# Emotion Recognition Based on the MoCap Data of Actors Expressing Emotions

handed in
PRACTICAL COURSE

B.Sc Selin Kesler

Human-centered Assistive Robotics
Technical University of Munich

Univ.-Prof. Dr.-Ing. Dongheui Lee

Supervisor: Dr. Hyemin Ahn

I have fully read and acknowledge the above information and guidelines for the practical course.

**Matriculation Number:** 03694002

**Full Name (First Name, Last Name):** SELIN KESLER

**Date, Place:** 03.05.2021, MÜNCHEN

**Signature:**

TECHNISCHE UNIVERSITÄT MÜNCHEN

**Human-centered Assistive Robotics**

UNIV.-PROF. DR.-ING. DONGHEUI LEE

April 18, 2021

PRACTICAL COURSE
for
Student's name, Mat.-Nr.

## Emotion recognition based on the MoCap data of actors expressing emotions

Problem description:

Understanding human emotion is a crucial factor when intuitively communicating with others. Recognizing human emotion has been considered a lot in various research fields, including the human-robot-interaction [2, 1]. The goal of this course is to build a neural network based classifier, which can recognize various human emotions (i.e., sad, happy, surprise) based on the MoCap full body human pose data (`https://www.nature.com/articles/s41597-020-00635-7`), which is collected from actors expressing emotions. [3]. The trained classifier should be able to understand the emotional state of humans when a time-series of MoCap data has been given as an input.

Tasks:

- Preprocessing the dataset of "Kinematic dataset of actors expressing emotions" (`https://www.nature.com/articles/s41597-020-00635-7`).
- Train a neural network which can classify different emotional expressions (i.e., sad, happy, surprise).

Bibliography:

[1] Luefeng Chen, Mengtian Zhou, Wanjuan Su, Min Wu, Jinhua She, and Kaoru Hirota. Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. *Information Sciences*, 428:49–61, 2018.

[2] Nourhan Elfaramawy, Pablo Barros, German I Parisi, and Stefan Wermter. Emotion recognition from body expressions with a neural network architecture. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 143–149, 2017.

[3] Mingming Zhang, Lu Yu, Keye Zhang, Bixuan Du, Bin Zhan, Shaohua Chen, Xiuhao Jiang, Shuai Guo, Jiafeng Zhao, Yang Wang, et al. Kinematic dataset of actors expressing emotions. *Scientific data*, 7(1):1–8, 2020.

Supervisor: Dr. Hyemin Ahn

(D. Lee)
Univ.-Professor

## Abstract

Due to the growing popularity of Human Robot Interaction (HRI), not only in academic researches but also in the industry, emotion recognition algorithms gained importance and popularity. As different fields require different methods, there is not a single best solution for the task.

This project focuses on the general pattern recognition for emotions from body language with the kinematic dataset, which should work independently from the actions in the given scene. For this purpose Long Short Term Memory is implemented to capture temporal and spatial information with the help of its memory cells. The first method of using pure positional joint information through the frames is substituted with spatial and temporal information extraction. The use of magnitude of the change in joint coordinates and particular spatial information extracted from head and shoulder led the accuracy of the network to rose from %51.8 to %69.04.

