Modeling Multiple Temporal Scales of Full-body Movements for Emotion Classification

Cigdem Beyan, Sukumar Karumuri, Gualtiero Volpe, Antonio Camurri, and Radoslaw Niewiadomski

Abstract—This work investigates classification of emotions from full-body movements by using a novel Convolutional Neural Network-based architecture. The model is composed of two shallow networks processing in parallel when the 8-bit RGB images obtained from time intervals of 3D-positional data are the inputs. One network performs a coarse-grained modelling in the time domain while the other one applies a fine-grained modelling. We show that combining different temporal scales into a single architecture improves the classification results of a dataset composed of short excerpts of the performances of professional dancers who interpreted four affective states: anger, happiness, sadness, and insecurity. Additionally, we investigate the effect of data chunk duration, overlapping, the size of the input images and the contribution of several data augmentation strategies for our proposed method. Better recognition results were obtained when the duration of a data chunk was longer, and this was further improved by applying balanced data augmentation. Moreover, we test our method on other existing motion capture datasets and compare the results with prior art. In all experiments, our results surpassed the state-of-the-art approaches, showing that this method generalizes across diverse settings and contexts.

Index Terms—Emotion recognition, convolutional neural network, full-body movements, kinematics, multiple temporal scales, motion capture

1 Introduction

Several studies have acknowledged the importance of expression dynamics for perception and automatic recognition of emotions [1], [2], [3], [4]. In particular, the expressive qualities of full-body movements, i.e., how a movement is performed, provide significant information about the emotional state of a person. Among many others, Wallbott [5] showed that emotions such as "hot anger" and "elated joy" are characterized by high movement activity and dynamics as well as expansive movements while emotions such as "contempt" and "sadness" are characterized by low movement activity and dynamics. Similarly, it is possible to recognize emotions just from point-light displays of arm movements, i.e., from movement dynamics, as shown in [1].

Extracting expressive qualities of a movement conveying an emotion requires temporal analysis. At the same time, there is no gold standard regarding the minimal observation time needed to perceive an expressive quality nor to detect it automatically. Regarding that, Camurri et al. [6] presented a conceptual framework for the analysis of expressive qualities of the movements. Inspired by previous research on human movement perception and dance theories (e.g., Laban Effort [7]), the authors postulate that computational models of expressive qualities should operate on different temporal scales. The first layer of their framework [6] consists of low-level features (e.g., velocity) computed instantaneously while middle and high-level features (such as impulsivity and fluidity) are computed on larger temporal scales, varying from 0.5 to 5-seconds. That framework [6] finds empirical confirmation in a recent functional magnetic resonance imaging (fMRI) study [8], showing that low-level features are processed by a different part of the brain than mid-level features. Other recent works (see for example the European FET PROACTIVE Project EnTimeMent, http://entimement.dibris.unige.it) also refer to the importance of different temporal scales in movement analysis and prediction. For example, processing the data at short time intervals is sufficient to detect hand tremors or trembling (e.g., in anxiety [9]) whilst longer time intervals are required to identify fluid and large full-body movements (e.g., in lightness [10]).

Motivated by these findings, in this paper, we propose a novel approach to modeling the dynamics of full-body movement data represented on multiple temporal scales for the emotion recognition task. A motion capture (MoCap) system was used to collect positional data, which contained short excerpts of the performances of professional dancers who interpreted four affective states: anger, happiness, sadness, and insecurity. We focus on dancers' improvised movements as they are characterized by high complexity and versatility and involve a much larger set of movements compared to the regular day-to-day activities. The dancers use their physical and motor abilities to endow emotional meaning to the movements through the modulation of movement dynamics. Dancers' movements are here context-free, i.e., they are not constrained or limited by the context and the surrounding objects, which is a common issue in everyday activity datasets. Dancers were asked to express emotions without using specific actions or stereotype emotion emblems, and they did not perform any specific action.

In our approach, movement dynamics at two different but related temporal scales are processed jointly by a two-branch neural network architecture. Our proposed method learns simultaneously both features at a middle-level (i.e., fine-grained such as 0.5-, 1-

- C. Beyan is with the Department of Information Engineering and Computer Science (DISI), University of Trento, Povo-Trento, Italy. E-mail: cigdem.beyan@unitn.it.
- S. Karumuri is with Casa Paganini - InfoMus, DIBRIS, University of Genoa, Genoa, Italy. Email: kai.sukumar@gmail.com.
- G. Volpe is with Casa Paganini - InfoMus, DIBRIS, University of Genoa, Genoa, Italy. Email: gualtiero.volpe@unige.it.
- A. Camurri is with Casa Paganini - InfoMus, DIBRIS, University of Genoa, Genoa, Italy. Email: gualtiero.camurri@unige.it.
- R. Niewiadomski is with the Department of Psychology and Cognitive Science, University of Trento, Rovereto, Italy, and COgNiTive Architecture for Collaborative Technologies (CONTACT) Unit, Istituto Italiano di Tecnologia, Genoa, Italy. E-mail: r.niewiadomski@unitn.it.

Manuscript received xx xx, 2020; revised xx xx, xx.

seconds temporal scale) and at a high-level (i.e., coarse-grained such as 4-seconds scale). The network architecture consists of two shallow Convolutional Neural Networks (CNN) processing in parallel, where inputs are 8-bit RGB images obtained from various time intervals of 3D-positional data. We investigate the effect of data chunk duration, various data augmentation strategies for the classification of the aforementioned four emotions as well as the performance of the proposed method compared to the prior art. Additionally, the proposed method was tested on two other datasets, which a) contain more emotion classes including non-basic emotions (from 8 to 12 emotions), b) were captured in different contexts (i.e., contemporary dance and daily-living actions) and c) contain full-body motion performed by multiple participants (from 6 to 12 participants).

The rest of the paper is organized as follows. Related works on automatic full-body movements classification and particularly emotion recognition are discussed in Section 2. Our dataset and the data representation method are introduced in Sections 3 and 4, respectively. Section 5 describes the proposed method. The details of the experimental analysis is given in Section 6 while Section 7 includes an ablation study and discusses the performance of the proposed method within a comparative study performed on our as well as other available datasets. Finally, in Section 8 we conclude the paper with a summary, list of findings and discussions including future research.

2 Related Work

There has been a growing interest in *automatic classification of full-body movements*. Majority of the works have focused on automatic recognition of a pre-defined set of *activities* or *gestures*. However, in this paper, we target classification of *emotions* from full-body movement data. Thus, below, we only briefly discuss the action and gesture recognition studies and mainly focus on the prior art of emotion recognition. Finally, we review the affective computing studies for modeling multiple temporal scales.

2.1 Action and Gesture Recognition

Ha and Choi [11] presented a CNN-based human activity recognition method that performs better than Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) for 12-activity classes (e.g., standing still, sitting and relaxing, lying down, walking) whose data was collected by accelerometers and gyroscopes. In [12], a CNN model is used for action recognition from MoCap data. The captured data is represented as images such that the joint positions constitute the x-axis and the time information constitute the y-axis of the image. That method [12] was applied to two standard datasets while CNN showed significantly better results compared to hand-crafted features. In [13], a deep encoder-decoder architecture was applied for classification and prediction of activities in the CMU MoCap database, which contains 2230 recordings of different physical activities such as walking, running, and punching, resulting in 78% accuracy for nine classes. Wan et al. [14] present a Bidirectional Long Short-Term Memory (Bi-LSTM) network for large-scale isolated and continuous gesture recognition, showing a remarkable performance. There is a large number of other studies performing action and/or gesture recognition from full-body movements. Interested readers can refer to the survey papers in which the data is captured by RGB cameras [15], depth sensors [16], [17], and wearable inertial sensors [18].

2.2 Emotion Recognition

Emotions expressed through full-body can be perceived from 1) a static full-body pose (e.g., a forward head and chest bend expressing the sadness [19]), 2) specific gestures being emotion emblems (e.g, raising arms and hands-on-hips are the gestures of pride [20]), and 3) expressive quality of the movement (e.g., performing expanse movement in anger [3]). This work focuses on the third aspect while the second aspect can be addressed with the methods described in Section 2.1, and the first aspect does not rely on the temporal data.

