81.41%	35	81.39%	111	81.38%	8	81.38%
3	81.37%	101	81.37%	63	81.37%	14	81.35%
55	81.34%	47	81.32%	89	81.28%	114	81.27%
34	81.26%	2	81.25%	19	81.25%	51	81.23%
49	81.23%	22	81.22%	80	81.22%	120	81.21%
17	81.20%	93	81.20%	58	81.18%	115	81.17%
109	81.16%	68	81.15%	44	81.13%	45	81.12%
46	81.12%	113	81.12%	83	81.09%	20	81.06%
60	81.05%	97	81.03%	25	81.03%	67	81.02%
102	81.01%	96	80.98%	7	80.98%	100	80.95%
52	80.94%	48	80.89%	107	80.86%	16	80.81%
78	80.80%	85	80.78%	94	80.77%	41	80.77%
42	80.72%	112	80.62%	81	80.58%	10	80.51%
5	80.50%	32	80.45%	92	80.35%	28	80.31%
104	79.96%					None	81.70%

TABLE A2 The classification results after removing features using the backward elimination

Removed f_i	Acc.
31, 11	81.98%
31, 11, 106	82.01%
31, 11, 106, 105	82.26%
31, 11, 106, 105, 103	81.07%