# Contents

# Chapter 1

# Background

## 1.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are dynamic systems, which have an internal state at each time step by means of their circular connections between higher- and lower-layer neurons and optional self-feedback connections. Through these connections, RNNs can attain data from earlier events in current steps and therefore they can form a memory of the previous events. [SM19]
A simple vanilla RNN could be imagined as a neural network with a series of inputs over time, where the network has a single layer of neurons for each input. Each layer of neurons generates a hidden state in addition to their output, which enables them to transfer information from previous steps to the current state. During the training process, not only the weights in each neuron are assigned but also the parameters in hidden states are defined, which determines how much of the information from previous time steps should be carried with. Two of the most common gradient-based training methods for RNN, Back-Propagation Through Time (BPTT) [WZ95] and Real-Time Recurrent Learning (RTRL) [RF87], share a limitation problem of long term memory due to the magnitude of the error signal propagated back in time, which depends exponentially on the weights and has a tendency either to vanish or blow-up [GSC00]. Due to these limitations, standard RNNs can't connect more than five to ten time steps without losing information [GSC00].
To solve this long-time memory limitation because of the vanishing error problem, a new approach named LSTM (Long-Short Term Memory) [HS97a] was proposed in the late 1990s by German researchers Hochreiter and Schmidhuber, which can expand the memory up to 1000 time steps by using constant error carousels (CECs) in special units called memory cells [SM19]. CECs are recurrently self-connected linear units with a fixed-weight, which leads to constant, non-vanishing and non-exploding error flow in the memory cell, whose activation state is called cell state [HS97b]. CEC is protected from both forward and backward flowing error by closing input and output gates so that noise and irrelevant input can't access to the cell [GSC00].

# Chapter 2

# Introduction

Expression of emotions played always an important role in the social community as they could mean so much more than pure verbal communication [Pic97]. Emotions are not only expressed through facial expressions but also through eye movements, gestures, and most importantly through body language. The focus of this project lies in finding a general pattern of 6 basic emotions in the body language of humans for various scenarios.
For this purpose, a dataset containing joint coordinates and rotational information of actors expressing emotions in short clips has been selected, which consists of 1402 clips of actors performing various emotion scenarios [Zha20]. The selected emotions for classification are sad, happy, angry, surprised, disgust, fear, which are first categorized by Charles Darwin back in 1872 [New13] and are scientifically accepted in many kind of literature today [PW17]. A seventh state, neutral, was also used in the dataset to prevent the constant need of detecting an emotion.

## 2.1 Problem Statement

As emotion expression is more a subjective topic in comparison to the tasks with a certain problem definition that has clearly defined boundaries such as object detection, it brings along some additional problems and challenges with itself. The major challenges could be classified into two sub-groups :

- Different cultures tend to show different emotional expressions[EF71].

  While some cultures could use their hands and full body to express their feelings, others could do minimal or local movements, which would make emotion pattern recognition a more difficult and complex task with diverse levels.

- Body language for the same emotion could vary between different action scenarios.

  Although the spatial information for specific emotions could have common features in different scenarios, e.g. a sad person would have his shoulders down as he eats or walks while the posture of a happy person would be more straight and tend to have broad shoulders. Extraction of temporal information in these scenarios on the other hand would be more challenging, as a happy person could swing his arms more during walking while a sad person tends to not swing his arms widely. However, this temporal information of arm swings would not be helpful in an eating scenario, which makes it difficult to find a general pattern for all actions.

## 2.2 Related Work

As the importance of human-robot communication rose as a consequence of the popularity of machine learning and robotics in the previous years, the need for intelligent systems which are capable of understanding human psychology has shown an increase. The researchers approached this challenge from different perspectives while some of them decided to focus on facial expressions [KMK$^+$15], [TGK20], [SPG19] such as capturing the essentials of gestures and mimics while the others chose to target body language [EBPW17], [RG19], [SKPA19], [SCHD19]. A fusion of speech analysis with both of these methods to improve performance is also to be seen in many works [SLII20], [Kam19], [SKP$^+$18].

The movement of the body is considered to be a reflection of our feelings [BL82]. This variance in the body could be examined directly in specific actions or a general pattern of these gestures could be searched. As the implementations for action recognition algorithm using skeletal information [ZLX17], [SLX$^+$16] has become more demanding with the rising quality of kinematic datasets, the same approaches were adapted to emotion recognition tasks [SKPA19], [SLII21].

The study by Tanmay Randhavane *et. al.* focused on identifying emotions from walking [RBK$^+$19] with the help of LSTM networks, in which they proposed to divide the extracted features into two segments; posture features and movement features. Posture features consisted of the spatial information about angles and distances between specific joints while movement features focused on features such as speed, acceleration magnitude, and movement jerk for the temporal information extraction.

