Generic Body Expression Recognition Based on Synthesis of Realistic Neutral Motion

Arthur Crenn, Alexandre Meyer, Hubert Konik, Rizwan Ahmed Khan, Saïda Bouakaz

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 23, 2020, accepted November 1, 2020, date of publication November 17, 2020,
date of current version November 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3038473

Generic Body Expression Recognition Based on Synthesis of Realistic Neutral Motion

ARTHUR CRENN[1], ALEXANDRE MEYER[1], HUBERT KONIK[3], RIZWAN AHMED KHAN[2], AND SAIDA BOUAKAZ[1]

[1]LIRIS, CNRS, Univ Lyon, Université Claude Bernard Lyon 1, 69100 Villeurbanne, France
[2]Faculty of IT, Barrett Hodgson University, Karachi 74900, Pakistan
[3]LHC, UMR5516, Université de Saint-Etienne, F-42000 Saint-Etienne, France

Corresponding author: Alexandre Meyer (alexandre.meyer@univ-lyon1.fr)

This work was supported by the Region Auvergne-Rhône-Alpes.

ABSTRACT Most automatic expression analysis systems attempt to recognize a conventional set of expressions such as happiness, sadness, anger, surprise and fear, etc. Although this set of expressions is the most typical of the face, it is not the most representative/relevant for what the body expressions tell us. This paper presents a novel and generic approach for the recognition of body expressions using human postures. Our method is based on the notion of neutral motion generated from a given expressive one. In a second time, we estimate a residue function, as the difference between the two associated motions, namely the expressive and the neutral motion. More precisely, this function that is inspired by studies from psychology domain, gives a ''neutrality'' score of a motion. Using this ''neutrality score'', we propose a cost function which enables to synthesis the neutral motion from any input expressive motion. The synthesis of neutral motion process is based on two nested Principal Component Analysis providing a space where moving and selecting realistic human animations become possible. Proposed approach is evaluated on four databases with heterogeneous movements and body expressions and it achieved recognition results for body expression recognition that exceed state of the art.

INDEX TERMS Computer vision, body expression, automatic recognition, 3D skeleton, classification.

I. INTRODUCTION

Emotion is a complex phenomenon difficult to formalize. Interpretation of an emotion is subjective as two different people can perceive and interpret the same emotion differently [1]. Likewise, perception of emotion creates expressions which change from one culture to another [2], [3]. Furthermore, the complexity of an expression increases even more as humans express it through different channels such as facial expressions, speech, postures and movement [4]. Several studies from various domains have shown that body expressions are as powerful as facial expressions [5]. Nevertheless, if facial expression recognition was widely studied [6], [7], body expression recognition is still an emerging area. Furthermore, with the growth and easy access of devices that track 3-dimensional body, like the Kinect [8], [9] or accelerometer [10] based motion capture system, different applications will emerge based on body expression recognition. Our assumption is that many applications would benefit from the ability to understand human emotional state in order to provide more natural interaction, e.g. video games, video surveillance, human-computer interaction, artistic creation, etc.

This article presents method to detect and classify body expressions through sequence of 3D skeleton-based poses. The challenge is to propose a body expression recognition method invariant to body movement. For example, expression of happiness could be shown while the action is running, jumping, kicking, etc. The movement is composed by the action enhanced by the expression. Whereas, the state of the art approaches aim at recognizing body expression using motion-dependent features.

Our goal in this research work is to propose a generic approach for body expression recognition as illustrated in Figure 1 by separating expression from the action. Inspired by the field of animation synthesis, we propose to automatically create a neutral motion in order to extract body expression from any kind of input motion. This method allows to separate motion from expression, making proposed expression recognition method invariant to motion. Thus, main challenge we have tackled in this research work is to produce automatically neutral motion. We have derived

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

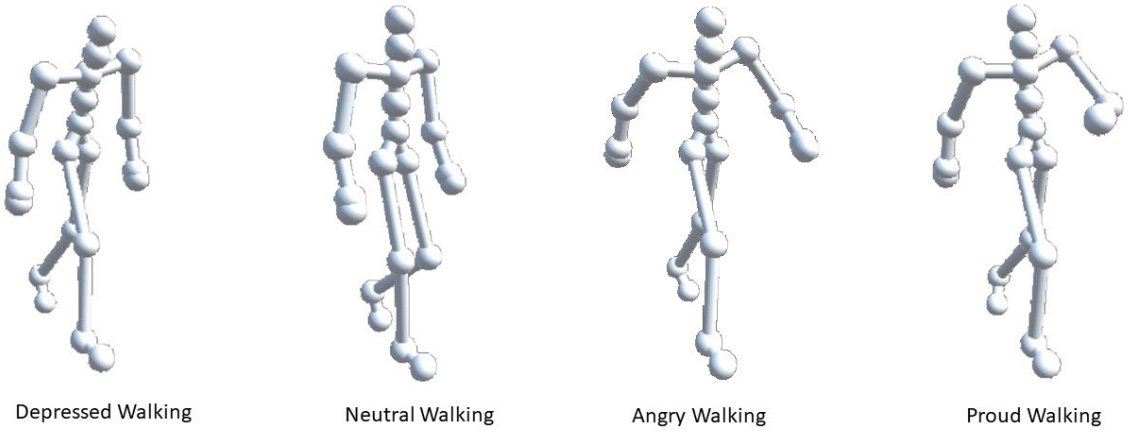

FIGURE 1. Our method automatically recognizes the expression of a human 3D skeleton animation. In the databases used to evaluate our method, the skeletal movement was captured by Kinect or any other motion capture systems that are now more and more accessible.

and aggregated different features from the state of the art in motion analysis, from psychology and computer vision domains in order to precisely characterize neutral motion.

Following the same paradigm as our previous work [11], the body expression is obtained by providing to a classifier the difference between the expressive motion and the corresponding neutral motion in the frequency domain using the Fast Fourier Transform [12]. As improvement to our previous work, the synthesis of neutral motion is automatically computed by a new method providing more realistic motions, and more efficient for the classification task. To the best of our knowledge, state-of-the-art approaches focus mostly on databases containing similar movements. We extended the evaluation phase by testing four different databases containing heterogeneous motions and expressions.