Several challenges have been mentioned in the literature regarding emotion recognition from full-body MoCap data [21], [22], [23]. The contextual and interpersonal differences in expressing and perceiving affect makes emotion recognition complex [22], [23]. Consequently, it becomes harder to obtain reliable ground-truth data, which is needed to develop automatic recognition methods. Full-body affective expressions may differ between individuals in both intensity and quality level due to numerous factors, e.g., personality, physical capacity, and personal experience. Such differences might cause low accuracy for person-independent automatic recognition [22]. Additionally, existing MoCap datasets are usually rather small, due to the effort needed to collect and most importantly annotate such data with a high reliability. This requires not only relatively expensive and sophisticated hardware but also a lot of post-processing, which might even include manual data cleaning. Consequently, it is very important to develop shallower machine learning methods that are able to efficiently deal with limited data.

Indeed, the majority of the studies in this context still rely on hand-crafted features and apply learning methods such as SVMs and Random Forests [24], [25], [26], [27]. For example, Castellano et al. [24] classify *acted* emotional states using the movement features (motion quantity, velocity, movement fluidity and so forth) extracted from visual data. A set of temporal aggregators is applied to these low-level features to describe their dynamics, which are later classified in terms of four emotions. Piana et al. [27] use 3D-motion data of full-body movements and defines a number of low-level (e.g., kinematics of a single joint) and high-level (e.g., contraction index, impulsiveness) features, which are modelled by an SVM classifier. The contribution of temporal features (e.g., regularity of a motion profile, overall or single gesture phase impulsiveness) and multi-level body cues (e.g., based on Body Action and Posture Coding System [28]) to automatic emotion classification were investigated by Fourati et al. [25] on a dataset composed of 8 daily-life actions (e.g., walking with/without objects in hands, moving books on a table) performed with 8-states (anxiety, pride, joy, sadness, panic fear, shame, anger and neutral).

Daoudi et al. [29] represent the 3D-skeleton data in the Riemannian manifold by integrating covariance operator and use this representation with a Nearest Neighbour classifier to differentiate between angry, fearful, joyful, neutral and sad walks. Kacem et al. [30] adapt the idea of using representations parametrized in the Riemannian manifold, followed by a temporal warping and SVM. That method was used for action recognition from 3D-data, emotion recognition from 3D-body movements and 2D-facial expression recognition, showing boosted performances in all these cases. In [31], a 3-layered Recurrent Neural Network (RNN) is used to perform emotion classification from MoCap data of daily activities such as clapping, drinking, throwing, and waving associated with four-emotions: happy, angry, sad, and neutral. Creen et al. [32] present a method, which synthesizing neutral motion, is used to detect body expression from the 3D-skeleton provided by MoCap data. Neutrality in a motion is quantized through a cost function, and then the difference between body expression of other emotions and synthesized neutral emotion is calculated.

2.3 Modeling Multiple Temporal Scales

The existing affective computing works integrating multiple temporal scales, exclusively rely on analyses of *facial expressions* in *videos* or by *audio processing*. For instance, Yun et al. [33] present an engagement detection method processing facial videos with a CNN architecture that includes a layer modeling the temporal long and short-term data dynamics. Chanti et al. [34] use the combination of 3D-CNN to model short-term spatio-temporal features and Convolutional-LSTM to learn global spatio-temporal features for video-based facial expression analysis. Similarly, a combination of CNN and LSTM models are recently tested for affect recognition from audio and video facial expressions data [35], [36].

In terms of architecture design, targeted problem and dataset, the most similar work to our study is [37]. The authors [37] compare data representation methods: coarse position format, fine position format, logistic position format, and logistic velocity format by applying a shallow CNN architecture for classification of affect from full-body movements. Herein, we use an extended version of the dataset introduced in that study [37] and rely only on the logistic position format (described in Section 4) as it performed the best out of all others analyzed in [37]. Unlike [37], we explore the effectiveness of using *multiple temporal scales* for emotion recognition.

3 Our Dataset

Our dataset is composed of four affective states: angry, happy, sad and insecure. The choice of labels was inspired by previous studies such as [38] where the images displaying bodily emotions of four basic emotions: anger, happiness, sadness, and fear were correctly categorized at least 85% of all the cases. The same set of four labels was also considered in other studies on perception of emotions from static images and videos [39], [40]. We replaced “fear” with “insecure”, which is not among basic emotions, but shares some characteristics with the former (e.g., both are reactions to threats). The advantage of "insecurity" (i.e., a reaction to abstract threat) is that it can be easier to express with dance than fear (i.e., a reaction to immediate and concrete threat).

TABLE 1: The total number of segments and overall segment duration in seconds for each emotion class in our dataset.

Emotion	Number of Segments	Total Duration (seconds)
Angry	16	176
Happy	17	334
Insecure	18	292
Sad	10	283
Total	61	1085

Two professional dancers participated in the data collection. They were asked to portray an emotion in a free movement improvisation, not necessarily dance, avoiding stereotype movements and specific actions.[1] Each recording session was 1-minute long, on average. A team of three experts selected segments of various duration from each recording. Segments that display 1-type of emotion as agreed by all experts were kept, while segments that do not display any clear emotion were discarded. The selected segments have an average duration of 17.8 seconds. The number of selected segments and their duration are summarized in Table 1.

TABLE 2: The order of the markers (1-30) and their correspondence with the body parts.

Body Part	Order index - Short Label (Description)
Head & Torso	1 - ARIEL (Top Head), 2 - C7 (7th Cervical Vertebra), 3 - T5 (5th Thoracic Vertebra), 4 - STRN (Sternum), 5 - CLAV (Xiphoid Process), 6 - BWT (Sacrum Bone), 7/8 - LBWT/RBWT (Left/Right Pelvic Bone),
Left Arm	9 - LSHO (L.Shoulder), 10 - LBUPA (L. Upper Arm), 11 - LELB (L. Elbow), 12 - LIWR (L. Wrist), 13 - LPLM (L. Palm), 14 - LINDX (L. Index Finger)
Right Arm	15 - RSHO (R. Shoulder), 16 - RBUPA (R. Upper Arm), 17 - RELB (R. Elbow), 18 - RIWR (R. Wrist), 19 - RPLM (R. Palm), 20 - RINDX (R. Index Finger)
Left Leg	21 - LFTHI (Left Thigh), 22 - LKNI (Left Knee), 23 - LANK (Left Ankle), 24 - LHEL (Left Heel), 25 - LMT1 (Left 1st Meta Tarsal)
Right Leg	26 - RFTHI (Right Thigh), 27 - RKNI (Right Knee), 28 - RANK (Right Ankle), 29 - RHEL (Right Heel), 30 - RMT1 (Right 1st Meta Tarsal)

4 Data Representation

A Qualisys MoCap System was used for creating the dataset. 30 markers were attached to the body of the dancers. Sample rate was 100 frames per second (fps). Markers were split into five sets: head and torso, left arm, right arm, left leg, and right leg. The markers were re-ordered within each group according to their position in the body (from top to bottom of the body). These were then arranged as shown in Table 2. Missing values were interpolated by using polynomial interpolation.

4.1 Image Construction

Data consisting of the 3D-positions of 30 markers at 100 fps was converted into RGB images, which is a common input format for CNNs. This includes dividing the MoCap segments as identified by the experts, which have a variable duration, into chunks of fixed duration. Then, a chunk of data is converted into an RGB image. Various values were tested for the duration of a single chunk while overlapping in time was also applied. The resulting number of images for each setting is given in Table 3. For example, for the data chunk having a duration of 4-seconds and applying a 0.5-seconds overlapping to the whole dataset results in 1683 RGB images, in total.

TABLE 3: The total number of images obtained from a certain chunk duration and overlapping. NA stands for "not applied".

Chunk Duration (seconds)	Overlapping (seconds)	Number of Images
0.5	NA	2169
1	0.5	2068
2	NA	513
2	0.5	1946
4	0.5	1683

[1] A video containing sample visualizations of the MoCap data is provided as the supplementary material.

The procedure for constructing RGB images includes body-centered relative normalization, which can be described as follows. The value of marker CLAV at the first frame of each chunk is taken as a point of reference. CLAV is situated on the lower part of the chest on the xiphoid process. So, in the first frame, the position of CLAV is zero. The positions of the other markers are taken with respect to this new origin. In the second data frame, if the dancer moves, the position of CLAV changes, i.e., it is no longer at the origin. This allows us to model not only the movement of the joints with respect to the CLAV, but also the movement of the whole body with respect to its initial position. Additionally, by using body-centered relative positioning, the range of the marker values is reduced, thus, it is no longer required to map all the positions of the work-space. An advantage of this normalization is that it allows the creation of images with an overlap. For example, if we take a 1-second chunk with 0.5-seconds overlap, in the first chunk the origin corresponds to the CLAV position at frame one, while in the second chunk, the origin corresponds to the CLAV position at frame 51 of the whole sequence. Hence, the overlapping portion of two consecutive images contain different values.

