Identifying Dancers and Style from Motion Capture Data Using ResNet

By

ALICIA B. ALARIE

THESIS

Submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA,
DAVIS

Approved:

Michael Neff, Chair

Kwan-Liu Ma

Yong Jae Lee

Committee in Charge

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT OF THE THESIS

Identifying Dancers and Style from Motion Capture Data Using ResNet

This work aims to apply advancements in deep learning for image classification to improve the recognition of movement style in motion capture data. A RESNET architecture is used to classify individual dancers based on clips of their movement and to predict style based on clips of various motions in 7 different style categories -angry, childlike, depressed, proud, etc. Motion capture clips from trained dancers at George Mason University performing the same choreographic sequence several different times were used for a dancer identification task. A data set of actions performed with different labeled styles such as proud, depressed, angry, old, and childlike created by [39] was used for a style identification task. Results were compared using Quaternion, scaled positional coordinates, and Euler angle representations of the motion capture clips supplied to the network for learning.

Chapter 1

Introduction

Motion capture technology has allowed for very realistic motions of animated characters. The capability to string together multiple clips of motion capture movement for a seamless effect can save animators lots of time. But efficient techniques for labeling and indexing these clips within large databases has been a challenge. Often it's helpful to be able to search for multiple motion clips of a similar style (sitting down, happy, depressed, child character) when composing a scene. But it can be difficult to tag and classify these motions in an automated way because motions that seem similar to the human eye may not have numerically similar coordinates for each pose in 3d space. For example, with a raw 3d space coordinate representation, the action of reaching forward with one's arm would not be the same numerically as reaching forward after a 90 degree rotation of the entire body.

To this end, many techniques have been explored as a means to identify similarities in style of movement with motion capture, including machine learning. With the rise in popularity of machine learning, a variety of different neural network techniques and architectures have been applied to the problem of identifying similar movement with motion capture data. Many advancements in machine learning architecture have come through applications of these networks to image classification and segmentation problems, and residual networks (ResNets) have recently been shown to achieve better performance in these tasks. Advancements in deep

learning have allowed for higher performance in image classification by solving the vanishing gradient problem, and allowing the use of deeper architectures [12]. Residual networks have also improved performance by implementing "shortcut connections" that fit residuals rather than full functions [13] . These methods have resulted in much higher performance in image classification and image segmentation tasks, and have recently been applied to problems in the field of animation.This work explores whether they can achieve better performance than non residual network architectures in tasks of identifying similar motion capture.

This work applies a ResNet deep learning architecture to motion capture data for two different classification tasks on motion capture data.The classification accuracy of the ResNet model is compared to that of a human performing the same task. In addition, this work compares the performance of neural networks on motion capture data in joint angle representation vs in quaternion representation, and finds similar performance using each type of representation.

Specifically, this work uses data-driven methods for identifying style of individuals during dance motion and style of movement on two different data sets using a ResNet network. The network is trained to predict labels for each individual and for each movement style by learning its own representations of similarity within label groups. A 50-layer ResNet architecture is used for these two different style classification tasks. The first task is to identify individual dancers and the second task is to classify motion styles (angry, childlike, depressed, old, proud, etc.). The tasks are performed on two different data sets. The data set for identifying individual dancers consists of ten trained dancers performing the same choreography on ten different training runs. The data set for identifying style of motion consists of clips created by [39] containing many different actors performing a variety of actions (walking, jumping, punching) in 7 different style categories (angry, proud, childlike, old, etc.). Smith [33] used deep neural networks for efficient real-time motion style transfer. As part of his work, a human subject study was conducted to test how easily each style could be identified by a person, and how clear each affect was to the human eye. This human

subject study is shown in this work as a comparison to the performance of our network on the same data set. The purpose of this is to gauge the performance of the network vs human performance.

An investigation into motion capture representations is also done. Positional and joint angles representations are compared. As well as Euler angle and Quaternion representations of joint angles. Raw motion capture data can consist of one point for each joint in the skeletal model with x, y, and z coordinates in space over time for each joint. Another way of representing motion capture data is using joint angles, which only convey information about the angle of each joint relative to each other and do not have information about distance traveled in space directly. Both Euler Angles and Quaternions have been used to represent joint angles, however, Euler angles inherently can have more discontinuities in the representations, which motivates our comparison of the two representations.