This article is organized as follows. Section II presents the state of the art on emotion analysis. Section III describes databases that are used to evaluate proposed algorithm. Section IV presents proposed novel method. Proposed method extracts body expression based on the residue formed by the difference between the original motion and a synthesized neutral motion. The synthesis of neutral motion is based on two nested Principal Component Analysis providing a space where transforming realistic human animations become possible. Results obtained on different challenging databases are discussed and compared with state-of-the-art methods in Section V. Finally, Section VI concludes the paper and presents future work.

II. RELATED WORK

Extensive research has been conducted to develop and evaluate methods for automatic expression recognition, mostly facial expression recognition. Generally, abundant literature exists, analyzing different input modalities such as speech, text, physiological signals, images, video, etc. The recognition techniques come from multiple research domains, such as signal processing, machine learning, computer vision, speech processing, etc. The goal of these techniques is to predict labels/class that would match the label a human would have perceived in the same situation. Even if verbal expression recognition [13] provides an important support of expression recognition, it is now widely accepted that nonverbal behavior constitutes an important medium of communication in addition to speech. Moreover, images of a person are as easy to record as their speech, since a camera record both image and sound whereas physiological measures are clearly less convenient to be recorded. The accuracy of expression recognition is usually improved when it combines the analysis of human expressions from multi-modal forms all together [14]. However, considering only a sub-part of channels such as only the face of a person or the posture stays an important way to improve knowledge of the whole domain since multi-modal approaches generally combine specific approaches dedicated to a single channel [15]. Moreover, the representations and descriptors designed for recognition may be extended to propose approaches for generating virtual agents with various perceived emotions [16].

For many years, expression recognition in computer vision mainly focused on automatic facial expression recognition [6], [17], [18]. Nevertheless the field of psychology has shown that body expression is as powerful as facial expressions in expressing emotions [19]. At the same time, with the increasing proliferation and popularity of human skeletal capture devices i.e. Kinect [9], accelerometers-based [10], webcam-based [20], etc., researchers have also focused on the problem of recognition of actions based on skeleton analysis [21], [22]. And more recently, research community has shown interest in the recognition of expression based on body movements [23].

Skeleton-based action recognition and expression recognition share the common point of analyzing a body human movement [24]. Both problems share the challenge of having to deal with high dimensional space, with many correlated degrees of freedom of a motion. Nevertheless, actions recognition approaches can not be used directly for recognizing body expressions. An expression can be seen as an enrichment of an action/gesture. In body expressions recognition, the actions and the expressions are overlapping.

Expression recognition is an orthogonal problem to action recognition, since the expression changes the gesture in a subtle and barely perceptible way. Even if separating the action and the expression seems non trivial, it is intuitively an interesting challenge to tackle. To the best of our knowledge, there was no literature available except our previous approach [11] that tried above stated approach.

In the areas of action recognition and facial expression recognition, very recent advances are mostly based on deep learning approaches [18]. Nevertheless, for body expression recognition, there is a lack of large datasets, which are required for deep learning. Secondly, actions and expressions are blended and most of the current literature is specialized in one specific type of motion. Lastly, model-based approaches that propose efficient analysis on the skeleton motion are still preferred [23], in comparison to brute force approaches letting a network explores a very large movement dataset. Thus, many articles focus on the most common motions such as walking [25]–[30], action of knocking [31], [32], talking persons [33], or artistic performance [34]. These methods are adapted on specific motions, but they fail when the challenge is to recognize body expressions from different scenarios.

Few articles tackled challenge of analysis of heterogeneous body movement. Kleinsmith *et al.* [2] proposed UCLIC database featuring 13 participants from different cultural regions, portraying four emotions (anger, fear, happiness and sadness). Wang *et al.* [35] proposed a real-time system that recognizes emotions from body movements. They used a combination of 3D postural features, kinematic and geometrical features. Truong *et al.* [36] proposed a new set of 3D gesture descriptors based on a generalization of the Laban descriptor model proposed by Rudolf Laban for gestures expressiveness. They evaluated their classification approach on their own database which contains 882 gestures. Dewan *et al.* [37] extended the Laban movement analysis by combining it with a temporal window, which can be seen as a more conceptually elaborate formulation of Laban theory. They evaluated their algorithm on the UCLIC database [2]. These approaches succeed in tackling the difficult problem of expression recognition of heterogeneous movements, but they all have the disadvantage of having to deal with expression mixed with action.

Our study aims to separate the expression from the action by modeling an expression as an enhancement of a gesture. Expression is what differs between two gestures performing the same action but with a different mood or style. We argue that separating the gesture from the expression is an important aspect for the analysis of body expression. This idea has been developed and proposed, with limited scope, in our previous work [11]. In this work, we combine knowledge from the field of psychology and animation synthesis in order to propose a new formalization of a neutral motion.

III. MATERIALS

The proposed approach is evaluated on four databases that are listed in Table 1. A brief description of the databases is presented below. The first three databases are real actors recorded by motion capture, while SIGGRAPH database (referred with same title in the remainder of this work) consists of synthetic animations generated by human animator extended by the method of Xia *et al.* [38]. Brief description of these four databases is given below.

TABLE 1. Description of the databases used for the evaluation of proposed method.

Data Base	Number of movements	Number of expressions
Emilya [39]	10001	8
UCLIC [2]	183	4
MPI [33]	1447	11
SIGGRAPH [38]	572	4 and 4 styles

1) Emilya Database [39]: this recorded database contains 10001 motions of 7 actions: walking, sitting down, knocking at a door, lifting and throwing an object with one hand and moving objects with two hands. It includes 8 expressions (joy, anger, panic fear, anxiety, sadness, shame, pride and neutral). The expressive motions are performed by 11 actors (6 females and 5 males) with an average age of 26 years.
2) UCLIC Affective Body Posture and Motion [2]: this database consists of 183 real motions with 4 expressions (fear, sad, happy, angry). The movements are carried out by 13 individuals from different cultural backgrounds. The subjects were asked to perform the emotion postures in their own way. The movements are acquired by a motion capture system which delivers a set of 32 body joints (3D positions).
3) MPI Emotional Body Expressions Database for Narrative Scenarios [33]: this database gathers 1447 motions. The scenarios were performed by 8 actors, 4 females and 4 males with an average age of 25 years. Actors were asked to imagine that they were narrating several stories to children. It contains 11 expressions (amusement, anger, disgust, fear, joy, neutral, pride, relief, sadness, shame, surprise). Database is recorded using motion capture system, which collects 3D postures leading to a set of 22 joints. Unfortunately, it is important to highlight that this database is highly imbalanced in terms of expressions. Joy is the most represented expression with 227 instances while shame is the less represented one with only 58 instances.
4) SIGGRAPH database [38]: this database contains synthetic sequences. It contains 572 animations with 8 body expressions or style (angry, childlike, depressed, neutral, old, proud, sexy, strutting). The motivation of using the SIGGRAPH database is that it includes a large range of movements: jump, run, kick, walk, punching and transitions between these motions.