Following that, an 8-bit RGB image format is used to represent the data based on the method presented in [12]. In detail, the X, Y and Z coordinates of the markers are associated with the R, G and B layers, respectively. Markers are represented on the y-axis, while the consecutive frames of the sequence are represented on the x-axis. For example, a row of the R layer in the resulting image represents the temporal evolution of the X coordinate of the marker associated with that row. Then, logistic position (LP) is used to fit the information in this 8-bit image format, as this method was shown to be the most efficient mapping in our previous work [37].

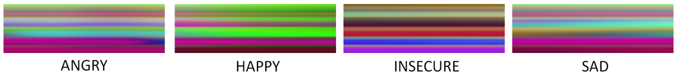


Fig. 1: RGB images in *logistic position format* with their emotion labels. These examples encode 1-second of data (x-axis), and 30 markers (y-axis).

Logistic Position Image Format. While an 8-bit image allows 256 values in the range of 0 to 255, the marker positions are provided in millimeters for high accuracy, hence even with relative positioning, the range of 256 values is insufficient to fit all the data. The approach that we use is based on human perception. In detail, humans are quite capable of noticing differences in lower frequencies, but not so good at identifying high frequency components. Hence, high frequency components can be mapped to a single quantum value and lower frequency components can be mapped to a larger number of values. Inspired by this observation, we use a logistic function that maps the positions into the -127 to +127 interval. The function is given as follows:

$$R = \left\lceil \frac{255}{1+e^{-L(Q)}} \right\rceil \tag{1}$$

where R represents the new marker value, Q represents the original value obtained from relative position extraction and L was a constant selected empirically. Shortly, the input values closer to the origin are mapped to a larger range of output values. Some examples of the resulting images in logistic position format for the 4-emotion classes of our dataset are given in Fig. 1.

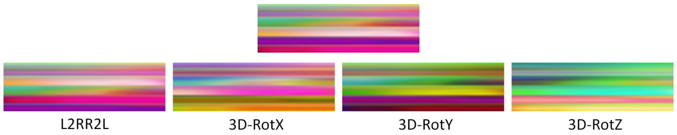


Fig. 2: Original image in *logistic position format* (top) and the corresponding new images obtained by applying data augmentation: (a) L2RR2L, (b) 3D-RotX, (c) 3D-RotY, (d) 3D-RotZ (bottom).

4.2 Data Augmentation

Given a data chunk D represented by the X, Y and Z coordinates of the 30 markers, we obtain a new chunk by:

- Swapping the data associated to the left-side body markers with the right-side body markers (L2RR2L): The values of the left-arm markers (i.e., LSHO, LBUPA, LELB, LIWR, LPLM, LINDX) are assigned to the right-arm markers (i.e., RSHO, RBUPA, RELB, RIWR, RPLM, RINDX), while the values of the right-arm markers are assigned to the left-arm markers. Similarly, the values of the left-leg markers (i.e., LFTHI, LKNI, LANK, LHEL, LMT1) are exchanged with the values of the right-leg markers (i.e., RFTHI, RKNI, RANK, RHEL, RMT1) as well as the values of the LBWT marker is exchanged with the values of the RBWT marker and vice versa.
- Applying 3D rotation around the X-axis (3D-RotX): We obtain a new data Dx', which results from the rotation of the original data D around the X-axis for 90 degrees.
- Applying 3D rotation around the Y-axis (3D-RotY): A new data Dy', which results from the rotation of the original data D around the Y-axis for 90 degrees, is obtained.
- Applying 3D rotation around the Z-axis (3D-RotZ): A Dz', which results from the rotation of the original data D around the Z-axis for 90 degrees, is obtained.

L2RR2L assumes that the person can display the same expressive quality by moving left and right part of his/her body. Strategies similar to 3D-RotX, 3D-RotY, and 3D-RotZ have been recently used in [41] for automatic detection of reflective thinking. Herein, we have adapted them for emotion classification, and they are based on the assumption that emotion recognition should be invariant to 3D-rotations. In other words, augmentations applied add some variability to the data, which can be corresponding to the potential real-life situations, e.g., various viewpoints that might not be covered during data collection. The aforementioned augmentation strategies have been designed considering the content (i.e., full-body movements), instead of applying traditional methods such as image cropping, padding, or horizontal flipping. Fig. 2 shows examples of the original and the corresponding augmented data represented in the logistic position image format. The data obtained as a result of the aforementioned augmentations retain the labels of the original chunks since the expressed affective states are unaltered.

Given one of these four strategies, we apply:

- Augmenting every training image (A_ALL): This creates the biggest training set out of all.
- Augmenting 10% of the images belonging to each emotion class (A_10%).
- Balanced Training (A_BALANCED): We calculate the number of images belonging to each emotion class in the original training set. The majority class (i.e., the class having the maximum number of data) is not augmented, but all other classes are

augmented, resulting in a final training set containing an equal number of images from each emotion class.

In case of using all four augmentation strategies (L2RR2L, 3D-RotX, 3D-RotY, 3D-RotZ) together, we randomly determine the strategy to be applied for the generation of a single new image.

One reason to apply data augmentation is to increase the size of the training set, which might result in a better performing model (e.g., as applied in [42]). But it is often difficult to foresee what the quantity of the augmented data should be. Among A_ALL, A_10%, and A_BALANCED the highest number of augmented data is obtained by applying A_ALL while the smallest number of augmented data is obtained by applying A_BALANCED. Another important issue is having class imbalanced data that causes a tendency of the trained model to bias towards the majority class [43]. The possible negative effects of having class imbalance are handled by applying data augmentation with the A_BALANCED strategy where quantity of the augmented data is not a parameter (as in A_10%), but instead it is in terms of the quantity of the data belonging to the majority class. Overall, by applying the aforementioned data augmentations, we aim to improve the overall classification performance while performing equally well prediction for each emotion class.

5 Proposed Method

In this study, we have spatio-temporal data to be processed and classify in terms of some set of emotion classes. While CNN is best known for its application to the classification of static images (i.e., only spatial data), it is also an appropriate technique to process temporal data [33], [12]. There are some benefits of using a CNN model for our task as compared to other machine learning methods. For example, similar dynamic patterns observed in different parts of the body can be identified using the same filter. Thus, we do not need to train separately the network to detect the same features in different parts of the body as it could be in case of other machine learning methods. Also, by using CNN, the filter is able to detect a quality (i.e., emotion cue) in the data irrespective of when the corresponding motion occurs. CNN allows us to bypass the manual extraction of movement features and let the network decide the best features for classification. Moreover, compared to other popular deep learning methods such as RNNs, CNNs typically require less computation and memory and can provide better classification results for a smaller data size [37].

Our proposed method employs a two-branch architecture. It consists of two CNNs, each of them is composed of three convolutional layers followed by fully connected layers. It is illustrated in Fig. 3. One key feature in our architecture is the shape of the convolutional filters. Instead of using square filters, which is more common, our filters are extended along the time-axis to form 3×5 rectangles. The reason for having rectangular filters is that we expect the network to learn and extract features in the time domain rather than among successive markers. It is also important to highlight that the input image is always rectangular. The first convolution is applied to the input image with 16 filters. The "same" padding, which makes the size of outputs the same as that of inputs, is used. A max pooling operation with a stride of two is performed, which reduces both x and y dimensions by half. The obtained result is given to the next layer after applying a ReLU function.

This series of operations is repeated in the next two convolutional layers, but with increasing numbers of filters. In other words, in the second layer 32 filters are used while in the third layer 64 filters are used. Such increase in the number of filters allows us to identify more complex features in the deeper layers. After the third and final max pooling and ReLU layer, the image size is reduced to 3×12, but with 64 layers. The output is then flattened out. In this two-branch architecture, separate convolutional layers are used before the weights are flattened out and these weights are added together in a fully connected layer. Finally, another fully connected layer of dimension 4×1 is used as the output layer such that each output value corresponds to a single emotion class. A softmax function in the output layer determines the final emotion class for the given input image.

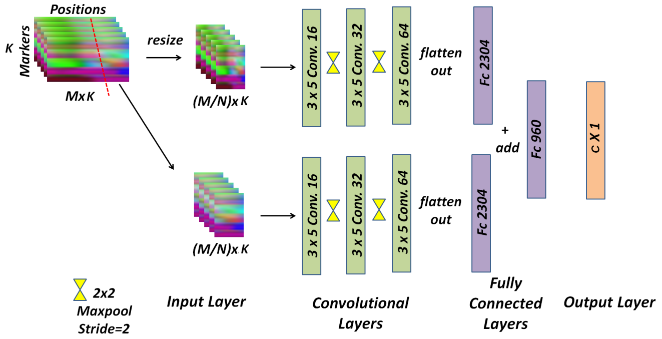


Fig. 3: The architecture of our proposed method. K stands for the number of markers and C represents the number of emotions.

