From Motions to Emotions: Classification of Affect from Dance Movements using Deep Learning

Sukumar Karumuri
University of Genoa
Genoa, Italy
kai.sukumar@gmail.com

Radoslaw Niewiadomski
University of Genoa
Genoa, Italy
radoslaw.niewiadomski@dibris.unige.it

Gualtiero Volpe
University of Genoa
Genoa, Italy
gualtiero.volpe@unige.it

Antonio Camurri
University of Genoa
Genoa, Italy
antonio.camurri@unige.it

ABSTRACT

This work investigates classification of emotions from MoCap full-body data by using Convolutional Neural Networks (CNN). Rather than addressing regular day to day activities, we focus on a more complex type of full-body movement - dance. For this purpose, a new dataset was created which contains short excerpts of the performances of professional dancers who interpreted four emotional states: anger, happiness, sadness, and insecurity. Fourteen minutes of motion capture data are used to explore different CNN architectures and data representations. The results of the four-class classification task are up to 0.79 (F1 score) on test data of other performances by the same dancers. Hence, through deep learning, this paper proposes a novel and effective method of emotion classification which can be exploited in affective interfaces.

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312910

KEYWORDS

Emotion Recognition, Convolutional Neural Network, Motion Capture, Dance

INTRODUCTION

Most existing works on emotion recognition from full-body movements (e.g., [2, 3]) use procedural approaches which involve the creation of hand-crafted algorithms for the computation of high-level features and machine learning algorithms such as a Support Vector Machine for the classification. In this paper, by adopting a deep learning approach, and in particular a Convolutional Neural Network (CNN), we depend on the network to learn features through training on the dataset. Inspired by the works of Wallbott [12] and Pollick [10], we focus on the dynamic aspects of the movement and we do not aim to recognize and classify particular gestures (e.g., hand raising) which may accompany the emotional states. While CNN is popular due to its applications on classification of static images (e.g., [8]), it is also an appropriate technique to process temporal information and hence we try to make the network learn features which are meaningful for emotion classification. Compared to other popular deep learning methods such as Recurrent Neural Networks (RNN), the CNN often requires less computation and can provide good classification results for a smaller data size.

In this paper, a motion capture system is used for data collection due to its high accuracy. The reason dance movements have been chosen is that they are characterized by high complexity and versatility, and involve a much larger workspace of movements compared to regular day-to-day activities. This work, however, is the first stage of a larger research project in which we aim to show that the CNN trained on a specific domain such as dance, can be adapted to recognize emotions in other domains (e.g., every day activities) through transfer learning.

DATASET

Four classes of emotional states were considered in this study: angry, happy, insecure, sad. Three of them are present in the set of so called *basic emotions*, i.e., emotions that are claimed to be universally recognizable across different cultures [6]. The fourth one, insecure, is not among basic emotions but is closely related to fear. Recently it was shown that images displaying bodily emotions of anger, happiness, sadness and fear were correctly categorized in at least 85% of the cases [4].

Two professional dancers participated for the creation of the dataset. The dancers were asked to portray an emotion in a dance sequence of their choice. A team of three experts in affective computing cut down recordings into episodes of variable sizes. The parts with no clear emotion were discarded and only those episodes, which displayed one emotion (agreed upon by all the experts) were selected. The final dataset is depicted in Table 1. The Qualisys motion capture system was used for the creation

PREVIOUS WORK

A few works exist on classification of different physical activities and/or emotions from full-body data using deep learning algorithms:

(1) In [7], the authors apply a CNN for human activity recognition from the data collected using three accelerometers and two gyroscopes.
(2) In [5], CNN architecture is used for action recognition from motion capture data.
(3) In [1], the authors use a deep Encoder-Decoder architecture for both classification and prediction of various actions.
(4) In [9], a three layered RNN was used to perform the classification of emotions from motion capture data of daily activities.