Appositely, Tomasz Sapinski *et. al.* present an idea for general pattern detection for emotions, for which they created a dataset with 7 emotional states similar to the one that is used for this project [Zha20]. They examine the joint positions of actors in different neural networks such as CNN, RNN, and RNN-LSTM, in which they got the best results with RNN-LSTM using upper and lower body position information together in comparison to the experiments with separate body parts or with orientation information. They have determined that happiness, sadness, and anger have a higher rate of recognition in comparison to other emotions while the neutral state causes confusion. [SKPA19]

# Chapter 3

# Technical Approach

## 3.1 Design of your solution

The goal of this project was to develop and integrate an interpreter for emotion recognition with kinematic data, which focuses on RNN models that are capable of transferring memory. For this purpose different RNN models such as LSTM and GRU had been implemented and optimized with different logical approaches for emotion detection. These methods could be divided into 2 subsections, where the first subsection includes more general video processing approaches such as full-video and clips processing, while the second subsection focuses on the extraction of the most relevant and important information from joints.

### 3.1.1 Preprocessing of the Dataset

Kinematic dataset [Zha20] consists of 1402 samples of actors expressing emotions in multiple scenes. Each video was recorded with 125 FPS, in which each frame contains the 3-dimensional coordinate information and 3 rotation angles of 72 joints. For simplicity, only the coordinates of the joints will be considered for the given task.

As each video had different lengths, fixing the input length of the RNN model became necessary. For this purpose average video length of 902 frames was divided to 12 so that the FPS of the video would approximately be reduced from 125 to 10 FPS and the average video length would be equal to 75 frames, which would be a more convenient input to RNN so that it would not lose memory information while processing the frames. The number of frames to be skipped is then calculated dynamically as the total video length is divided to 75 for each video sample, which would make the 75 chosen frames always equally distant from each other for the processed sample. 3 coordinates of 72 joints were then saved for each of the 75 frames. This preprocessing approach forms the baseline for further preprocessing ideas with different kinds of information extraction methods.

## 3.2 Implementation

### 3.2.1 Video Processing

2 kinds of approaches for modeling the LSTM input have been tried, the full video approach and the clips approach. For the full video approach, extracted 75 frames had been given to LSTM directly and the output of the last cell is given as an input to the fully connected layer, which then outputs the predictions of emotion ID. The final emotion ID is then chosen by taking the index of maximum prediction. (Figure 3.1)

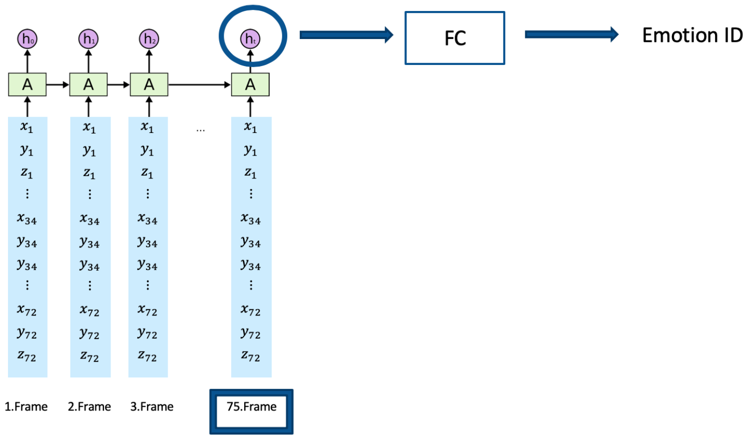

Figure 3.1: Full Video Method

For the clips approach, 75 frames were divided into 5 subgroups of 15 frames, which were then used as inputs for LSTM. Each of the 5 LSTM outputs its own emotion ID, from which the most wanted ID is selected as the final prediction of the given sample. (Figure 3.2)