The input of our model at a time is a part of a data segment in the form of two RGB images in logistic position image format (see Fig. 3). The L constant introduced in Eq. 1 was taken as 0.0035, which was decided empirically, for the experiments applied on our dataset. The two-branches of the proposed architecture take images I'' and I', both having the size $M/N \times K$ (which is determined empirically in Section 7 while K is defined by the number of markers). Starting from an image I (having the size $M \times K$) corresponding to a certain data chunk duration and I' that is a part of I corresponding to the last portion of I (e.g., the last quarter), thus its size is $M/N \times K$ (e.g., $M/4 \times K$), where $N \in \mathbb{Z}$ and $1 < N < M$, first, image resizing with bi-cubic interpolation is applied to I, resulting in I'', which is $M/N \times K$. Then, I'' and I' are given to the network as the inputs, simultaneously.

Our architecture as well as the way of pre-processing the input data are inspired by the human perception and the studies in [6], [8]. In detail, one branch learns the temporal patterns of the longer chunks (so-called global learning) while the other branch learns more local temporal patterns, which are shorter (so-called local learning). In other words, the branch that processes I'' performs a coarse-grained modelling in the time domain while the other branch applies a fine-grained modelling by processing I'. It is important to highlight that by applying overlapping, both branches process all the data of the given segment. We illustrate this in Fig. 4.

Below, we investigate whether emotion classification from full-body movements can be performed more accurately as compared to not considering multiple temporal scales by applying our proposed two-branch architecture as well as we compare the performance of the proposed method with the prior art.

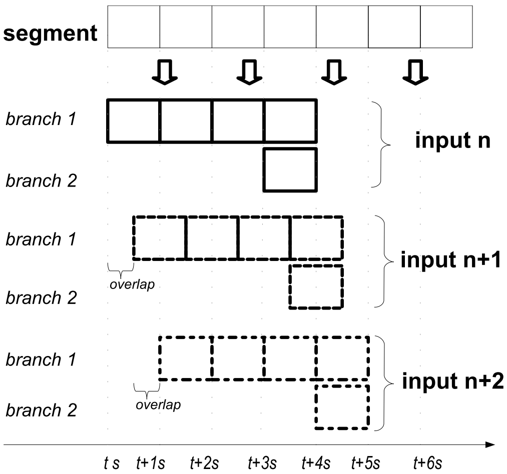


Fig. 4: Illustration of data segmentation and overlapping

6 Experimental Analysis

To train the proposed method, mini-batch gradient descent was used with Adam optimizer. Dropout regularization was performed on the fully connected layers. Early stopping was applied to determine the number of epochs such that we stopped the training where the validation error started to increase significantly. In the end, the learning rate was fixed to 0.0009 and increased up to 0.00135, mini-batch size was fixed to 32 or 64, the number of epochs was fixed to 120 or 200 together with the values of filter sizes and convolution operations mentioned in Section 5. The values of all these parameters were set empirically.

A hybrid cross validation was performed. This hybrid method involves a k-fold cross validation and a Monte Carlo cross validation. In the inner loop, a 5-fold cross validation was applied. In the outer loop, the data was shuffled randomly and sent to the inner loop. This enables the creation of different segments in the 5-fold and hence provides a measure of randomness akin to the Monte-Carlo method. This was performed for ten times bringing on 50 (10×5) results. The results given in the next Section are the average of the 50 results of the cross validation.

7 Results

In this section, first, we discuss the effect of chunk duration and overlapping together with the effect of the size of the input images (Section 7.1). Then, the contribution of the applied data augmentation strategies is investigated (Section 7.2). All of these analyses were performed by using a single branch of the proposed architecture (illustrated as the top branch in Fig. 3 with or without resizing), and the corresponding results are given in Table 4 in terms of F1-score. Following that, we report the performance of the proposed method, which analyses the data at multiple temporal scales with a two-branch CNN architecture (Section 7.3, Table 5) and compare its performance with the prior art (Section 7.4, Table 6). All these aforementioned experiments were realized on our dataset (Section 3). Finally, we tested the proposed method on two other datasets containing full-body MoCap data of multiple persons performing dance and daily actions associated with several emotion classes (Section 7.5).

TABLE 4: F1-score of each setting. "NO" stands for not applying augmentation. "w/ ALL" refers to randomly applying one of the four augmentation strategies: L2RR2L, 3D-RotX, 3D-RotY, 3D-RotZ by using a single data chunk. The best of all results is emphasized in bold.

Duration	Overlap	Image Size	Augmentation	F1-score
1	0.5	100x30	NO	77%
1	0.5	100x30	A_ALL	76%
1	0.5	100x30	A_10%	77%
1	0.5	100x30	A_BALANCED w/ L2RR2L	79%
1	0.5	100x30	A_BALANCED w/ 3D-RotX	79%
1	0.5	100x30	A_BALANCED w/ 3D-RotY	78%
1	0.5	100x30	A_BALANCED w/ 3D-RotZ	79%
1	0.5	100x30	A_BALANCED w/ ALL	80%
0.5	No	50x30	NO	74%
0.5	No	50x30	A_BALANCED w/ALL	76%
0.5	No	50x30 to 100x30	NA	74%
0.5	No	50x30 to 100x30	A_BALANCED w/ALL	76%
2	0.5	200x30	NO	80%
2	0.5	200x30	A_BALANCED w/ ALL	81%
2	0.5	200x30 to 100x30	NA	81%
2	0.5	200x30 to 100x30	A_BALANCED w/ALL	83%
4	0.5	400x30	NO	87%
4	0.5	400x30	A_BALANCED w/ ALL	89%
4	0.5	400x30 to 100x30	NO	90%
4	0.5	400x30 to 100x30	A_ALL	89%
4	0.5	400x30 to 100x30	A_10%	90%
4	0.5	400x30 to 100x30	A_BALANCED w/ L2RR2L	91%
4	0.5	400x30 to 100x30	A_BALANCED w/ 3D-RotX	91%
4	0.5	400x30 to 100x30	A_BALANCED w/ 3D-RotY	90%
4	0.5	400x30 to 100x30	A_BALANCED w/ 3D-RotZ	91%
4	0.5	400x30 to 100x30	A_BALANCED w/ ALL	**92%**

7.1 The effect of chunk duration, overlapping, and input image resizing

The duration that is equal to 1-second and overlapping that is equal to 0.5-seconds, resulting in images having size 100×30 was taken as the *baseline* setting since this was the best performing configuration in our earlier work [37]. It is important to recall that choosing a chunk duration influences the size of the images (see Table 3). Consequently, direct comparison of the classification results for different chunk durations might be unfair when the same architecture with the same filter sizes is used, thus we also applied image resizing such that the input images became the same size with the baseline dimensions (100×30).

Applying image resizing did not affect the classification performance when the chunk duration was taken equal to 0.5. However, when the chunk duration was increased to 2- and 4-seconds, 1% (from 80% to 81% F1-score) and 3% (from 87% to 90% F1-score) classification improvements were obtained respectively, after image resizing.

Shorter chunks result in higher numbers of images (see Table 3) while applying overlapping increases the number of images as well. Typically, having more images in the training set might result in better classification performance. It is however important to highlight that longer chunks contain more information and although the final model is trained with the same data, it might be able to learn better. Indeed, the model trained with longer chunks resulted in improved performance (without image resizing: 74%, 77%, 80%, 87% F1-scores for 0.5-, 1-, 2- and 4-seconds data, respectively; with image resizing: 74%, 77%, 81%, 90% F1-scores for 0.5-, 1-, 2- and 4-seconds data, respectively).

7.2 The effect of augmentation

The effects of applying data augmentation according to different strategies were examined for 1) the baseline setting and 2) the setting having data chunk duration equal to 4-seconds and overlapping equal to 0.5-seconds, resulting in images having size 400×30.

For the baseline setting, the A_BALANCED strategy performed the best (the performance changing from 78% to 80% F1-score). Following that, not applying data augmentation (shown as NO) and 10% data augmentation for each emotion class (shown as A_10%) performed equally well (77% F1-score), whilst augmenting every training data performed the worst (76% F1-score) out of all. Balanced training with a randomly selected strategy per data chunk (A_BALANCED w/ ALL: L2RR2L, 3D-RotX, 3D-RotY, or 3D-RotZ) showed the best performance: 80% F1-score. There is no statistically significant ($p-value > 0.05$) performance difference between L2RR2L, 3D-RotX, 3D-RotY, and 3D-RotZ when they are applied individually (78-79% F1-scores).