IV. PROPOSED METHOD

We propose a method able to recognize body expressions from any input human motion represented as a skeleton motion as illustrated in Figure 1. The proposed novel approach is based on the principle of recognition of

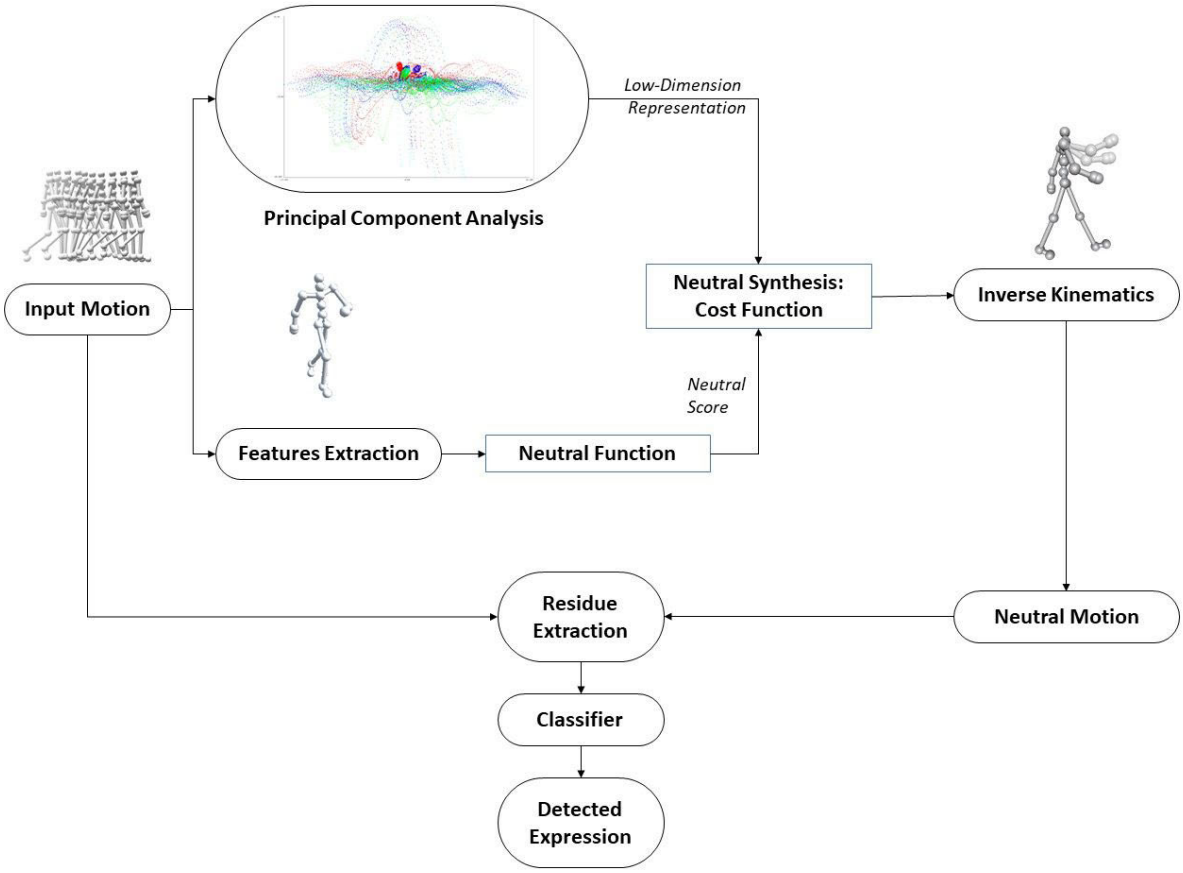

FIGURE 2. Overview of the proposed method. From the expressive input motion, a neutral motion is automatically synthesized using the formalism proposed in this paper. The residue between the synthesized neutral motion and the input motion is computed in the frequency domain. From this residue, a classifier learns on a sub-part of dataset of expressive motions how to recognize the expression. To evaluate the quality of our approach, the classifier is tested on the second part of the dataset. All this learning phase is repeated several times by mixing the dataset.

expressions by modeling the separation between the input movement and its neutral movement. This neutral motion is automatically computed using technique of a neutral motion synthesis, introduced in this paper. Compared to our previous work [11], the neutral synthesis step is more elaborated and generates neutral animations that greatly improves the recognition rate as illustrated in the result section V.

The overview of our approach is shown in Figure 2. We propose a novel way to synthesize neutral motion based on the optimization of a cost function. This cost function includes three terms. The main term is computed by a function, entitled the neutral function, that characterizes a neutral motion to evaluate the neutrality of the motion synthesized. Intuitively, this function returns a high value for a neutral movement and a near-zero value for a movement with a very strong expression, such as exaggerated ones existing in cartoons. In the Russell Circumplex Model (RCM) with the two axis of arousal and valence, a neutral animation would be placed near the center. This function is based on the formalization of several features from the field of psychology and is described in greater detail in the Section IV-A3. The second term of the cost function is a data attachment term in order to respect the original motion. The third term is the constraints term which ensures different constraints on the synthesized motion in order to generate realistic human motion. As illustrated on the Figure 2, a succession of two Principal Component Analysis (PCA) steps [40] are proposed to represent a motion in a low dimensional space and makes the optimization step tractable.

A. NEUTRAL MOTION SYNTHESIS

The principle of using a neutral motion synthesis for expression recognition has been validated in our previous work [11]. Even if our first method produced robotic and non-realistic neutral motion, it was sufficiently encouraging to achieve promising results. We argue that the quality of the generated neutral motion influences body expression recognition rates. Since the more realistic and convincing the neutral motion is, the better the separation between body expression and motion will be. In this work, we propose a novel way to synthesize the neutral motion. This section presents proposed new cost function used in an optimization

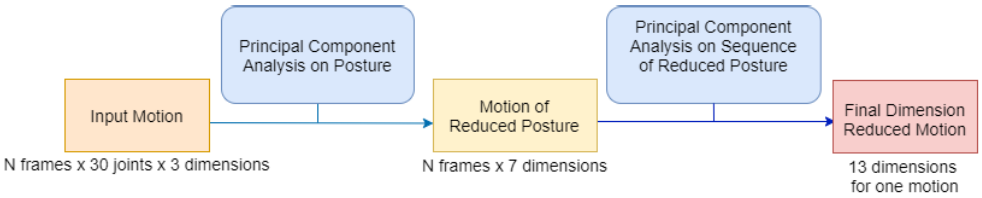

FIGURE 3. Process to reduce the dimension of one motion.

process to generate a neutral motion from the input expressive one.