### 3.2.2 Information Extraction from the Joints

Several information extraction methods have been tried to find the best results. First approach to be tried was using directly the 3-dimensional coordinates of 72 joints as suggested by T. Sapinski *et. al.* [SKPA19], with the goal of LSTM learning the spatio-temporal information. As 72 joints in 3 dimensions result in large input for LSTM, the number of the joints had been reduced by removing the z-coordinates to examine the effect of the depth information. The detailed joint information from both hands has been also reduced to a single hand joint, which culminated in a total of 21 joints in comparison to 72 joints. The impacts of hand gestures on emotion detection have been studied with these settings and examined further in chapter 4. Another method was separating the spatial and temporal information. The magnitude and speed of the actions could be good indicators [RBK$^{+}$19] for emotion

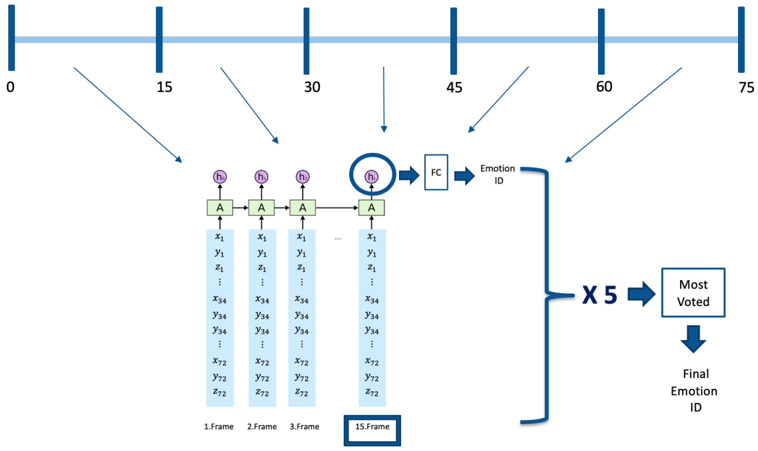

Figure 3.2: Clips Method

recognition, as for instance, people tend to make bigger moves while they are happy or make faster movements if they got surprised [MGC87]. In order to capture this information, the change of the coordinates between consecutive frames has been saved as magnitude, while for the speed information, the calculated magnitude was divided by the number of frames that have been jumped. The network is trained with magnitude and speed information separately but also together as the outputs of both networks had been combined for a single prediction.