For the 4-seconds duration with 0.5-seconds overlapping and without image resizing, applying A_BALANCED w/ ALL improved the results from 87% to 89% F1-score. When image resizing was applied, applying data augmentation type A_BALANCED w/ ALL showed the best performance out of all, which improved results from 90% to 92% F1-score. This performance was followed by not applying data augmentation (90% F1-score) and by A_10% for each emotion class (90% F1-score) while A_ALL (89% F1-score) performed the worst. Shortly, the data augmentation trend observed for the baseline setting was the same for the data represented as 4-seconds chunks with 0.5-seconds overlapping.

To sum up, results (Table 4) show that data augmentation has potential to improve the emotion classification performance and the key point is balancing the training data. On the other hand, longer data chunks (implies less training images) bring in better emotion classification results, whilst image resizing applied to longer chunks also contributes positively to the classification. Consequently, for our dataset, the best recognition was observed for 4-seconds chunks resized to 100×30, when balanced data augmentation was applied out of all combinations tested.

7.3 The effect of multiple temporal scales

Given the setting having chunk duration equal to 4-seconds and overlapping equal to 0.5-seconds, resized to 100×30, performed better than any other settings of shorter chunks, we tested our proposed two-branch architecture for that setting. In detail, one branch of the proposed method was fed with images lasting 4-seconds with 0.5-seconds overlapping. Meanwhile, the other branch was fed with the last quarter of the aforementioned images (i.e., they encode the data occurring in the last 1-second of the 4-seconds chunk and image size is equal to 100×30). The images in the former branch were resized to 100×30, while for both branches the same types of data augmentation were applied. The performance boost occurred by the inclusion of the multiple temporal scales is reported in Table 5.

TABLE 5: Performances in terms of F1-score. NA and A_BALANCED stand for "not applied" and "balanced training", respectively. ALL refers to randomly applying one of the four augmentation strategies: L2RR2L, 3D-RotX, 3D-RotY, and 3D-RotZ using a single data chunk.

Method	F1-score
Proposed Method w/out multiple temporal scales	**90%**
Proposed Method w/out multiple temporal scales w/ A_BALANCED ALL	**92%**
Proposed Method w/out augmentation	92%
Proposed Method w/ A_BALANCED L2RR2L	94%
Proposed Method w/ A_BALANCED 3D-RotX	94%
Proposed Method w/ A_BALANCED 3D-RotY	93%
Proposed Method w/ A_BALANCED 3D-RotZ	94%
Proposed Method w/ A_BALANCED ALL	**95%**

Our proposed method performs better than its single-branch version (i.e., without multiple temporal scales) when augmentation is not applied (90% vs. 92% F1-score) as well as when augmentation is applied (92% vs. 95% F1-score). The performance gain which is obtained by integrating multiple temporal scales (from 92% to 95% F1-score) is statistically significant (p-value < 0.05, measured with a t-test on 10 executions) compared to processing the data at a single temporal scale. These results clearly show that processing full-body movement data at multiple temporal scales improves emotion classification.

7.4 Comparison with the prior art

In Table 6, the performance of the proposed method (with the setting described in Section 7.3) is compared with the best results of:

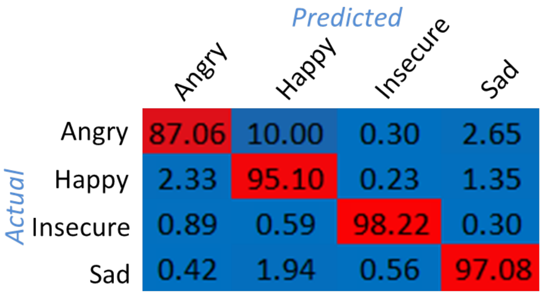
- Our earlier work [37]: A multiple input CNN-based architecture taking the logistic position and the corresponding logistic velocity based image as the simultaneous inputs when the data chunk length is 1-second and 0.5-seconds overlapping is applied.
- A Bi-LSTM Network [14] getting the raw positional data as the input: The input data size at a time corresponds to a vector of 30×3 (30 markers and 3D-positional data) and the length of the data chunk is 1-second while 0.5-seconds overlapping is applied. As a pre-processing step, we applied body-centered relative normalization (see Section 4.1 for its definition). We designed two-hidden layers in the Bi-LSTM network having 64 or 128 or 256 hidden units (when the length of the input data is smaller than the number of hidden units used, it might result in under-fitting the training data. In that case, we also tested having an additional fully connected layer before the LSTM layer to augment to data, which improved the results). These layers were followed by a dense layer and a softmax (giving equal weights to each emotion class) to obtain the probability for each class. We used the Adam optimizer with the batch size 32 or 64. The model was trained for up to 100 epochs with a learning rate of 0.001. As we noticed that this network might be prone to over-fitting, we tested to apply a 50% dropout in the Bi-LSTM layers too.
- Support Vector Machine (SVM) with the flattened images in the logistic position format as the input, and data chunk duration is {1-, 2-, or 4-} seconds with 0.5-seconds overlapping. We used radial basis function (RBF) kernel when the penalty parameter C of the error term is ranging from 0.001 to 10000, and γ kernel coefficient is ranging from 0.001 to 1000.

Bi-LSTMs and SVM-RBF have been frequently applied to process MoCap data of nonverbal behaviors in various contexts including emotion classification [26], [14], [41], [25], therefore, we included them to the comparisons.

Our method outperforms the prior art: [37], [14] and SVM-RBF. Performing better than [37] once again shows the benefits of processing the full-body movement data at *multiple temporal*

TABLE 6: Performance comparisons among proposed method and the prior art in terms of F1-score. The best performance is emphasized in bold.

[37]	Bi-LSTM [14]	SVM	Proposed Method
89%	82%	80%	**95%**



Actual \ Predicted	Angry	Happy	Insecure	Sad
Angry	87.06	10.00	0.30	2.65
Happy	2.33	95.10	0.23	1.35
Insecure	0.89	0.59	98.22	0.30
Sad	0.42	1.94	0.56	97.08

Fig. 5: Confusion matrix (%) of our proposed method in testing.

scales. Additionally, surpassing SVM shows that the designed filters of the proposed method are good at capturing the *spatio-temporal* relationship of the data. Performing better than the Bi-LSTM [14] network shows that our image representation: *logistic position format* can be preferable to using raw 3D-positional data.

In Fig. 5, we also report the confusion matrix of the proposed method corresponding to its performance given in Table 6. For happy, insecure, and sad classes the Correct Classification Rates (CCR) of the proposed method are all above 95% while the lowest CCR, which is 87.06%, was obtained for angry. Angry was mis-classified as happy with 10% classification rate. The relatively higher confusion rate for this pair can be explained by the fact both these emotions are characterized by expansive movements and high movement dynamics [5].

7.5 Experiments on other datasets

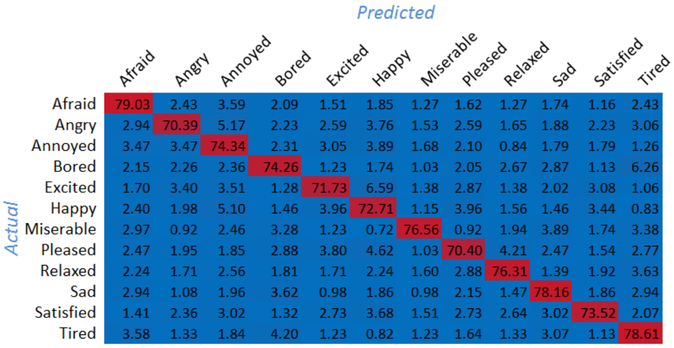
Several multimodal datasets for emotion recognition are publicly available, e.g., [44], [45], [46], [47], but only few contains full-body MoCap data. We evaluate the proposed method on DMCD [48] and Emilya [49] datasets as they contain relatively larger number of full-body motion data. By using them, we aim to show generalizability of our approach within the same (e.g. dance) and different domains. Both dataset also include non-basic emotions (e.g., satisfied, excited and miserable), allowing us to test our method on diverse emotion classes. Below, we explain the applied experimental analysis and then discuss the corresponding results.

