pHealth 2022
B. Blobel et al. (Eds.)

doi:10.3233/SHTI220967

Assessing Human Mobility by Constructing a Skeletal Database and Augmenting it Using a Generative Adversarial Network (GAN) Simulator

Yoram SEGAL [a,1], Ofer HADAR [a] and Lenka LHOTSKA [b]
[a] *Department of Systems and Communication Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel*
[b] *Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Prague, Czech Republic*

Abstract. This paper presents a neural network simulator based on anonymized patient motions that measures, categorizes, and infers human gestures based on a library of anonymized patient motions. There is a need for a sufficient training set for deep learning applications (DL). Our proposal is to extend a database that includes a limited number of videos of human physiotherapy activities with synthetic data. As a result of our posture generator, we are able to generate skeletal vectors that depict human movement. A human skeletal model is generated by using OpenPose (OP) from multiple-person videos and photographs. In every video frame, OP represents each human skeletal position as a vector in Euclidean space. The GAN is used to generate new samples and control the parameters of the motion. The joints in our skeletal model have been restructured to emphasize their linkages using depth-first search (DFS), a method for searching tree structures. Additionally, this work explores solutions to common problems associated with the acquisition of human gesture data, such as synchronizing activities and linking them to time and space. A new simulator is proposed that generates a sequence of virtual coordinated human movements based upon a script.

Keywords. OpenPose, Rehabilitation, Generative Adversarial Network (GAN), Siamese twins Neural Network, Simulator, Human body movements

Introduction

Remote medicine utilizes human gestures to conduct real-time medical treatments (1) (2). For example, The COVID-19 pandemic demonstrated the importance of remote diagnosis and treatment. In the modern age, it is now possible to utilize a camera video stream to collect, analyze, and interpret human emotions in a remotely located 3D environment by using artificial neural networks (3). Vocabulary Lexicon's purpose with this Research work is to present an indexing system for our predefine body movements that will enable us to recognize and describe physical actions using machine learning algorithms. As an outcome of this research, developers will transform posture into text

[1] Corresponding Author. Yoram Segal, Department of Systems and Communication Engineering, Ben-Gurion University, Beer-Sheva, Israel; E-mail: yoramse@post.bgu.ac.il.

and vice versa. Our objective is to characterize human motion by using neural network architectures such as Autoencoder (4), Siamese twins (5), and DWT-NN (6) in conjunction with OpenPose (7) and Mediapipe (8). Remote therapy may be used when many patients recuperate after hip, knee, elbow, or shoulder surgery (9) (10). A variety of non-contact medical treatments might be developed by utilizing a family of neural network designs resulting from this research.

1. Literature Review

This research work dissertation proposes a solution to enrich and enhance skeletal data veracity, by providing accurate and specific data tailored to research requirements using the GAN deep-learning method (11). In the articles (12) (13) (14), some databases contain video clips of human movements divided into a variety of classes. To begin with, they process the data through the OpenPose software, translating the video frames into skeletal pose sequences, which are then analyzed. A three-dimensional matrix represents each skeletal pose. To preserve the relationship between the skeleton joints, the authors reordered every pose as part of Deep First Search (DFS). Our movement generator is based on skeletal data that provide spatial and temporal information. Several studies have investigated the issue of recognizing human movement using skeleton-based neural networks (CNNs) (12) (15) (16). Therefore, Deep Convolutional Generative Adversarial Networks (DC-GANs) use CNN layers as their generator and discriminator [8]. It is proposed in (12) (14) (16) (17) to use an image format (TSSI - Tree Structure Skeleton Image) to generate a tree structure skeleton image based on the collection of N tree structure sequences. Therefore, we utilized Deep First Search (DFS) to restructure and create tree structure skeletons.