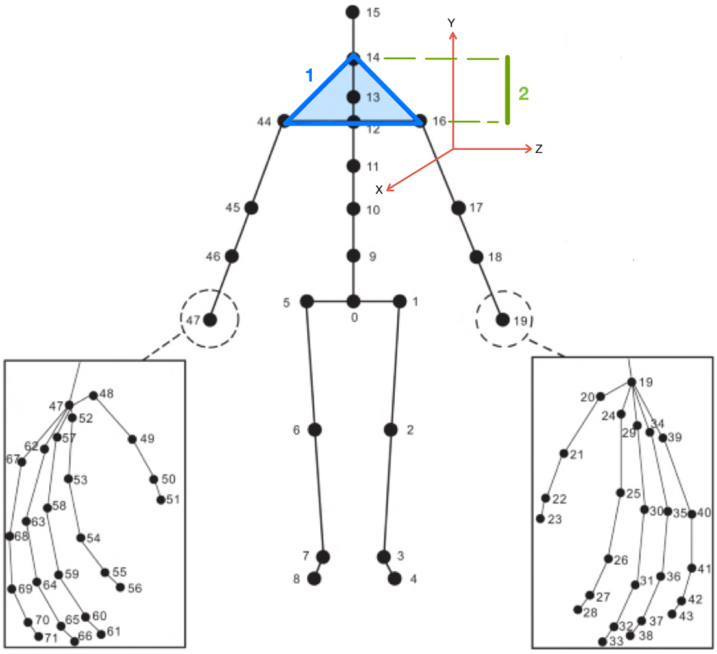

Figure 3.3: Skeleton Representation

For the extraction of spatial information, the focus has lied on the shoulder and head ratio. As the dataset consists of different kinds of scenarios and the goal of the project is to implement an algorithm capable of detecting emotions independently from the actions, it is not possible to examine the hand feet ratio or hand hip ratio, which could have given information about swinging of the arms by walking. The shoulder and head ratio could on the other hand reveal knowledge about the state of happiness or sadness due to the fact that people tend to hold up their heads high and shoulders broad when they are happy and drop their shoulders and bend their necks when they are feeling sad. In order to express these conditions mathematically, the area of the triangle between the shoulder points and head was calculated and the vertical distance between the head joint and shoulder joints was also used in each frame. The calculated area is shown in the Figure 3.3 with the number 1, while the vertical distance is represented with y-coordinate differences between joints and presented with the number 2. These extracted additional 3 spatial information was then added to the magnitude of the 72 joints, making the input length $72 + 3 = 75$ for each input.
Finally, in addition to LSTM [HS97a], GRU [CvMBB14] was also implemented and tested. The outcomes are further discussed in the Evaluation chapter 4.

# Chapter 4

# Evaluation and Discussion

In the following chapter, the impacts of various combinations of video processing and different feature extraction methods will be examined and discussed. The chapter starts with video processing methods and uses the best result in the experiments to follow such as hand joint effects, hyperparameter variation, and spatio-temporal information extraction.

## 4.1 Experimental Results

Experimental results are divided into subsections. First of all video processing with clips approach and full video approach is examined with different optimizer and hyperparameter settings. The extracted features are fixed to the pure 3D positional information of 72 joints in order to focus only on the video processing methods.

Table 4.1: Clips Method and Full Video Method

| Approach | BS | Optimizer | lr | lr Decay | Hidden Dimension | Accuracy |
|---|---|---|---|---|---|---|
| Full | 32 | Adam | 0.1 | NaN | 100 | %32.1 |
| Full | 32 | SGD | 0.1 | 0.5 / 50 | 100 | %36.7 |
| Clip | 16 | SGD | 0.1 | NaN | 100 | %29.6 |
| Clip | 16 | SGD | 0.1 | 0.5 / 50 | 100 | **%51.8** |
| Clip | 16 | Adam | 0.1 | 0.5 / 50 | 256 | %44 |

The results in the Table 4.1 indicate that giving clips rather than full video as an input has a positive effect on accuracy. Since in longer inputs the memory information could get lost or lose importance while the most voted emotion ID form clip accuracy might be more robust. Best results were obtained with the optimizer SGD with a learning rate of 0.1 and a learning decay of 0.5 in 50 epochs. In further experiments clip approach has been used for different feature settings.

Table 4.2 points out the significance of the depth information and the importance of hand gestures for emotion detection. Since hand gestures play an important role in

Table 4.2: Hand Joints and Z-Coordinates Importance

| Approach | BS | Optimizer | LR | LR Decay | Hand Info | Z-Coord | HD | Accuracy |
|---|---|---|---|---|---|---|---|---|
| Clip | 16 | SGD | 0.1 | 0.5 / 50 | Yes | Yes | 100 | **%51.8** |
| Clip | 16 | SGD | 0.1 | 0.5 / 50 | Yes | No | 100 | %37.3 |
| Clip | 16 | SGD | 0.1 | 0.5 / 50 | No | Yes | 100 | %31.6 |

emotion representations in many cultures, the remarkable decrease in the accuracy is feasible.

Effects of different hyperparameter settings are represented in the Table 4.3. Adding Gaussian noise with mean = 0 and variance = 0.01 to samples resulted in an accuracy decrease, while it showed positive effects on reducing the overfitting, as the gap between test and train losses came closer.