7.5.1 Dance Motion Capture Database [48]

DMCD consists of various dance performances recorded with PhaseSpace Impulse X2 MoCap system. We used the contemporary dance sequences performed by six participants (Andria, Elena, Olivia, Sophie, Theodora, and Vasso), having different dance-related backgrounds (theatrical, ballet, gymnastic, and so on). They performed a choreography, each associated to one of 12 emotions: excited, happy, pleased, satisfied, relaxed, tired, bored, sad, miserable, annoyed, angry, and afraid. There are in total 108 performances (12 emotions × 9 since 3 performers did two trials per emotion) corresponding to 614898 3D-points captured with 38 markers.

It is important to note that the number of markers and the position of the markers in the DMCD database [48] are slightly different than our dataset (Section 3). We used 26 markers arranged as follows: 1- Head, 2- Neck1, 3- Neck, 4- Spine1, 5- Spine, 6- Hips, Left Arm: 7- LeftShoulder, 8- LeftArm, 9- LeftForeArm, 10- LeftHand, 11- LeftHandIndex1, 12- RightShoulder, 13- RightArm, 14- RightForeArm, 15- RightHand, 16- RightHandIndex1, 17- LHipJoint, 18- LeftUpLeg, 19- LeftLeg, 20- LeftFoot, 21- LeftToeBase, 22- RHipJoint, 23- RightUpLeg, 24- RightLeg, 25- RightFoot, and 26- RightToeBase. This arrangement is like the one given in Table 2.

We segmented the continuous DMCD data with a window of 100, 200 and 400 frames. This segmentation resulted in images having the width of 100, 200, 400 pixels, respectively, while the height of the images is defined by the number of the selected markers (i.e., 26). These images were given as the inputs to the first branch of the proposed method, which were further resized to the dimension of the images that were the inputs of the second branch (i.e., 25×26, 50×26 and 100×26, respectively). Thus, the proportions between the image sizes of the first (before resizing) and second branch were kept the same as in the previous experiments in Section 7.4 (i.e., the second branch gathers the data which is corresponding to the last quarter of the data given to the first branch). As these settings resulted in training/test images having a size that is similar to the size of the training/test images obtained from our dataset, we were able to use the exact same architecture introduced in Fig. 3 without changing the size of the filters. We kept all the hyper-parameters (mini-batch, epoch, learning rate, dropout and so forth) the same as defined in Section 6. During the creation of images in logistic position format, we applied 50 frames overlapping when L (Eq. 1) was taken as 0.1. We only used the world coordinates data from the DMCD dataset and applied body-centered relative normalization using the *Spine* marker as the point of reference. We applied the cross-validation method described in Section 6 and the balanced data augmentation method in which the four augmentation strategies (L2RR2L, 3D-RotX, 3D-RotY, and 3D-RotZ) were randomly applied to the minority classes.



Actual \ Predicted	Afraid	Angry	Annoyed	Bored	Excited	Happy	Miserable	Pleased	Relaxed	Sad	Satisfied	Tired
Afraid	79.03	2.43	3.59	2.09	1.51	1.85	1.27	1.62	1.27	1.74	1.16	2.43
Angry	2.94	70.39	5.17	2.23	2.59	3.76	1.53	2.59	1.65	1.88	2.23	3.06
Annoyed	3.47	3.47	74.34	2.31	3.05	3.89	1.68	2.10	0.84	1.79	1.79	1.26
Bored	2.15	2.26	2.36	74.26	1.23	1.74	1.03	2.05	2.67	2.87	1.13	6.26
Excited	1.70	3.40	3.51	1.28	71.73	6.59	1.38	2.87	1.38	2.02	3.08	1.06
Happy	2.40	1.98	5.10	1.46	3.96	72.71	1.15	3.96	1.56	1.46	3.44	0.83
Miserable	2.97	0.92	2.46	3.28	1.23	0.72	76.56	0.92	1.94	3.89	1.74	3.38
Pleased	2.47	1.95	1.85	2.88	3.80	4.62	1.03	70.40	4.21	2.47	1.54	2.77
Relaxed	2.24	1.71	2.56	1.81	1.71	2.24	1.60	2.88	76.31	1.39	1.92	3.63
Sad	2.94	1.08	1.96	3.62	0.98	1.86	0.98	2.15	1.47	78.16	1.86	2.94
Satisfied	1.41	2.36	3.02	1.32	2.73	3.68	1.51	2.73	2.64	3.02	73.52	2.07
Tired	3.58	1.33	1.84	4.20	1.23	0.82	1.23	1.64	1.33	3.07	1.13	78.61

Fig. 6: The confusion matrix corresponding to 74.68% F1-score, obtained by applying the proposed method to the DMCD database.

The proposed method tested on the DMCD database results in the confusion matrix given in Fig. 6. It corresponds to the best performance that has 74.68% F1-score. The highest class CCR was obtained for afraid (79.03%), while the CCR obtained for tired (78.61%) and sad (78.16%) classes are the followers. Afraid was mis-classified as annoyed with 3.59% classification rate and it was recognized as other emotion classes with classification rate less than 2.5%. The lowest CCR were obtained for pleased (70.40%) and annoyed (70.39%) classes. Happy, which was classified with 72.71% CCR, was mis-classified as annoyed, excited, pleased, and satisfied with 5.10%, 3.96%, 3.96%, and 3.44% classification rates, respectively. These findings are in line with [50] that applied the Pearson's Correlation coefficients analysis to a subset of the data we used in our analysis, showing that excited, happy, pleased and satisfied are highly correlated and happy is correlated to annoyed and angry as well.

The other classes that were found highly correlated in [50] are: bored, sad and miserable, while satisfied, relaxed and tired are mildly correlated. The proposed method classified sequences labelled as bored with 74.26% CCR, while it mis-classified bored as tired class with 6.26% classification rate. Sad sequences, which were classified with 78.16% CCR, were mis-recognized as bored with 3.62% classification rate. Miserable emotion class that was detected with 76.56% CCR, was recognized as sad with 3.89%, as bored with 3.28% and as tired with 3.38% classification rates. To sum up, our results follow the previous findings [50] such that highly correlated emotion classes were mis-classified with each other more than they were mis-classified with other emotion classes.

7.5.2 *Emilya Dataset [49]*

This is a 3D-MoCap dataset of emotional body expressions during 8 daily actions: simple walking (SW), walking with an object in hands (WH), moving books on a table (MB), knocking (KD), sitting down (SD), being seated (BS), lifting (Lf) and throwing (Th). In total, 12 persons were asked to perform all these actions with 8 states: anxiety, pride, joy, sadness, panic fear, shame, anger and neutral.

The data was captured with 28 markers. To generate images in logistic position format, we ordered the markers as follows: 1- EndSiteHeadX, 2- HeadX, 3- NeckX, 4- Chest4, 5- Chest3, 6- Chest2, 7- Chest, 8- Hips, 9- LeftCollar, 10- LeftShoulder, 11- LeftElbow, 12- LeftWrist, 13- EndSiteLeftWrist, 14- RightCollar, 15- RightShoulder, 16- RightElbow, 17- RightWrist, 18- EndSiteRightWrist, 19- LeftHip, 20- LeftKnee, 21- LeftAnkle, 22- LeftToe, 23- EndSiteLeftToe, 24- RightHip, 25- RightKnee, 26- RightAnkle, 27- RightToe, and 28- EndSiteRightToe. The body-centered relative normalization was applied using the *Chest2* marker. The training/test images (input to first branch) have the size of either 100×28 or 200×28 while 50 frames overlapping was applied. Except *L* (Eq. 1) that was taken as 0.01, other settings (regarding architecture, hyper-parameters, augmentation, second branch image proportions) were all kept the same as those applied on DMCD [48] and our dataset.

We tested the proposed method on individual action classes as well as all sequences of all actions. We compare our results with the state-of-the-art methods: [25], [51] and [32]. It is important to highlight that [25], [51] and [32] utilized two slightly different cross-validation set-ups. We followed the same set-ups as them in one-to-one comparisons. The corresponding results in terms of accuracy are given in Table 7 where we also report the Correct Classification Rates (CCR) of the proposed method for each emotion class.

When tested on individual action classes, the proposed method (PM) surpasses [25], [51] with a considerable extent. This refers to performance improvements with a margin of 2.29-28.59% accuracy. The results of the PM are above 87% accuracy in all actions while it performs the best for BS action (96.59% accuracy). Given the CCR results, one can observe that sadness was recognized with the highest rate for 6 out of 8 actions. On the other hand, the lowest CCR results were obtained for anger, anxiety, and joy. The PM emotion recognition performance, when all actions are considered altogether (ALL1 and ALL2), achieves remarkable results, which are 15.48% better than [25], [51] and 9.11% better than [32]. Joy (87.89% and 85.21%) has the lowest CCR results in overall while Neutral has the highest CCR results (95.22% and 94.78%).