1) OPTIMIZATION OF THE OVERALL COST FUNCTION

For the synthesis of the neutral motion, a cost function given in the Equation 1 with three weighted terms is introduced.

$$Cost(Motion) = \lambda Neutral(Motion) + \gamma Data(Motion) + \beta Penalty(Motion) \quad (1)$$

The first term named *Neutral* is the neutral score, explained in Section IV-A3. It evaluates the neutrality of a given motion. The second term named *Data* evaluates the distance between the generated neutral motion and the expressive input motion. This term of attachment to the data is used to avoid generating an animation too far from the original, where the action would be different than the original one. The last term is a regularization term, named *Penalty*, it adds penalties when the motion does not respect several constraints that would cause unrealistic movement. This term guarantees that the distances between joints will remain mostly constant and that the feet will not slip on the floor once in contact. We have designed the function in such a way that it returns high values for expressive motion, and value near zero for ''neutral'' animation, meaning animation including only an action without expressiveness. Although this is subjective, we have tried to formalize this with equations. The cost function minimization process aims to find the best parameters in the motion space build by the two steps of Principal Component Analysis (PCA), presented further in the Section IV-A2. Indeed, the number of degree of freedoms of human motion is too large to have a fast and accurate optimization process.

We optimize this cost function using the optimization algorithm called Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [41] on a subpart of the whole datasets to avoid overfitting as explained in the Section V and in Figure 5. CMA-ES presents the advantage to be derivative-free methods for numerical optimization of non-linear or non-convex continuous optimization problems, which is our case as well. CMA-ES is mainly used in animation with great success. We have carried out experiments/simulations with different values of λ, γ, β and we have empirically found that $\lambda = 0.8$, $\gamma = 0.6$ and $\beta = 0.5$ provides good default setting for all databases.

2) REPRESENTATION OF A MOTION IN A LOW DIMENSIONAL SPACE

This section explains how motion in a low dimensional space is represented with few optimization parameters. Optimizing every angle of every joint of every posture of an animation is too large to be reasonably tractable. Motion is then represented in low dimensional space using two steps, each one based on PCA:

1) the first PCA works on the posture, and
2) the second works on the temporal sequence.

Figure 3 shows the pipeline of this process. By applying two PCAs on input motion, we obtained projection space that represents the motion.

In addition to reduction of dimensions, PCA allows to define a space where a movement is relatively realistic and not really different from the original data. Comparing to a deep-learning-based approach, it has the advantage that it can be computed even with a limited amount of animation data.

For these two PCAs and the weight optimization presented in Section IV-A1, we isolated a subset of all animations of the databases in order to limit the risk of overfitting. The first PCA is computed on all the postures of this subset. We have re-targeted different skeletons topologies of the databases in one skeleton in a similar way as proposed by Holden *et al.* [42]. The unified skeleton is composed of 30 joints. Before the PCA, using 3D position for joints, a posture is described by 90 float values. By applying PCA to all the postures of the different databases, we evaluated that for our purpose a posture can be represented by 7 values, the 7 first components of the PCA. These 7 values preserve 95% of the original variance. We have tried to keep more components of the PCA without significant improvement in the recognition rates whereas less values decrease the rate. From this representation, we can represent a full motion with a temporal sequence by vector of 7 components. Then, we projected each posture of the motion on the first learned PCA, thus, reducing the dimension of a motion from $90 \times n$ frames to $7 \times n$ frames where n is the number of frames of the motion.

In a second step, another PCA is applied on the motion representation described above. This second PCA is applied in order to reduce the $7 \times n$ dimension. By applying this second PCA on all animations of the subset of the databases, we have evaluated that a motion can be represented by only 13 components while preserving 95% of the original variance. As for the posture, we have tested multiple values and found that 13 provides the best accuracy on the results.

Figure 4 shows the projection of the PCA applied on all the postures of the SIGGRAPH database. To illustrate the PCA in 2D, we have represented two main components of the PCA. Even with this 2D projection, we can see the different motions in the database. Each point represents a posture and one can see a motion by following a set of postures in a curve. The color represents different body expressions of the

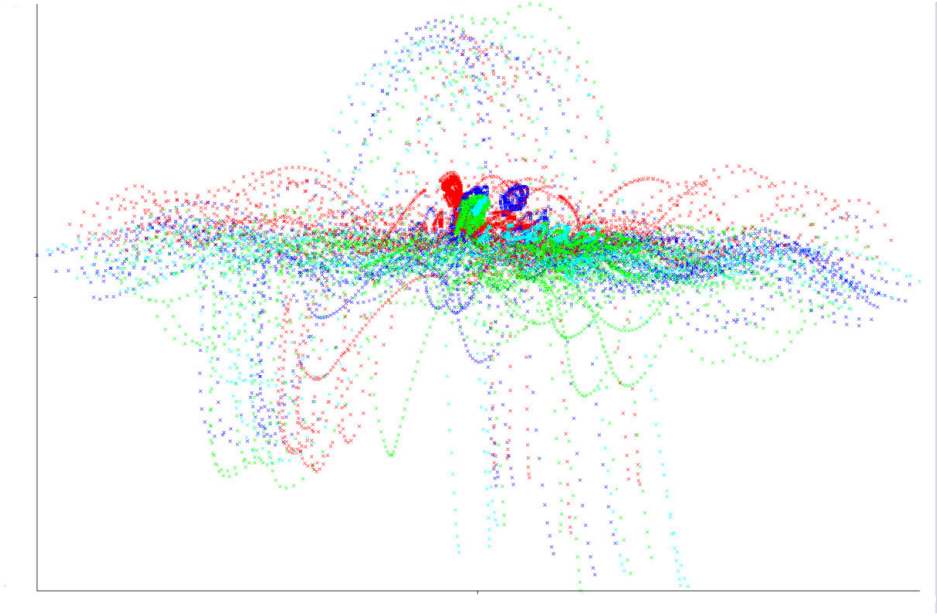

FIGURE 4. Projection on two axis of the PCA calculated on all the postures of the SIGGRAPH database. Each curve is a motion. The color represents different body expressions. Objectively the posture is well clustered in motions.