Table 1: Total segment duration for each emotion

Emotion	No. of segments	Total duration
Angry	14	142s
Happy	15	199s
Insecure	17	248s
Sad	9	244s

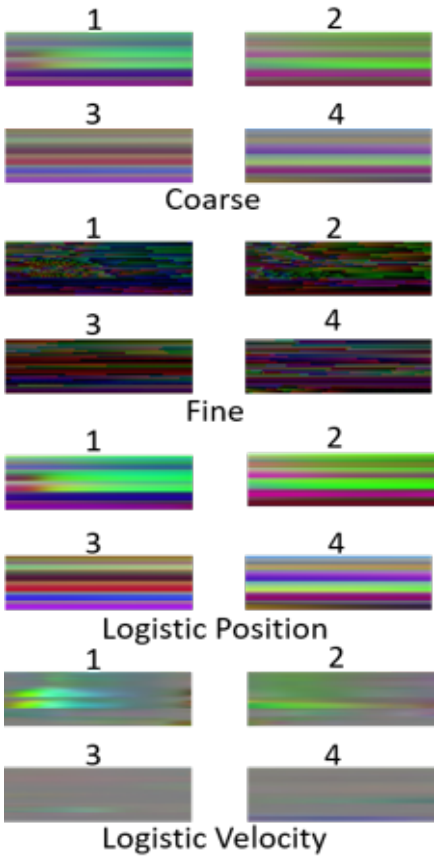

Figure 1: Visualization of images created using different formats: 1) angry, 2) happy, 3) insecure and 4) sad

of the dataset at the frame rate of 100 fps. 30 markers were used. Missing values were interpolated using either a polynomial or linear interpolation.

DATA REPRESENTATION

The 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). An 8 bit RGB image format was used to represent the data based on the method proposed in [5]. The five sets of marker positions constituted the y-axis while the consecutive frames of the sequence were represented on the x-axis. The X, Y and Z positions of the markers were grouped into the R, G and B frames respectively. This conversion required segmenting the episodes into shorter sections of fixed duration. For this purpose three window lengths were considered of six-, three- and one-second respectively. A body-centered relative normalization was applied: the value of marker that is situated on the xiphoid process at the first frame of each segment, was taken as the origin.

Four different formats were proposed to fit information in an 8-bit image. In subsequent equations, R represents the new marker value, while Q represents the original value obtained from relative position extraction. Examples of generated images are given in Figure 1.

The coarse position format (Coarse). This method involves extracting the coarser movement and rounding up the finer ones, using the Equation 1.

$$R = \left\lceil \frac{Q}{10} \right\rceil + 127 \tag{1}$$

The fine position format (Fine). While some emotions are often expressed by very large movements (e.g., anger or happiness [12]) others are characterized by subtle and/or repetitive movements such as trembling (e.g., fear [11]). We introduce a format which focuses on the finer details of movement. The markers undergo the operation shown in Equation 2.

$$R = \left\lceil \left(\frac{Q}{10} \right)_{mod10} \cdot 10 \right\rceil + 127 \tag{2}$$

The logistic position format (LP). A third format uses a logistic function to map the positions into the -127 to +127 interval as shown in Equation 3:

$$R = \left\lceil \frac{255}{1 + e^{-0.0035(Q)}} \right\rceil \tag{3}$$

The logistic velocity format (LV). In this format, due to the small size of the dataset, we extract a low-level feature from the data. Next, we compute frame-by-frame velocity, apply Savitzky-Golay filter, apply logistic function (Equation 3) and shift values to the 8-bit range.

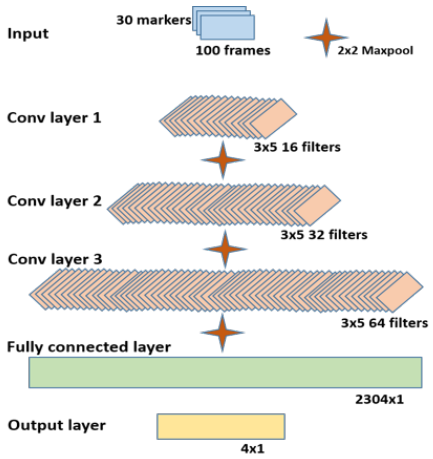

Figure 2: Single Input Architecture

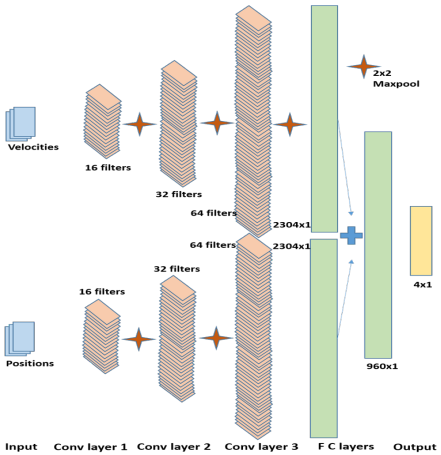

Figure 3: Multiple Input Architecture

In total, 1556 images were created for the data segment with length 1 second using a 0.5 second overlap. 1440 images were created for 3 second segment length with 2 second overlap, and 691 images were created for 6 second segment length with 3 second overlap.