Table 4.3: Hyperparameter Settings

| Approach | BS | Optimizer | LR | LR Decay | Bi-D | LD | Detach | Noise | Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| Clip | 16 | SGD | 0.1 | 0.5 / 50 | Yes | 2 | Yes | No | %50.17 |
| Clip | 16 | SGD | 0.1 | 0.5 / 50 | No | 2 | No | No | %49.93 |
| Clip | 16 | SGD | 0.1 | 0.5 / 50 | No | 3 | Yes | No | **%51.69** |
| Clip | 16 | SGD | 0.1 | 0.5 / 75 | No | 3 | Yes | No | %42.26 |
| Full | 32 | SGD | 0.1 | 0.5 / 50 | No | 2 | No | No | %33.65 |
| Clip | 16 | Adam | 0.1 | NaN | No | 2 | Yes | (0.01) | %45.96 |
| Clip | 16 | SGD | 0.1 | 0.5 / 50 | No | 2 | Yes | (0.01) | %46.11 |

For investigating the impacts of different feature extraction methods, the hidden dimension is fixed to 100, while the batch size of clips is set to 16. Table 4.4 focuses on the effects of various spatial and temporal information extraction approaches. Best result from the previous approach with only skeletal position information (%51.8) is used to compare with new results. For temporal information, the magnitude and speed of the joints have been examined separately and together, while the best results were then obtained with pure magnitude information. In order to add spatial information to the temporal information, the shoulder-head triangle area and distances described in the subsection 3.2.2 were added to the input of magnitude, which led to the best result with a relatively small improvement in comparison to the pure magnitude information.

For the fusion of the spatial and temporal information, skeletal positional information was added to the magnitude input, making the input for each cell twice as long. This approach reached an accuracy of %63.7 as represented in the Table 4.4 with the name Magnitude + Pos (1). On the other hand Magnitude + Pos (2) trains 2 different LSTMs for magnitude and position separately and combines the outputs for emotion identification. This approach achieved a lower accuracy, as the results from position LSTM could be causing limitations on the predictions from magnitude LSTM.

Table 4.4: Spatial and Temporal Information

| Approach | BS | Optimizer | LR | LR Decay | HD | Noise | Accuracy |
|---|---|---|---|---|---|---|---|
| Position | 16 | SGD | 0.1 | 0.5 / 50 | 100 | No | %51.8 |
| Magnitude | 32 | Adam | 0.1 | NaN | 100 | No | %57.25 |
| Magnitude | 16 | SGD | 0.1 | 0.5 / 75 | 100 | No | %68.53 |
| Magnitude | 16 | SGD | 0.1 | 0.5 / 75 | 100 | Yes (0.01) | %66.26 |
| Speed | 16 | SGD | 0.1 | 0.5 / 50 | 100 | No | %61.86 |
| Magnitude + Speed | 16 | SGD | 0.1 | 0.5 / 50 | 100 | No | %65.12 |
| Magnitude + Trio | 16 | SGD | 0.1 | 0.5 / 75 | 100 | No | **%69.04** |
| Magnitude + Trio | 16 | SGD | 0.1 | 0.5 / 100 | 100 | No | %66.55 |
| Magnitude + Trio | 16 | SGD | 0.1 | 0.1 / 75 | 100 | No | %61.21 |
| Magnitude + Pos (1) | 16 | SGD | 0.1 | 0.5 / 75 | 100 | No | %63.7 |
| Magnitude + Pos (2) | 16 | SGD | 0.1 | 0.5 / 75 | 100 | No | %47.69 |

Last of all, performances of different RNN models, LSTM and GRU, had been examined. It was decided to use GRU, which also has long- and short term memory dependencies, that are not transmitted in 2 separate ways as in LSTM but in a combined direction thus, so the network is considered even lighter than LSTM. Due to the lighter structure of the GRU, an accuracy of %60.5 has been observed while with the same hyperparameter settings an accuracy of %69.04 was observed with LSTM.