SIGGRAPH database. This figure shows that the posture is well clustered in motions.

3) MOTION NEUTRALITY QUANTIZATION

To produce a neutral motion from an expressive one, the input motion is first represented in the low dimensional space previously presented in Section IV-A2. The optimization step changes the representation parameters in order to minimize the cost function presented in Section IV-A1. This section focuses on the term ''Neutral'' of this cost function. In order to appreciate neutrality of a motion, we formalize several qualitative features from the domain of the psychology [43]. These features describe qualitatively the expression of a motion. We use them to quantify the neutrality of a motion, separating them into two clusters: the posture features and the temporal features.

Here we will explain how we designed the function that characterize the neutrality of a motion based on the formalization of features previously described. The neutrality function is defined as a weighted arithmetic mean given in the equation 2.

$$Neutral(Motion) = \frac{\sum_{j=1}^{n} \alpha_j f_j}{\sum_{j=1}^{n} \alpha_j} \tag{2}$$

where α_j corresponds to the weight of the j^{th} feature f_j and n is the number of features. Weights α_j are computed by an optimization on a set of animations. To avoid overfitting it is necessary to extract a subset of animations from the different databases described in Section V. The same optimal weights are then used for all experiment on the rest of animations. This optimization seeks to produce a score that tends towards zero for all neutral animations extracted from the datasets. All posture and temporal features f_j formalized in this study are described below.

- **Posture features**
  1) The Body openness $f_{bodyOpenness}$ measures whether arms are far from the body and/or feets are farm from each other.
  2) The Sagittal body leaning $f_{bodyLeaning}$ measures forward/backward leaning.
  3) The Body straightness $f_{bodyStraightness}$ measures the bending of the head/trunk/knees.
- **Temporal features**
  1) Movement power $f_{mvtPower}$ represents the amount of force involved in the movement.
  2) Movement fluidity $f_{mvtFluidity}$ represents the continuity of the movement.
  3) Movement speed $f_{mvtSpeed}$ measures the speed with which the movement is performed.
  4) Quantity of arms movement $f_{quantityArmsMvt}$ measures the amount of arms movement.
  5) Regularity of arms movement $f_{regularityArmsMvt}$ represents the variation in motion pattern.

After extraction of posture features on each frame, mean and standard deviation for each feature are computed. Temporal features are calculated in frequency domain using Fast Fourier Transform (FFT) [12]. FFT is a classical technique to extract power and regularity of temporal sequence. Proposed method uses Discrete Fourier Transforms (DFT) to extract temporal features as DFT is extremely useful in revealing periodicities in discrete input data.

Let x_n be a discrete time domain signal of one of the degrees of freedom (DOF) of a human motion data. The Discrete Fourier Transform X_k of x_n is given by:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \tag{3}$$

where N is the length of the signal and $i^2 = -1$. The single-sided spectrum X_ω is given by:

$$X_\omega = \frac{2}{N} X_k \quad k = 0, \ldots, N/2 \tag{4}$$

where x_s is the sampling frequency of the original time domain signal x_n, $\omega = (x_s/N)k$ is the frequency transformed from the samples k in the spectral space. Only single-sided spectrum in the positive frequency range ($\omega = 0 : x_s/2$) is used in proposed method.

From this spectral representation, magnitude and phase of the spectra are calculated. The magnitude defines existence and intensity of motion, whereas phase describes the relative timing. DFT is calculated on the movement of the joints, i.e. on the rotation signal of each joint.

Movement power is defined as the sum of magnitude spectrum of each body joints rotation signal.

$$f_{mvtPower} = \sum_{j=1}^{\theta} \sum_{q=1}^{4} \sum_{k=1}^{N/2} R[j, q, \omega_k] \tag{5}$$

where θ is the number of joints in a skeleton, the term q iterates on the 4 values of quaternion representing the rotation of the joint, ω_k is the k_{th} frequency, $R[j, q, \omega_k]$ is the magnitude for the joint j, the q^{th} term of the quaternion, at the frequency ω_k. Indeed, we represent rotation of each joint with quaternion, so with 4 values q_1, q_2, q_3, q_4 and the quaternion is given by $q_1 + q_2 i + q_3 j + q_4 k$.

The frequency of the Fourier transform provides an indication on the **regularity of the arms movement**. This regularity feature is the sum of the differences between the magnitude spectrum of each arm (shoulder, elbow and hand).

$$f_{regularityArmsMvt} = \sum_{j}^{\theta_{arms}} \sum_{q=1}^{4} \sum_{k=1}^{N/2} |R[j_{left}, q, \omega_k] - R[j_{right}, q, \omega_k]| \tag{6}$$

with θ_{arms} is *{shoulder, elbow, hand}* the three joints of an arm. The notation j_{right} (resp. j_{left}) provides the joint number of the three joints for the right arm (resp. left arm).

Third temporal feature is the **movement speed**, which measures the speed with which the movement is performed. It is given by the sum of the average speed of each joint.

$$f_{mvtSpeed} = \sum_{j=1}^{\theta} \frac{1}{N} \sum_{t=1}^{N} \frac{dP_j(t)}{dt} \tag{7}$$

where $P_j(t)$ is the world position of the joint j at the time t. Speed, the first derivative of the position, is computed by finite difference.

The **quantity of arms movement** is defined by the mean of the cumulative distance covered by every joint in both arms.

$$f_{quantityArmsMvt} = \sum_{j=1}^{\theta_{arms}} \sum_{t=1}^{N} P_j(t) \tag{8}$$

Last temporal feature is **movement fluidity**. It is computed by calculating mean of the acceleration for both hands and both feet.

$$f_{mvtFluidity} = \sum_{j=1}^{\theta_{endEffectors}} \frac{1}{N} \sum_{t=1}^{N} \frac{d^2 P_j(t)}{dt^2} \tag{9}$$

with $\theta_{endEffectors}$ is *{left_hand, right_hand, left_foot, right_foot}* .

B. NEUTRAL VS EXPRESSIVE MOTION: RESIDUE EXTRACTION

Section IV-A explained how neutral motion corresponding to the original input motion is computed. Proposed framework for body expression recognition theorizes that body expression is present in the residue formed by the difference between the neutral and the expressive motion and that a spectral representation of a motion is well adapted to separate the expression from the gesture. This theory/assumption is supported by the work of Crenn *et al.* [11] and Yumer and Mitra [44] that managed to extract expression of a motion with this formalism.