CNN ARCHITECTURE

With a total duration of about 14 minutes, the dataset created a constraint on the size of the network and hence a rather simple network inspired by the LeNet-5 architecture [8] was designed. Three convolutional layers were found to be a good trade-off number. Two architectures were proposed:

(1) *Single Input Architecture* (SIA-CNN) uses one out of four data representation formats at once,
(2) *Multi Input Architecture* (MIA-CNN) allows multiple input formats to be applied simultaneously.

The SIA-CNN design for a one second segment is shown in Figure 2. A key feature is the shape of the convolutional filters which are extended along the time axis to form 3x5 rectangles. (It is important to note that the input image is also rectangular). The reason for this is that we expect the network to learn more features over time rather than between successive markers.

The input image is first convoluted using 16 filters of size 3x5. "SAME" type of padding is used to obtain 16 output matrices of same dimensions as input image. A 2x2 maxpool operation with a stride of 2 is done to reduce the dimensions. The new input is then passed onto the next layer via a ReLU function. This series of operations is repeated again in the next two convolutional layers but with increasing number of filters - 32 filters in the second layer and 64 in the third. After the third and final maxpool and ReLU layer, the image size is reduced but it contains 64 layers. The output is then flattened out and converted into a fully connected (FC) layer. A 4x1 FC layer is used as the output layer and a softmax function on this determines the final emotional class of the given image.

In the Multi Input Architecture (MIA-CNN), separate convolutional layers are used on the data representation formats before the weights are flattened out and added together in the first fully-connected layer (see Figure 3). The filter size and convolution operations are similar to the SIA-CNN.

Training and Validation

Mini-batch gradient descent is used to train the data. Adam optimization has been used as the optimization algorithm. Dropout regularization has been implemented on the FC layers and a form of early stopping was used to determine the number of epochs for training. The key hyper-parameter choices finalized using a grid search method are mentioned in Table 2. A hybrid cross validation was used to validate the training. In the inner loop, a 5-fold cross validation is applied. In the outer loop, the data is shuffled randomly and sent to the inner loop (Monte Carlo validation). This is done twice to provide a final of 10 (2x5) results.

Table 2: Hyper-parameter choices

Hyper-parameter	Value
Learning rate	0.0009
Minibatch Size	64
Number of epochs	Depends on format
Activation function	ReLU
Number of layers	3 Conv. + 2 FC

Table 3: Results for different segments

	LP		LV	
Length	Acc.	F1	Acc.	F1
1 sec	0.85	0.81	0.70	0.68
3 sec	0.73	0.71	0.65	0.62
6 sec	0.64	0.62	0.59	0.59

Table 4: Results in terms of F1 score for different data representation

Data Format	Validation	Test
Coarse format	0.79	0.67
Fine format	0.57	0.57
Logistic position	0.81	0.71
Logistic velocity	0.68	0.62

Table 5: Results in terms of F1 score for four different versions of MIA-CNN

Combination	Validation	Test
C1: Coarse + Fine	0.74	0.76
C2: LP + LV	0.87	0.79
C3: Coarse + Fine + LV	0.76	0.76
C4: Coarse + LP	0.79	0.70

ANALYSIS AND RESULTS

Two methods were used to evaluate the algorithms. In *Validation phase* the hybrid cross validation is used and the final results are computed as the average of the 10 readings.

Testing phase involves using additional 8 episodes, two for each emotion (a total of 105 1 second images, corresponding to around 6% of the original dataset) taken from other recordings of the same dancers which were not used at all in the dataset.

Sequence Length Evaluation

Among the considered sequence lengths, the shorter segments most of the time provided better results (Table 3) and hence only the results computed on 1 second segments are shown for comparison in the subsequent section. To check whether the size of the dataset lead to this result, the quantity of the 1 second segments was reduced to be the same as that of a 3 second segments dataset. Consequently, the accuracy for 1 second segments slightly decreased from 0.85 to 0.81 for the LP format but it was still much better than the accuracy of 3 second segments (0.73 as seen in Table 3). This showed that the reason behind better results was the way the network learned from the data.