# Chapter 5

# Conclusion

From various proposed approaches for examining the emotional stand of humans, a skeletal data-based emotion recognition algorithm with LSTM network was chosen to be implemented and developed for this project as they work well with the computational cost limitations and are suitable for real-time applications if necessary. Skeletal data brings the necessary human information in a light representation in comparison to RGB image methods. The network is trained with a Kinematic dataset.

During the implementation of this project, many experiments for different methods were used to optimize the performance of the emotion classification network. The first approach was using the pure position information of the 72 joint coordinates to LSTM as an input sequence, from which a maximum accuracy of %51.8 was achieved. A remarkable performance boost is achieved, as temporal information extraction was introduced which used the magnitude of joint coordinate differences between the frames and achieved an accuracy of %68.53. Lastly, spatial information was combined with the temporal information and ended up with an accuracy of %69.04.

It should be noted that the Kinematic dataset for emotion detection is not specified for actions and contains various scenarios, which increases the difficulty level for emotion prediction as actions are strongly correlated with body language.

## 5.1 Future Work

Since emotions are a wide and stratified topic even for humans, implementation of a sufficient emotion detection algorithm considering varying features is difficult. In order to capture as much as information from the available data, a fusion of different neural networks could be implemented, such as using 2D CNNs on RGB images or 3D CNNs directly on videos and merging their results with the current skeletal data network. In addition to the positional information of the joints, the effect of rotation angles of the joints could be further investigated.

Adding new dimensions to the project is also another possibility. Human body could be divided and examined in smaller parts such as facial expressions or hand gestures. A combination of these specific parts with the skeletal information could lead to better results. A study for searching the most essential and critical body parts for emotion recognition could be used to weigh the parts of the body with their importance score.

Another further study possibility could be classifying emotions with respect to actions. A separate network is then needed to recognize actions, from which the output should be used as input to the emotion detection system or even could direct the task to the specific emotion recognition network which is specifically trained for the given action.

# List of Figures

# Acronyms and Notations

**RNN** Recurrent Neural Network

**LSTM** Long Short-Term Memory

**GRU** Gated Recurrent Units

**SGD** Stochastic Gradient Descent

**Adam** Adaptive Moment Estimation

**LR** Learning Rate

**LR Decay** Learning Rate Decay

**BS** Batch Size

**HD** Hidden Dimension

**Bi-D** Bi-Directional

**Hand Info** Hand Information

**Z-Coord** Z Coordinates

**RGB** Red Green Blue

**HRI** Human Robot Interaction

**ID** Identification

# Bibliography

[BL82] I. Bartenieff and D. Lewis. *Body Movement: Coping with the Environment*. Gordon and Breach Science Pub., 1982. URL: `https://books.google.de/books?id=Kvb9kQAACAAJ`.

[CvMBB14] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014. URL: `http://arxiv.org/abs/1409.1259`, `arXiv:1409.1259`.

[EBPW17] Nourhan Elfaramawy, Pablo Barros, German Parisi, and Stefan Wermter. Emotion recognition from body expressions with a neural network architecture. pages 143–149, 10 2017. `doi:10.1145/3125739.3125772`.

[EF71] P. Ekman and W. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17 2:124–9, 1971.

[GSC00] Felix A. Gers, Jurgen Schmidhuber, and Fred A. Cummins. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12:2451–2471, 2000.

[HS97a] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. `doi:10.1162/neco.1997.9.8.1735`.

[HS97b] Sepp Hochreiter and Jurgen Schmidhuber. Lstm can solve hard long time lag problems. In *Advances in Neural Information Processing Systems 9*, pages 473–479. MIT Press, 1997.

[Kam19] D. Kaminska. Emotional speech recognition based on the committee of classifiers. *Entropy*, 21, 2019.

[KMK$^+$15] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. 11 2015. `doi:10.1145/2818346.2830596`.