The residue between the neutral animation and the expressive animation is an array of values computed for each degree of freedom (DOF) of each joint in the skeleton independently from the others. It consists in a subtraction between the neutral spectral magnitude and the expressive spectral magnitude. In a formal manner, it is described by the Equation 10.

$$\begin{aligned} \textbf{Residue} = array[\ldots,\ & |R_o[j, q, \omega_k] - R_s[j, q, \omega_k]|,\ \ldots] \\ & j \in \theta \\ & l \in \text{DOF (the 4 quaternion values)} \\ & k \in 1..N/2 \quad \text{where N is the signal length} \end{aligned} \tag{10}$$

where $R_s(j, q, \omega_k)$ (resp. $R_o(j, q, \omega_k)$) is spectral magnitude for the joint j and for the DOF q at the frequency ω_k during an action computed on the synthesized neutral movement (resp. on the original movement). The magnitude contains the information about the motion and the expression of an animation. The residue forms the feature vector, which is used as input data to the classifiers in order to get body expression class.

V. EXPERIMENTS AND RESULTS

To evaluate proposed approach, several experiments were carried out on several databases presented in Section III. To generate neutral animations, our algorithm optimizes a neutral function computed by a weighted average of different terms as explained in the Section IV. This optimization is done in a space reduced by two steps of PCA. The weights are the results of a preliminary optimization. To compute the PCA and to determine the adapted weights reducing overfitting as much as possible, 25% of animations are randomly extracted from all databases by keeping the same ratio of each expressions as in the whole datasets as illustrated in Figure 5. These 25% of animations are used only for the weights computation and the PCA, they are not used in the recognition process. Table 2 presents recognition rate achieved by proposed approach and compares performance of proposed method with different classifiers in Figure 6 :

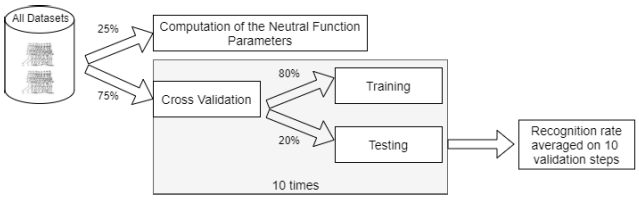

FIGURE 5. 25% of all animations are randomly extracted from all databases for computing the PCA and the weights of the neutral function. For computing recognition rate, the cross validation process is repeated 10 times with training the classifier on a random selection of 80% of the subset and tested on the remaining 20%.

TABLE 2. Recognition accuracy: comparison of proposed method with state-of-the-art methods. Our approach uses SVM classifier with a k-fold of 10.

Database	Results from state-of-the-art	Proposed Method	Kappa index
Emilya	–	**82.2**%	0.796
UCLIC	**87.30**% [37]	74%	0.653
MPI	50% [11]	**78.6**%	0.765
SIGGRAPH	98% [11]	**98.8**%	0.986

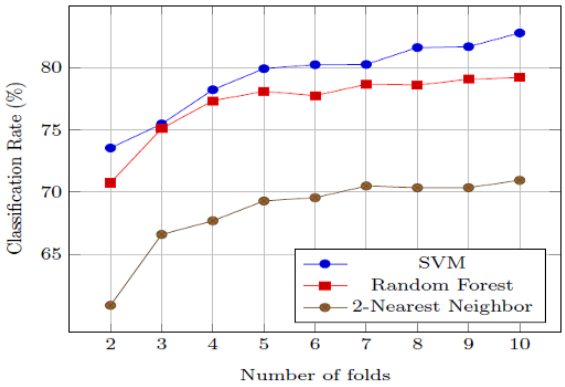

FIGURE 6. Evolution of the classification rate on the Emilya Database with the increasing number of folds for the *k*-fold cross validation method, with three different classifier algorithms. This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is trained on the remaining k-1 folds.

- Support Vector Machine (SVM) with χ^2 kernel;
- Random Forest with 100 trees;
- 2-Nearest neighbor based on Euclidean distance.

It compares with the state-of-the-art methods giving recognition accuracy and Kappa index.

Figure 6 shows the influence of the size of the training set on the performance of the three classifiers used in evaluation, precisely achieved on Emilya database.

Table 3 shows confusion matrix of the different emotions and motions from the Emilya database. This table illustrates that proposed feature vector correctly discriminates different expressions across heterogeneous motions. Moreover, it can be observed that neutral expression is the best recognized emotion from the Emilya database. This result validates proposed approach that theorizes to synthesize neutral motion from the expressive motion. ''Sad'' expression achieves the lowest recognition rate. This could be explained as expression of ''Sad'' is not well represented in the Emilya database.

TABLE 3. Confusion Matrix for the Emilya database. Abbreviations are N for Neutral and PF for Panic Fear. Results are obtained with SVM classifier with a k-fold of 10, giving best results in previous Figure 6.

	Angry	Anxious	Joy	N	PF	Proud	Sad
Angry	**0.76**	0.05	0.07	0.02	0.06	0.04	0
Anxious	0.04	**0.79**	0.04	0.02	0.07	0.04	0
Joy	0.05	0.03	**0.77**	0.02	0.08	0.05	0
N	0.02	0.01	0.03	**0.88**	0.02	0.04	0
PF	0.03	0.04	0.05	0.03	**0.84**	0.03	0
Proud	0.03	0.03	0.07	0.04	0.03	**0.80**	0
Sad	0.03	0.06	0.10	0.04	0.05	0.08	**0.64**

TABLE 4. Confusion Matrix for the Emilya database with a resampling filter in order to have a balanced database. An oversampling method is used and results are obtained with SVM classifier with a k-fold of 10. Abbreviations are N for Neutral and PF for Panic Fear.

	Angry	Anxious	Joy	N	PF	Proud	Sad
Angry	**0.90**	0.01	0.02	0.02	0.03	0.02	0
Anxious	0.02	**0.90**	0.03	0	0.04	0.01	0
Joy	0.02	0	**0.90**	0.01	0.03	0.03	0
N	0.01	0.01	0.01	**0.95**	0.01	0.01	0
PF	0.01	0.02	0.02	0.01	**0.92**	0.02	0
Proud	0.02	0.02	0.04	0.02	0.01	**0.89**	0
Sad	0.02	0.03	0.06	0.02	0.03	0.04	**0.80**

To explore this influence of unbalanced database between emotions, Table 4 gives the confusion matrix of the different emotions and motions from the Emilya database when a common resampling filter is applied to the input data in order to balance the dataset. An oversampling method is used, balancing the dataset by increasing the size of rare samples. New rare samples are generated by using repetition on data. From this table, it can be observed that the recognition rate is better for ''Sad'' expression as the dataset is now balanced. Above presented results prove ability and robustness of proposed method in recognizing body expressions.