2. Patients Data Base

There are six basic physiotherapy exercises in the database, which have been carefully selected to be suitable for analyzing and processing with a single camera (two-dimensional processing), see (2). There are 30 participants in the database, who each perform six exercises. Ten cycles comprise each exercise (e.g. rotating the right arm). Exercises are performed once with a right tilt and once with a left tilt (for example, once with a right foot rotation and once with a left foot rotation). A total of about 7500 motion cycle videos have been tagged and timed in the database. This study included healthy subjects (volunteers - students) with no disability identified during tests to control postural stability. The subjects group comprised of 4 men and 26 women with an average age of 21.1 (SD 1.2) years, body weight 64,8 (SD 9,4) kg and body height 170 (SD 9) cm. One single measurement of each subject was taken during the session. The study was performed in accordance with the Helsinki Declaration and the study protocol was approved by the local Ethical Committee, by the Faculty of Biomedical Engineering, Czech Technical University in Prague. The entire database has been encoded as skeletons - a skeleton in every frame (Figure 1).

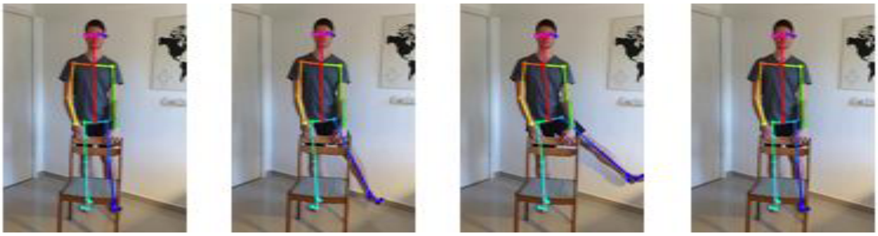

Figure 1. database has been encoded as skeletons - a skeleton in every frame

Performing exercises creates skeletal structures. The human body is represented by 25 vertices in each skeleton. The vertex has three components: Coordinate X, Coordinate Y, and Coordinate C, which indicates the level of certainty about each point in the skeleton on a scale from 0 to 1 (1-absolute certainty, 0 absolute uncertainty).

3. Gesture Generator Using Generative Adversarial Network

AI (artificial intelligence) is a revolutionary technology that is finding applications in a wide range of products and services that we use on a daily basis. An amazing application of AI technology is Deep Learning, which is based on neural networks. Deep learning-based application development requires a large dataset with a sufficient number of training examples. In the absence of such databases, researchers and data scientists use synthetic data to circumvent the problem. Through the use of synthetic data as an augmentation strategy, our research offers to demonstrate a solution for a particular database containing a limited number of video clips of human physiotherapy workouts. Using a gesture generator, skeletal gestures are generated from the videos (see Figure 2). OpenPose is utilized for creating a representation of this existing database. OpenPose is the world's first real-time collaborative system for identifying the human body in images and videos. Raw images of the human body are converted into three-dimensional skeleton vectors in Euclidean space using skeletal algorithms such as OpenPose, AlfaPose, and Mediapipe. The vector represents one position of a human being. In order to produce new samples and to govern the generated motion type, we will demonstrate in this research how to utilize a Generative Adversarial Network (GAN) as a generator.

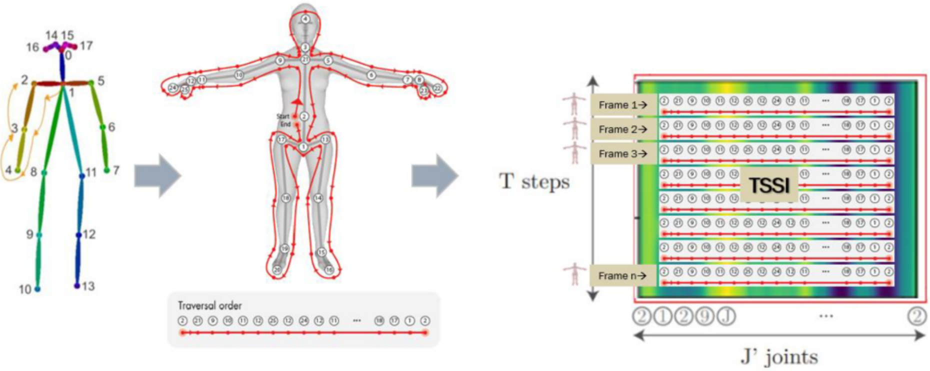

Figure 2. Tree Structure Skeleton Image (TSSI)