[MGC87] Joann Montepare, Sabra Goldstein, and Annmarie Clausen. The identification of emotions from gait information. *Journal of Nonverbal Behavior - J NONVERBAL BEHAV*, 11:33–42, 03 1987. `doi:10.1007/BF00999605`.

[New13] Catherine Newmark. *Charles Darwin: The Expression of the Emotions in Man and Animals*, pages 85–88. Springer Fachmedien Wiesbaden, Wiesbaden, 2013. `doi:10.1007/978-3-531-93439-6_11`.

[Pic97] Rosalind W Picard. Affective computing, 1997.

[PW17] Magda Piorkowska and Monika Wrobel. *Basic Emotions*. 07 2017. `doi:10.1007/978-3-319-28099-8_495-1`.

[RBK$^+$19] Tanmay Randhavane, Aniket Bera, Kyra Kapsaskis, Uttaran Bhattacharya, Kurt Gray, and D. Manocha. Identifying emotions from walking using affective and deep features. *ArXiv*, abs/1906.11884, 2019.

[RF87] AJ Robinson and Frank Fallside. *The utility driven dynamic error propagation network*. University of Cambridge Department of Engineering Cambridge, MA, 1987.

[RG19] Santhoshkumar Rajaram and M. Geetha. Deep learning approach for emotion recognition from human body movements with feedforward deep convolution neural networks. *Procedia Computer Science*, 152:158–165, 01 2019. `doi:10.1016/j.procs.2019.05.038`.

[SCHD19] Zhijuan Shen, Jun Cheng, Xiping Hu, and Qian Dong. Emotion recognition based on multi-view body gestures. *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3317–3321, 2019.

[SKP$^+$18] T. Sapinski, D. Kaminska, A. Pelikant, C. Ozcinar, Egils Avots, and G. Anbarjafari. Multimodal database of emotional speech, video and gestures. In *CVAUI/IWCF/MIPPSNA@ICPR*, 2018.

[SKPA19] T. Sapinski, D. Kaminska, A. Pelikant, and G. Anbarjafari. Emotion recognition from skeletal movements. *Entropy*, 21, 2019.

[SLII20] Jing Shi, Chaoran Liu, C. Ishi, and H. Ishiguro. 3d skeletal movement enhanced emotion recognition network. *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1060–1066, 2020.

[SLII21] Jing Shi, Chaoran Liu, C. Ishi, and H. Ishiguro. Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network. *Sensors (Basel, Switzerland)*, 21, 2021.

[SLX$^+$16] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *CoRR*, abs/1611.06067, 2016. URL: `http://arxiv.org/abs/1611.06067`, `arXiv:1611.06067`.

[SM19] Ralf C. Staudemeyer and Eric Rothstein Morris. Understanding lstm – a tutorial into long short-term memory recurrent neural networks, 2019. `arXiv:1909.09586`.

[SPG19] Akash Saravanan, Gurudutt Perichetla, and K. S. Gayathri. Facial emotion recognition using convolutional neural networks. *CoRR*, abs/1910.05602, 2019. URL: `http://arxiv.org/abs/1910.05602`, `arXiv:1910.05602`.

[TGK20] Thomas Teixeira, Eric Granger, and Alessandro Lameiras Koerich. Continuous emotion recognition with spatiotemporal convolutional neural networks. *CoRR*, abs/2011.09280, 2020. URL: `https://arxiv.org/abs/2011.09280`, `arXiv:2011.09280`.

[WZ95] Ronald J. Williams and David Zipser. Gradient-based learning algorithms for recurrent networks and their computational complexity, 1995.

[Zha20] Yu L. Zhang K. et al. Zhang, M. Kinematic dataset of actors expressing emotions. *PhysioNet*, 17 2, 2020.

[ZLX17] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. 03 2017. `doi:10.1109/WACV.2017.24`.

# License

# Todo list