In summary, the main conclusions that can be drawn from these experiences are:

1) Table 2 shows the comparison of the recognition rate of proposed framework with the state-of-the-art methods. It is noticeable that proposed method achieves body expression recognition accuracy that globally exceeds results obtained by state-of-the-art methods. Moreover, it is important to highlight the fact that we are comparing proposed method with specific methods developed for one specific database containing often only one type of movement. Whereas, proposed method is generic but still outperforms state-of-the-art results on SIGGRAPH database, MPI database and Emilya database.
2) Proposed method obtained recognition rate of 78.6% for MPI database which outperforms the state-of-the-art. The state-of-the-art on this database was our previous work [11] with a non-optimal synthesized neutral motion, i.e. robotic and with different artifacts. With this result, the idea that the more realistic the neutral motion is, the better the recognition rate is, is verified.
3) To the best of our knowledge, no method in the literature on body expression recognition has tested complete Emilya database. Proposed method achieved body expression recognition rate of 82% on it. This is an

encouraging result as Emilya database is relatively large and contains lot of different motions and expressions. This database presents a challenging problem as variety and number of motions in this database are heterogeneous.

4) Finally, for the UCLIC database, we compared proposed generic method to a specific approach [37] which uses a more conceptually elaborate formulation of the Laban theory combined with a temporal window. [37] achieved good results for UCLIC database 87.30% compared to our result 74%. This could be explained as [37] is well specifically suited for UCLIC database but with no genericity by nature.

We evaluated proposed method on different publicly available databases and outperformed state-of-the-art methods. Proposed approach is generic by nature and not tailor made for one type of motions, input data type or application. Achieved results on different databases prove that this approach is better than the features based methods for a body recognition expression in the wild. Features based methods are not generic and only suitable in a context of input a priori known motions.

VI. DISCUSSIONS AND CONCLUSION

In this paper a novel approach for automatic generic recognition of body expressions through 3D skeleton provided by motion capture data is presented. A novel method for synthesizing realistic neutral motion is introduced, which is used to detect body expression from an expressive input motion. Quantization of neutrality in a motion is formalized by a function. Novel cost function is proposed which is optimized for synthesizing the neutral motion from a given expressive motion. This cost function allows to extract the body expression by computing the difference, in the spectral domain, from the input motion and the synthesized neutral motion.

The results obtained by proposed method show its robustness as the evaluation is done on heterogeneous databases with different motions and expressions. Proposed generic method achieves better recognition rate for body expression on three databases compared to state-of-the-art methods which focused on one specific database. This performance highlights that proposed novel approach is invariant to the input motion as it is based on computing difference between the input motion and the neutral motion produced by proposed method. A slightly more precise analysis shows good recognition rate for positive expressions, these with high valence, whereas negative expressions are sometime confused.

In a more general objective of expression recognition, future work would be to use our approach as a brick in a multi-modal scheme. Body expression recognition could be used in conjunction with facial expression recognition, by synthesizing neutral facial expression as well. Mixing body and face approaches must clearly be more investigated by researchers [45], and moreover, approaches of data fusion with speech and physiological signals [14] could allow to formalize a more adapted framework for in the wild applications.

REFERENCES

[1] R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz, "A novel database of children's spontaneous facial expressions (LIRIS-CSE)," *Image Vis. Comput.*, vol. 83, pp. 61–69, Mar. 2019.
[2] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting Comput.*, vol. 18, no. 6, pp. 1371–1389, Dec. 2006.
[3] G. Bijlstra, W. Rob Holland, R. Dotsch, and H. J. Daniel Wigboldus, "Stereotypes and prejudice affect the recognition of emotional body postures," *Emotion*, vol. 19, no. 2, pp. 189–199, Mar. 2019.
[4] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Human vision inspired framework for facial expressions recognition," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 2593–2596.
[5] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 15–33, Jan. 2013.
[6] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 325–347, Jul. 2019.
[7] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Saliency-based framework for facial expression recognition," *Frontiers Comput. Sci.*, vol. 13, no. 1, pp. 183–198, Feb. 2019.
[8] *Kinect*, Microsoft, Redmond, WA, USA, 2017.
[9] R. Lun and W. Zhao, "A survey of applications and human motion recognition with microsoft kinect," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 29, no. 5, Jul. 2015, Art. no. 1555008.
[10] I. A. Faisal, T. W. Purboyo, and A. S. R. Ansori, "A review of accelerometer sensor and gyroscope sensor in IMU sensors on motion capture," *J. Eng. Appl. Sci.*, vol. 15, no. 3, pp. 826–829, Nov. 2019.
[11] A. Crenn, A. Meyer, R. A. Khan, H. Konik, and S. Bouakaz, "Toward an efficient body expression recognition based on the synthesis of a neutral movement," in *Proc. 19th ACM Int. Conf. Multimodal Interact. ICMI*, 2017, pp. 15–22.
[12] E. O. Brigham and R. E. Morrow, "The fast Fourier transform," *IEEE Spectr.*, vol. 4, no. 12, pp. 63–70, Dec. 1967.
[13] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.
[14] C. Maréchal, D. Mikolajewski, K. tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska, "Survey on AI-based multimodal methods for emotion detection," in *Proc. High-Perform. Modeling Simulation Big Data Appl.*, in Lecture Notes in Computer Science (LNCS), vol. 11400. Springer, Mar. 2019, pp. 307–324.
[15] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing body posture with facial expressions for joint recognition of affect in Child–Robot interaction," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4011–4018, Oct. 2019.
[16] T. Randhavane, A. Bera, K. Kapsaskis, R. Sheth, K. Gray, and D. Manocha, "EVA: Generating emotional Behavior of virtual agents using expressive features of gait and gaze," in *Proc. ACM Symp. Appl. Perception*, Sep. 2019, pp. 1–10.
[17] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Framework for reliable, real-time facial expression recognition for low resolution images," *Pattern Recognit. Lett.*, vol. 34, no. 10, pp. 1159–1168, Jul. 2013.
[18] S. Li and W. Deng, "Deep facial expression recognition: A survey," 2018, *arXiv:1804.08348*. [Online]. Available: https://arxiv.org/abs/1804.08348
[19] A. Mehrabian and J. T. Friar, "Encoding of attitude by a seated communicator via posture and position cues," *J. Consulting Clin. Psychol.*, vol. 33, no. 3, p. 330, 1969.
[20] Z. Cao, G. H. Martinez, T. Simon, S. E. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, Jul. 17, 2019, doi: 10.1109/TPAMI.2019.2929257.
[21] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image Vis. Comput.*, vol. 60, pp. 4–21, Apr. 2017.
[22] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *IEEE Trans. Image Process.*, vol. 29, pp. 15–28, 2020.