3.1. Network Architecture

Our system creates samples using a generator (Figure 3). A sequential layer model is designed using deep convolutional generative adversarial networks (DC-GAN). The input to the first layer consists of random noise (1x1xN) which is subsequently concatenated to the (1x1xC) class which controls our output movement type. Topology structure combines DC-GAN generator layers architecture and adaptation [3][8]. The discriminator in our system is responsible for distinguishing between the fake generator output and the true skeletal pseudo-image. The result is a value that lies within the interval [0,1] that gives the probability as follows, 0 for a phony pseudo-image and 1 for a real pseudo-image. It is suggested in papers (12) (18) that the DC-GAN discriminator might take a layered architecture. In this research we are using the Depth-First Search (DFS), a search technique for tree structures. We changed the order of joints in our skeletal model to highlight the connections between them. The samples will consist of only the gesture frames without any additional frames that are not essential.

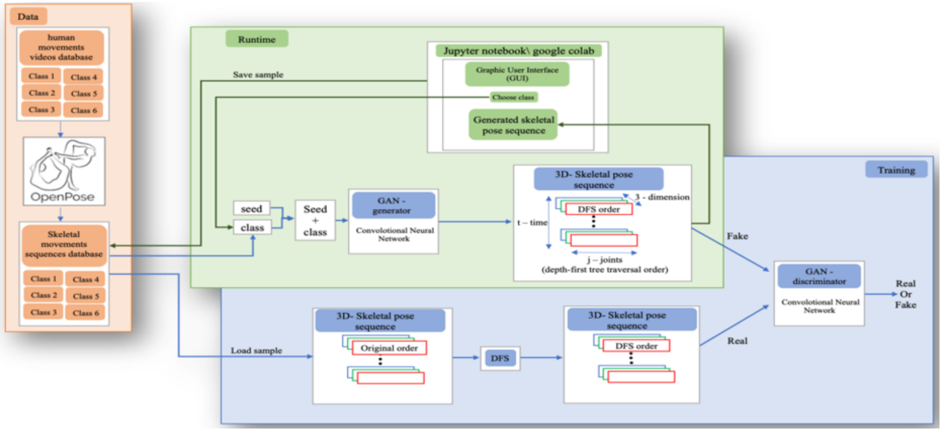
Data
human movements videos database
Class 1
Class 4
Class 2
Class 5
Class 3
Class 6
OpenPose
Skeletal movements sequences database
Class 1
Class 4
Class 2
Class 5
Class 3
Class 6
Runtime
Save sample
Jupyter notebook\ google colab
Graphic User Interface (GUI)
Choose class
Generated skeletal pose sequence
seed
class
Seed + class
GAN - generator
Convolotional Neural Network
3D- Skeletal pose sequence
3 - dimension
t – time
DFS order
j – joints
(depth-first tree traversal order)
Training
Fake
GAN - discriminator
Convolotional Neural Network
Real Or Fake
Load sample
3D- Skeletal pose sequence
Original order
DFS
3D- Skeletal pose sequence
DFS order
Real

Figure 3. GAN Concept Design

4. Demonstration of How to Measure Gesture Mimics via the Siamese Neural Network

The neural network chosen for this project is the Siamese neural network (5). The reason for selecting the Siamese network is its one-shot learning capability. The result is that once the network has been properly trained, it is possible to classify a new image into a class that was not included in the initial training. Integrated Pose Estimates (IPEs) images were developed to train the network. Using this new technique, we ware able to capture all the motion of the human body in one image. It is not necessary to use all of the frames within a time window to create a good representation of IPE. We conclude that different gestures require different time windows for optimal IPE representation.