Data Representation Evaluation

All four formats were compared using the SIA-CNN architecture on the validation and test sets (see Table 4). The best results were obtained for the LP format. Next, four different combinations of data formats were examined using MIA-CNN (Table 5). The combination of LP and LV (C2) provided the best results. This combination increased the F1 score of the test set to 0.79 which is 8 percent more than the maximum of the two individually (Table 4). This shows that different features were extracted and used by the network for LP and LV. The same conclusion can be said for the coarse and fine position formats (C1), while adding logistic velocity (C3) did not further increase the F1 score. Another interesting case was the combination of coarse position with LP. Individually these formats provided the best results but their combination did not improve these results.

In conclusion, the 1 second LP + LV format was the best combination, in terms of F1 score.

CONCLUSION AND FUTURE WORK

A novel approach for the classification of emotions from limited quantity of dance MoCap data using a CNN was presented. It is able to classify four emotions with an F1 score of 0.79. Two limitations should be mentioned 1) a rather small dataset size and 2) a test set coming from the same dancers. Thus, as a part of future works, we plan to apply the approach to other dance datasets. By extending the dataset to include more dancers, we hope to develop a robust emotion classification model which can be adapted to be used in other domains (e.g., to recognize emotions in every day full body activities)

The main contributions of this paper are:

(1) To the best knowledge of the authors, it is the first attempt to apply a CNN to classify emotions using full-body motion capture data.
(2) It is also one of the first works on the classification of emotions from dance movements using a deep learning approach.
(3) A novel data representation based on logistic position and velocity was proposed to get the most out of a limited dataset.
(4) A new dataset of dance sequences was created displaying four emotions which will be soon available for research purposes.

and applied to interactive systems such as social robots. Preliminary results (see Table 3) reveal that recognition accuracy differs for different length of input. Thus, we plan to extend the CNN architecture to be able to deal with the different temporal scales at once. Other research directions include the addition of Electromyography (EMG) sensor to better distinguish between anger and happiness (two emotions that were often confused by CNN). Last but not least, we will compare obtained results with human level of accuracy and other state of the art machine learning methods such as SVMs.

ACKNOWLEDGMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n. 824160 (EnTimeMent).

REFERENCES

[1] Judith Bütepage, Michael J. Black, Danica Kragic, and Hedvig Kjellström. 2017. Deep representation learning for human motion prediction and classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6158–6166.
[2] Ginevra Castellano, Santiago D. Villalba, and Antonio Camurri. 2007. Recognising Human Emotions from Body Movement and Gesture Dynamics. In *Affective Computing and Intelligent Interaction*, Ana C. R. Paiva, Rui Prada, and Rosalind W. Picard (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 71–82.
[3] Gokcen Cimen, Hacer Ilhan, Tolga Capin, and Hasmet Gurcay. 2013. Classification of human motion based on affective state descriptors. *Computer Animation and Virtual Worlds* 24, 3-4 (2013), 355–363. https://doi.org/10.1002/cav.1509
[4] Beatrice de Gelder and Jan Van den Stock. 2011. 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* 2 (2011). https://doi.org/10.3389/fpsyg.2011.00181
[5] Yong Du, Yun Fu, and Liang Wang. 2015. Skeleton based action recognition with convolutional neural network. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*. IEEE, 579–583.
[6] Paul Ekman and Wallace V. Friesen. 1975. *Unmasking the Face: A Guide to Recognizing Emotions From Facial Clues*. Englewood Cliffs, NJ: Prentice-Hall.
[7] Sojeong Ha and Seungjin Choi. 2016. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In *2016 International Joint Conference on Neural Networks (IJCNN)*. 381–388.
[8] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (Nov 1998), 2278–2324. https://doi.org/10.1109/5.726791
[9] Mohammad Reza Loghmani, Stefano Rovetta, and Gentiane Venture. 2017. Emotional intelligence in robots: Recognizing human emotions from daily-life gestures. In *Robotics and Automation (ICRA)*. 1677–1684.
[10] Frank E. Pollick. 2004. The Features People Use to Recognize Human Movement Style. In *Gesture-Based Communication in Human-Computer Interaction*, Antonio Camurri and Gualtiero Volpe (Eds.). Springer Berlin Heidelberg, 10–19.
[11] B. Schuller, M. Lang, and G. Rigoll. 2002. Multimodal emotion recognition in audiovisual communication. In *Proceedings. IEEE International Conference on Multimedia and Expo*, Vol. 1. 745–748 vol.1. https://doi.org/10.1109/ICME.2002.1035889
[12] Harald G Wallbott. 1998. Bodily expression of emotion. *European journal of social psychology* 28, 6 (1998), 879–896.