[23] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, early access, Oct. 16, 2019, doi: 10.1109/TAFFC.2018.2874986.

[24] Y. Chen, L. Yu, K. Ota, and M. Dong, "Hierarchical posture representation for robust action recognition," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 5, pp. 1115–1125, Oct. 2019.

[25] M. Karg, K. Kuhnlenz, and M. Buss, "Recognition of affect based on gait patterns," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1050–1061, Aug. 2010.

[26] A. Barliya, L. Omlor, M. A. Giese, A. Berthoz, and T. Flash, "Expression of emotion in the kinematics of locomotion," *Exp. Brain Res.*, vol. 225, no. 2, pp. 159–176, Mar. 2013.

[27] C. L. Roether, L. Omlor, A. Christensen, and A. Martin Giese, "Critical features for the perception of emotion from gait," *J. Vis.*, vol. 9, no. 6, pp. 1–32, 06 2009.

[28] L. Omlor and M. A. Giese, "Extraction of spatio-temporal primitives of emotional body expressions," *Neurocomputing*, vol. 70, nos. 10–12, pp. 1938–1942, Jun. 2007.

[29] T. Randhavane, A. Bera, K. Kapsaskis, U. Bhattacharya, K. Gray, and D. Manocha, "Identifying emotions from walking using affective and deep features," 2019, *arXiv:1906.11884*. [Online]. Available: https://arxiv.org/abs/1906.11884

[30] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "STEP: Spatial temporal graph convolutional networks for emotion perception from gaits," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 2, Apr. 2020, pp. 1342–1350.

[31] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *Affective Computing and Intelligent Interaction*, in Lecture Notes in Computer Science (LNCS), vol. 4738. Springer, 2007, pp. 59–70.

[32] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Methodology for assessing bodily expression of emotion," *J. Nonverbal Behav.*, vol. 34, no. 4, pp. 223–248, Dec. 2010.

[33] E. Volkova, S. de la Rosa, H. H. Bülthoff, and B. Mohler, "The MPI emotional body expressions database for narrative scenarios," *PLoS ONE*, vol. 9, no. 12, Dec. 2014, Art. no. e113647.

[34] S. Senecal, L. Cuel, A. Aristidou, and N. Magnenat-Thalmann, "Continuous body emotion recognition system during theater performances: Continuous body emotion recognition," *Comput. Animation Virtual Worlds*, vol. 27, nos. 3–4, pp. 311–320, May 2016.

[35] W. Wang, V. Enescu, and H. Sahli, "Adaptive real-time emotion recognition from body movements," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 1–21, Jan. 2016.

[36] A. Truong, H. Boujut, and T. Zaharia, "Laban descriptors for gesture recognition and emotional analysis," *Vis. Comput.*, vol. 32, no. 1, pp. 83–98, Jan. 2016.

[37] S. Dewan, S. Agarwal, and N. Singh, "Laban movement analysis to classify emotions from motion," in *Proc. 10th Int. Conf. Mach. Vis. (ICMV)*, Apr. 2018, Art. no. 106962Q.

[38] S. Xia, C. Wang, J. Chai, and J. Hodgins, "Realtime style transfer for unlabeled heterogeneous human motion," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–10, Jul. 2015.

[39] N. Fourati and C. Pelachaud, "Emilya: Emotional body expression in daily actions database," in *Proc. LREC*, 2014, pp. 3486–3493.

[40] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 2002.

[41] N. Hansen, "The CMA evolution strategy: A tutorial," 2016, *arXiv:1604.00772*. [Online]. Available: https://arxiv.org/abs/1604.00772

[42] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, Jul. 2016.

[43] N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 90–101, Jan. 2018.

[44] M. Ersin Yumer and J. Niloy Mitra, "Spectral style transfer for human motion between independent actions," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–8, Jul. 2016.

[45] T. Keshari and S. Palaniswamy, "Emotion recognition using feature-level fusion of facial expressions and body gestures," in *Proc. Int. Conf. Commun. Electron. Syst. (ICCES)*, Jul. 2019, pp. 1184–1189.

ARTHUR CRENN received the Ph.D. degree in computer vision from Université Claude Bernard Lyon 1, France, in 2019. His research interests include computer graphics and machine learning.

ALEXANDRE MEYER received the Ph.D. degree in computer science from Université Grenoble 1, France, in 2001. From 2002 to 2003, he was a Postdoctoral Fellow with the University College London. Since 2004, he has been working as an Associate Professor with Université Claude Bernard Lyon 1, France, and a member of the LIRIS research lab. His current research concerns computer animation and computer vision of characters.

HUBERT KONIK received the Ph.D. degree in computer science from Université Jean Monnet, in 1995. He is currently an Associate Professor with Télécom Saint-Etienne and a member of Image Science and Computer Vision team, Laboratoire Hubert Curien, Saint-Etienne, France. His research interests are focused on image processing and analysis, more particularly content aware image processing for new services and usages.

RIZWAN AHMED KHAN received the Ph.D. degree in computer vision from Université Claude Bernard Lyon 1, France, in 2013. He has worked as a Postdoctoral Research Associate with Laboratoire d'InfoRmatique en Image et Systemes d'information (LIRIS), Lyon, France. He is currently working as a Professor with Barrett Hodgson University, Karachi, Pakistan. His research interests include machine learning, computer vision, image processing, pattern recognition, and human perception.

SAIDA BOUAKAZ received the Ph.D. degree from Joseph Fourier University, Grenoble, France. She is currently a Full Professor with the Department of Computer Science, Université Claude Bernard Lyon 1, France. Her research interests include computer vision and graphics including motion capture and analysis, gesture recognition, and facial animation.