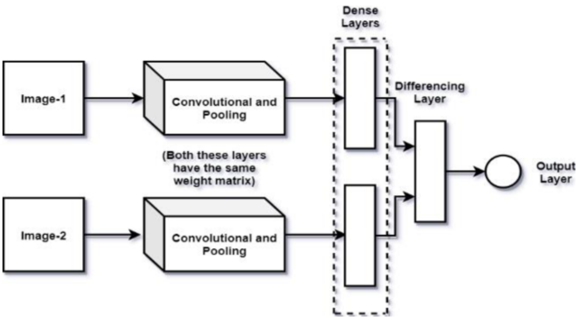
Dense Layers
Image-1
Convolutional and Pooling
Differencing Layer
(Both these layers have the same weight matrix)
Output Layer
Image-2
Convolutional and Pooling

Figure 4. Siamese Network Layout

The input to the network (see **Figure 4)** consists of a pair of IPEs with dimensions of 105x105 pixels each. Inputs are fed into the same convolutional and pooling layers and the output is a tensor with 4096 elements for each input, which can be considered as a type of code or latent of the IPE. These codes are fed into the differentiation layer, which computes their L1 distance, i.e.

$$D = flatten(|T_1 - T_2|) \quad (1)$$

where T_1, T_2 are the tensors obtained from the convolutional and pooling layers, respectively. There is only one neuron in the final dense layer that has a sigmoid activation function. We can model this layer mathematically:

$$y = \sigma(bias + \sum_i w_i D_i) \quad (2)$$

where:

- σ is the sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$(3)
- D_i is the i'th element of D
- w_i is the corresponding i'th weight

Accordingly, the output of the network is a number between 0 and 1, which correlates to the degree of similarity between the two inputted IPEs. The closer the output value is to zero, the higher the level of similarity predicted.

N-Way is a method of validating the one-shot learning process (5). We used a random subset of IPEs for every activation of a single N-Way evaluation. As part of a single N-Way evaluation, a specific IPE is randomly selected to function as the reference IPE, and its movement is considered to be the reference movement. Moreover, N other IPEs are selected at random with the constraint that only one belongs to the reference movement, while the remaining N-1 IPEs each belong to a different movement. Each of these N IPEs is paired with the reference IPE to create N pairs. Using the Siamese network, each of these pairs is converted into a corresponding output value. For an N-Way prediction to be considered accurate, the unique pair of movements must produce the minimum output value. By repeating the 'N-Val' process a number of times, the accuracy will be computed as follows:

$$Accuracy = 100 \cdot \frac{\#\, of\ correct\ predictions}{Nval}\ [\%] \quad (4)$$

5. Results

Over 4000 movements were generated for the Neural Network database and testing using a synthetic movement simulator. Through the simulator, the user is able to create choreography by combining seven predefined movements (see example in Figure 5). Synthetic movements were classified according to seven predefined categories in our system. There are five hidden layers in our classification network, which is a Siamese network. The classification of real data produces 86% accuracy, while the classification of our simulated data produces 91.82% accuracy.

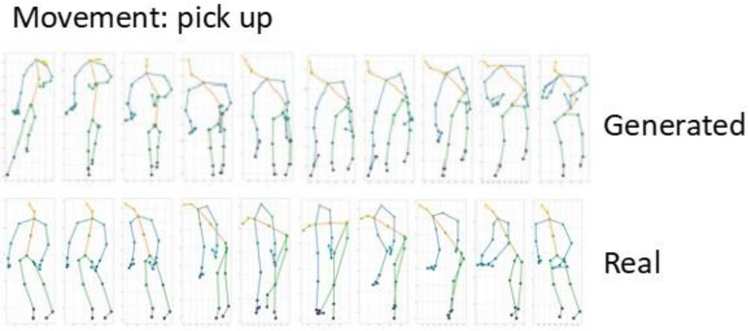

Figure 5. Results

6. Conclusions

We build a GAN simulator to feed a Siamese Network that measures, classifies, and infers human movement. The simulator produces video footage of actual physiotherapy treatment sessions. The video footages contain vectors representing skeletons, thus allowing us to recreate human movement sequences. We present the idea of how a Generative Adversarial Network (GAN) works and how we used it as a tool that makes it possible to create and control human body motion parameters. We reorganized the joints of our skeleton model in a tree structure search approach reorganized using DFS to emphasize their connections. Data collection of human gestures is usually fraught with challenges, such as maintaining synchronization between activities over time and space. The research investigates some of these and other standard stumbling blocks. It is interesting to note that a by-product of our study can be a virtual dance simulator that choreographers can use to translate choreographic sequences into dance steps.