TABLE 7: Performance comparisons (in terms of Accuracy) among the proposed method (PM) and the state-of-the-art approaches [25], [51], [32] on single actions: simple walking (SW), walking with an object in hands (WH), moving books on a table (MB), knocking (KD), sitting down (SD), being seated (BS), lifting (Lf) and throwing (Th) as well as all actions (shown as ALL1 and ALL2) of Emilya dataset [49]. Correct Classification Rates (CCR) of the PM for each emotion class, corresponding to the accuracy score of the PM, are also given. * and $^\times$ stand for the cross validation set-up applied in [25], [51] and [32], respectively. NA stands for "not available". The best performance for each category is emphasized in bold.

	SW*	WH*	MB*	BS*	SD*
[25], [51]	85.00%	84.00%	83.00%	68.00%	68.00%
[32]	NA	NA	NA	NA	NA
PM	**87.29%**	**87.35%**	**92.02%**	**96.59%**	**87.63%**
PM- Anger	85.58%	83.10%	77.09%	96.38%	90.00%
PM- Anxiety	85.11%	85.59%	89.97%	93.24%	82.61%
PM- Joy	85.59%	88.63%	80.45%	92.07%	85.83%
PM- Neutral	84.71%	86.89%	98.04%	96.05%	86.62%
PM- Panic Fear	92.17%	89.38%	89.07%	95.09%	85.81%
PM- Pride	90.78%	85.86%	96.44%	97.78%	86.93%
PM- Sadness	88.42%	92.23%	96.52%	99.70%	92.54%
PM- Shame	85.28%	85.73%	95.54%	98.01%	87.61%
	KD*	Lf*	Th*	ALL1*	ALL2$^\times$
[25], [51]	82.00%	78.00%	79.00%	75.00%	NA
[32]	NA	NA	NA	NA	82.20%
PM	**93.03%**	**90.24%**	**90.10%**	**90.48%**	**91.31%**
PM- Anger	94.55%	81.76%	78.13%	90.85%	90.92%
PM- Anxiety	91.37%	86.48%	86.84%	88.03%	91.04%
PM- Joy	81.06%	76.68%	78.13%	87.86%	85.21%
PM- Neutral	95.64%	93.94%	80.77%	95.22%	94.78%
PM- Panic Fear	97.65%	88.94%	89.33%	89.13%	91.39%
PM- Pride	89.01%	87.90%	86.79%	91.54%	90.48%
PM- Sadness	96.52%	96.81%	98.20%	91.17%	90.67%
PM- Shame	94.49%	95.75%	96.08%	89.79%	94.97%

8 Conclusions and Future Work

In this paper, a novel approach for the classification of emotions from MoCap data has been introduced. Inspired from a recent conceptual model [6] and fMRI studies [8], we introduced a two-branch neural network architecture for learning features from the data at multiple temporal scales, simultaneously. When we evaluated our proposed method on our dataset to classify four emotion classes, it achieved an average F1-score of 95%, presenting 3% improvement as compared to processing data at a single temporal scale. Additionally, the proposed method brings in 6% improvement compared to our previous work [37], which also relies on a multi-branch CNN architecture, but does not investigate the effect of processing multiple temporal scales of the data and instead examines different ways to represent MoCap data as RGB images. The proposed method outperforms Bi-LSTM [14] and SVMs with 13-15% improvement. Additional analysis was carried out to test the effectiveness of the proposed method on datasets having a higher number of emotion classes expressed by several persons. The proposed method achieved up to 91% average accuracy for classification of 8 emotions expressed during various daily actions and the classification accuracy increases up to 97% when considering these actions individually. In both cases our method surpasses remarkably the-state-of-the art. At the same time, the proposed method achieved on average 75% F1-score for 12 emotion classes expressed during contemporary dance.

The main contributions of this study can be summarized as follows:

- A novel two-branch neural network architecture to process full-body MoCap data at multiple temporal scales is presented. In detail, we jointly process movement data represented with two temporal scales. This outperforms its single temporal scale version as well as the state-of-the-art methods. The proposed method is able to achieve remarkable classification performances for a high number of classes including non-basic emotions captured in different contexts (i.e., contemporary dance and daily-living actions) and containing full-body motion performed by multiple participants.
- Several data augmentation methods were compared, showing the importance of balanced training for emotion classification from full-body movements.
- The effect of chunk duration and overlapping as well as the effect of the size of the input images were investigated. Results show that longer chunks improve the performance and image resizing applied to longer chunks contributes to the classification positively.

Motivated by the results showing that longer time observation intervals (such as 4-seconds) are boosting automatic emotion classification, one can speculate whether this is also valid for human perception of affective full-body continuous movements. The code of the proposed method is publicly available in: https://github.com/cbeyan/AffectiveBodyMovements.

Future work includes extending our dataset with performances of more dancers and with additional emotion classes. Adding more dancers would allow us to examine the generalization of the method across different participants, which is not covered in this study. To be able to deploy the proposed approach to interactive systems (e.g., social robots), data collected from RGB-D cameras will be used instead of MoCap systems. Last but not least, this work is a preliminary stage of a larger research project in which we aim to show that the proposed method trained on one domain (e.g., dance), can be adapted to recognize emotions elicited in other setups (e.g., daily actions).

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824160, FET PROACTIVE EnTimeMent (https://entimement.dibris.unige.it). We warmly thank Catherine Pelachaud (CNRS-ISIR, Sorbonne University, France) for providing the Emilya dataset and Andreas Aristidou (University of Cyprus, Cyprus) & Yiorgos Chrysanthou (University of Cyprus, Cyprus) for providing the Dance Motion Capture Database.

References

[1] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, no. 2, pp. B51 – B61, 2001.

[2] E. G. Krumhuber, A. Kappas, and A. S. R. Manstead, "Effects of dynamic aspects of facial expressions: A review," *Emotion Review*, vol. 5, no. 1, pp. 41–46, 2013.

[3] N. Dael, M. Goudbeek, and K. R. Scherer, "Perceived gesture dynamics in nonverbal expression of emotion," *Perception*, vol. 42, no. 6, pp. 642–657, 2013, pMID: 24422246.

[4] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," in *CVPR WORKSHOPS*, 2011, pp. 7–12.

[5] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[6] A. Camurri, G. Volpe, S. Piana, M. Mancini, R. Niewiadomski, N. Ferrari, and C. Canepa, "The dancer in the eye: Towards a multi-layered computational framework of qualities in movement," in *Proc. of the 3rd Int. Symposium on Movement and Computing*, ser. MOCO '16. New York, NY, USA: ACM, 2016.

[7] R. Laban and F. C. Lawrence, *Effort*. Macdonald & Evans, 1947.

[8] M. J. Vaessen, E. Abassi, M. Mancini, A. Camurri, and B. de Gelder, "Computational Feature Analysis of Body Movements Reveals Hierarchical Brain Organization," *Cerebral Cortex*, vol. 29, no. 8, pp. 3551–3560, 10 2018.

[9] P. H. Waxer, "Nonverbal cues for anxiety: An examination of emotional leakage," *Journal of Abnormal Psychology*, no. 86, 1977.

[10] R. Niewiadomski, M. Mancini, S. Piana, P. Alborno, G. Volpe, and A. Camurri, "Low-intrusive recognition of expressive movement qualities," in *ACM ICMI*, ser. ICMI 2017. ACM, 2017, pp. 230–237.

[11] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *Proc. of IJCNN*, July 2016, pp. 381–388.

[12] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Asian Conf. on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 579–583.

[13] J. Bütepage, M. J. Black, D. Kragic, and H. Kjellström, "Deep representation learning for human motion prediction and classification," *CoRR*, vol. abs/1702.07486, 2017.

[14] J. Wan, C. Lin, L. Wen, Y. Li, Q. Miao, S. Escalera, G. Anbarjafari, I. Guyon, G. Guo, and S. Z. Li, "Chalearn looking at people: IsoGD and ConGD large-scale RGB-D gesture recognition," *IEEE Trans. on Cybernetics*, pp. 1–12, 2020.

[15] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *IEEE Sensors Journal*, vol. 79, p. 30509–30555, 2020.

[16] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recogn. Lett.*, vol. 34, no. 15, p. 1995–2006, Nov. 2013.

[17] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *In: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer, 2013, pp. 149–187.

[18] I. H. López-Nava and A. Muñoz-Meléndez, "Wearable inertial sensors for human motion analysis: A review," *IEEE Sensors Journal*, vol. 16, no. 22, pp. 7821–7834, 2016.

[19] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of Nonverbal Behavior*, vol. 28, pp. 117–139, 01 2004.

[20] J. L. Tracy and R. W. Robins, "Show your pride: Evidence for a discrete emotion expression," *Psychological Science*, vol. 15, no. 3, pp. 194–197, 2004.

[21] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. on Affective Computing*, vol. 4, no. 1, pp. 15–33, Jan 2013.

[22] M. Karg, A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Trans. on Affective Computing*, vol. 4, no. 4, pp. 341–359, 2013.

[23] C. Corneanu, F. Noroozi, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. on Affective Computing*, no. 01, pp. 1–1, 2018.

[24] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2007, pp. 71–82.

[25] N. Fourati, C. Pelachaud, and P. Darmon, "Contribution of temporal and multi-level body cues to emotion classification," in *Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 116–122.

[26] G. Cimen, H. Ilhan, T. Capin, and H. Gurcay, "Classification of human motion based on affective state descriptors," *Computer Animation and Virtual Worlds*, vol. 24, no. 3-4, pp. 355–363, 2013.

[27] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *ACM Trans. on Interactive Intelligent Systems*, vol. 6, no. 1, pp. 6:1–6:31, Mar. 2016.

[28] N. Dael, M. Mortillaro, and K. Scherer, "The body action and posture coding system (BAP): Development and reliability," *Journal of Nonverbal Behavior*, vol. 36, pp. 97–121, 2012.

[29] M. Daoudi, S. Berretti, P. Pala, Y. Delevoye-Turrell, and A. D. Bimbo, "Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices," in *Image Analysis and Processing - ICIAP, Int. Conf., Part I*, ser. Lecture Notes in Computer Science, S. Battiato, G. Gallo, R. Schettini, and F. Stanco, Eds., vol. 10484. Springer, 2017, pp. 550–560.

[30] A. Kacem, M. Daoudi, B. B. Amor, S. Berretti, and J. C. Á. Paiva, "A novel geometric framework on gram matrix trajectories for human behavior understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 1–14, 2020.

[31] M. R. Loghmani, S. Rovetta, and G. Venture, "Emotional intelligence in robots: Recognizing human emotions from daily-life gestures," in *IEEE Conf. Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1677–1684.

[32] A. Crenn, A. Meyer, H. Konik, R. A. Khan, and S. Bouakaz, "Generic body expression recognition based on synthesis of realistic neutral motion," *IEEE Access*, vol. 8, pp. 207 758–207 767, 2020.

[33] W. Yun, D. Lee, C. Park, J. Kim, and J. Kim, "Automatic recognition of children engagement from facial video using convolutional neural networks," *IEEE Trans. on Affective Computing*, pp. 1–1, 2018.

[34] D. A. Al Chanti and A. Caplier, "Deep learning for spatio-temporal modeling of dynamic spontaneous emotions," *IEEE Trans. on Affective Computing*, pp. 1–1, 2018.

[35] C. Lu, W. Zheng, C. Li, C. Tang, S. Liu, S. Yan, and Y. Zong, "Multiple spatio-temporal feature learning for video-based emotion recognition in the wild," in *ACM ICMI*. New York, NY, USA: ACM, 2018, p. 646–652.

[36] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli, "Efficient spatial temporal convolutional features for audiovisual continuous affect recognition," in *Int. on Audio/Visual Emotion Challenge and Workshop*. New York, NY, USA: ACM, 2019, p. 19–26.

[37] S. Karumuri, R. Niewiadomski, G. Volpe, and A. Camurri, "From motions to emotions: Classification of affect from dance movements using deep learning," in *Extended Abstracts of the 2019 CHI Conf. on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2019, p. 1–6.

[38] B. de Gelder and J. V. den Stock, "The bodily expressive action stimulus test (BEAST). construction and validation of a stimulus basis for measuring perception of whole body expression of emotions," *Frontiers in Psychology*, no. 2, 2011.

[39] R. P. D. Silva and N. Bianchi-Berthouze, "Modeling human affective postures: an information theoretic characterization of posture features," *Computer Animation and Virtual Worlds*, vol. 15, no. 3-4, pp. 269–276, 2004.

[40] C. L. Roether, L. Omlor, A. Christensen, and M. A. Giese, "Critical features for the perception of emotion from gait," *Journal of Vision*, vol. 9, no. 6, p. 15, 2009.

[41] T. Olugbade, J. Newbold, R. Johnson, E. Volta, P. Alborno, R. Niewiadomski, M. Dillon, G. Volpe, and N. Bianchi-Berthouze, "Automatic detection of reflective thinking in mathematical problem solving based on unconstrained bodily exploration," *IEEE Trans. on Affective Computing*, 2020.

[42] C. Beyan, M. Shahid, and V. Murino, "RealVAD: a real-world dataset and a method for voice activity detection by body motion analysis," *IEEE Trans. on Multimedia*, pp. 1–1, 2020.

[43] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognition*, vol. 48, no. 5, pp. 1653 – 1672, 2015.

[44] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[45] T. Bänziger, M. Mortillaro, and K. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception," *Emotion (Washington, D.C.)*, vol. 12, pp. 1161–79, 11 2011.

[46] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[47] T. Sapinski, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, and G. Anbarjafari, "Multimodal database of emotional speech, video and gestures," in *Proc. of ICPR Workshops*, 2019.

[48] DMCD, "Dance Motion Capture Database: http://dancedb.eu/," 2021.

[49] N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 90–101, 2018.

[50] A. Aristidou, P. Charalambous, and Y. Chrysanthou, "Emotion analysis and classification: Understanding the performers' emotions using the lma entities," *Comput. Graph. Forum*, vol. 34, no. 6, p. 262–276, Sep. 2015.

[51] N. Fourati, "Classification and characterization of emotional body expression in daily actions. (classification et caractérisation de l'expression corporelle des emotions dans des actions quotidiennes)," Ph.D. dissertation, Télécom ParisTech, France, 2015. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01282785

Cigdem Beyan received her Ph.D. degree in Informatics (Computer vision and Machine Learning) from the University of Edinburgh, U.K., in 2015. She is currently an Assistant Professor in University of Trento. She has co-authored over 45 papers published in peer-reviewed journals and international conferences. Among her main research interest, there are social signal processing and multimodal data analysis. She is a reviewer of several journals including various IEEE Transactions, and IEEE/ACM conferences. She is in the Editorial Board of ICES Journal of Marine Science covering area of applications of computer vision and machine learning and was a Guest Co-Editor in Frontiers in Robotics and AI.

Sukumar Kurumuri received his Btech degree in electronics from Manipal Institute of Technology and later completed a master's degree in robotics engineering through the EMARO+ program, spending his first year at Warsaw University of Technology and second year at University of Genoa. He presently works at Miko, an Indian robotics company. His current research interests include multimodal emotion recognition and vision-based tracking.

Gualtiero Volpe received the M.Sc. degree in computer engineering in 1999 and the Ph.D. in electronic and computer engineering in 2003 from the University of Genoa, Italy. Since 2014, he is an Associate Professor at DIBRIS, University of Genoa. His research interests include intelligent and affective human-machine interaction, social signal processing, sound and music computing, modeling and real-time analysis of expressive content, and multimodal interactive systems.

Antonio Camurri received the Ph.D. degree in Computer Engineering and he is a full professor at DIBRIS (Polytechnic School, University of Genoa). His research interests combine scientific research in ICT with artistic and humanistic research and includes nonverbal multimodal interactive systems; computational models of nonverbal full-body expressive gesture, emotion, and social signals; interactive multimodal systems for performing arts, active experience of cultural content, wellness, therapy and rehabilitation. He is the scientific director of Casa Paganini - InfoMus Research Centre. He is the co-author of over 150 scientific publications. He is/was the coordinator of EU-funded projects in FP5 (IST MEGA), FP7 (ICT SAME, ICT FET SIEMPRE) and Horizon 2020 (DANCE, EnTimeMent), has been PI in about 20 EU-funded projects and has been in contracts with industry and cultural institutions. He is also the co-director of the Joint Research Laboratory ARIEL with Gaslini Children Hospital.

Radoslaw Niewiadomski received the Ph.D. degree in Computer Science from the University of Perugia, Italy. He is currently an Assistant Professor in University of Trento. His research interests include emotion recognition, nonverbal behavior synthesis and multimodal interaction. He has been involved in several EU research projects, e.g., FP6 CALLAS, FP7 ILHAIRE and H2020 DANCE and published over 75 peer-reviewed conference and journal papers.