7. Acknowledgment

This work was supported by the Israel Innovation Authority (Formerly the Office of the Chief Scientist and MATIMOP) & The Ministry of Education, Youth and Sports of the Czech Republic. The described research was supported by the project No. LTAIZ19008 Enhancing Robotic Physiotherapeutic Treatments using Machine Learning awarded in frame of the Czech-Israeli cooperative scientific research program.

References

[1] Segal Y, Yuval Y, Dahan O, Birman R, Hadar O, Kutilek P, et al. Camera Setup and OpenPose software without GPU for calibration and recording in telerehabilitation. In IEEE E-Health and Bioengineering; 2021; Lasi, Romania.

[2] Kutilek P, Hejda J, Lhotska L, Adolf J, Dolezal J, Hourova M, et al. Camera System for Efficient non-contact Measurement in Distance Medicine. In 2020 19th International Conference on Mechatronics - Mechatronika (ME); 2020; Prague. p. 1-6.

[3] Adolf J, Doležal J, Macaš M, Lhotská L. Remote Physical Therapy: Requirements for a Single RGB Camera Motion Sensing. In: 2021 International Conference on Applied Electronics. Plzeň: Západočeská univerzita v Plzni, 2021. ISSN 1803-7232.

[4] Nicolò Carissimi PRCBVM. Filling the Gaps: Predicting Missing Joints of of Human Poses Using Denoising Autoencoders. 2019 January 29.

[5] Koch. Siamese Neural Networks for One-Shot Image Recognition. thesis. Toronto: University of Toronto, Graduate Department of Computer Science ; 2015.

[6] Cai , Xu , Yi J, Huang J, Rajasekaran S. DTWNet: a Dynamic TimeWarping Network. Advances in neural information processing systems 32. 2019.

[7] Hidalgo G, Sheikh , Kitani K, Bansal A, Sanabria R, Xiang D, et al. OpenPose: Whole-Body Pose Estimation. April 2019.

[8] Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, et al. Mediapipe: A framework for perceiving and processing reality. In Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR).; 2019.

[9] Adolf J, Doležal J, Kutílek P, Hejda J, Lhotská L. Single Camera-Based Remote Physical Therapy: Verification on a Large Video Dataset. Applied Sciences. ; 2022: ISSN.

[10] Liao Y, Vakanski A, Xian M. A Deep Learning Framework for Assessing Physical Rehabilitation Exercises. IEEE Transactions on Neural Systems and Rehabilitation Engineering. Feb. 2020; 28(2): 468-477.

[11] Heaton. Artificial Intelligence for Humans: Amazon; 2021.

[12] Xi W, Devineau G, Moutarde F, Yang J. Generative Model for Skeletal Human Movements Based on Conditional DC-GAN Applied to Pseudo-Images. Algorithms. 2020; 13(12).

[13] Yang Z, Li Y, Yang J, Luo J. Action Recognition With Spatio–Temporal Visual Attention on Skeleton Image Sequences. IEEE Trans. Circuits Syst. Video Technol. 2019; 29(8): 2405–2415.

[14] Caetano C, Sena J, Brémond F, dos Santos JA, Schwartz WR. SkeleMotion: A New Representation of Skeleton Joint Sequences Based on Motion Information for 3D Action Recognition. [Online].; 2019. Available from: http://arxiv.org/abs/1907.13025

[15] Ren B, Liu M, Ding R, Liu H. A Survey on 3D Skeleton-Based Action Recognition Using Learning Method. [Online].; 2020 [cited 2021 Oct. 20. Available from: http://arxiv.org/abs/2002.05907.

[16] Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Van Gool L. Pose Guided Person Image Generation. [Online].; 2018 [cited 2021 Oct. 15. Available from: http://arxiv.org/abs/1705.09368.

[17] Caetano C, Brémond F, Schwartz WR. Skeleton Image Representation for 3D Action Recognition based on Tree Structure and Reference Joints. [Online].; 2019 [cited 2021 Oct. 20. Available from: http://arxiv.org/abs/1909.05704.

[18] Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. [Online].; 2016 [cited 2021 Oct. 21. Available from: http://arxiv.org/abs/1511.06434.