1.1 Background

Much work has been done on identifying similar movements in motion capture data. This is a fruitful field of research for animators as it allows for automation of some parts of the process of creating animated motion endowed with the style of a particular character, emotional emotional affect, or action. Once the features of a style are identified, motion retrieval and style transfer techniques allow for efficiently retrieving and stringing together motions in that style. Motion retrieval aims to retrieve a sequence of motion capture data in a certain style in order to generate animated movement by connecting sequences of movement that already exist in memory and share the desired characteristics of action or style. Styles can be similar in action (running), emotional affect (happy), or individual performer. Style transfer allows for editing/controlling affect in a clip of motion, such as walking, to transform it, for example, from a neutral walk to a happy or sad walk. Both of these methods avoid hours of labor costs for animators, and similar motion identification is a key building block

that's necessary to accomplish both of these tasks.

However, identifying similar motions can be difficult because often motions that look similar to a human observer may not be similar numerically on a skeleton representation. Data-driven methods have found success for this reason, as they are better able to identify similar motions even if the motions have some transform applied to them such as rotation, translation in space or are performed at slightly different speeds. One popular data-driven method of style representation is to use deep auto-encoders, which can create efficiently compressed representations of style. Wang used deep auto-encoders to create a 20-bit representation of features that show similarity in the action being performed and/or the style of movement [36]. Holden also used deep auto-encoders to identify create style representations [14]. Earlier motion retrieval works used shallower networks in PCA or ICA to identify features that could be used to identify similarity of motions [23] [11], but deeper networks were able to better identify motions that are similar to the eye but may not be numerically similar [36].

Prior to these techniques, many non-machine learning techniques were developed. Rather than using representations for similarity, some authors such as Kovar and Keogh used raw motion data in either positional or joint angle format to manually compute similarity via Euclidean Distance [19, 17] or weighted distance [27]. Other works identified motion by movement from separate body parts [38, 9, 22] or dynamic time warping to compare similarity on different time scales.

Other papers have been published using neural networks for a variety of tasks, such as pose estimation, with motion capture data, however none to current knowledge have used the ResNet architecture for style identification. Mehta and Xiao used ResNet as a performance baseline for the task of learning skeletal pose from images containing people [26, 40]. Xiao, Chen, and Papandreou have also used ResNet for pose estimation [41, 6, 28].

One goal of this work is to see if a neural network with improved performance on image classification and pose estimation (ResNet) can provide better performance in classifying

similar styles of movement in motion capture data. In 2015, deeper neural networks were made feasible by solving the vanishing gradient problem using "Relu" activation functions [12]. And the ResNet architecture achieved even better performance than regular deep networks in image classification tasks on COCO and ImageNet data sets in 2016 using "shortcut connections" to fit residual functions rather than attempting non-linear fitting without breaking the fit down into residual blocks [13].

The first style recognition task in this work is identifying individual dancers based on their movement in motion capture. Work on identification of people has been done for security/surveillance purposes as early as 1977 [8]. Research in this area continues today with the use of data-driven methods to achieve more accurate results [3]. Many works have been able to use motion capture specifically of gait to identify individuals [20] [2]. Carlson focused on individual identification during dancing activity specifically, but this was a "free dancing" activity, so movements were not choreographed [5] . Instead, each person moved in their own chosen way. Thus, the classification was based on each individual's personal embodiment of the music. In contrast, this work identifies individual dancers who are performing the same choreography (movements) at the same time.

The second task the network is trained to do is identify the style of motion (such as "happy","depressed","childlike","strutting"). Early work in style classification was algorithmic in nature. Laban Movement Analysis is a system for determining emotional affect in movement that was created prior to 1958 [21]. Russell's Circumplex model classifies emotion on a 2D spectrum [31] and was used by [1]. Early work in being able to create a quantifiable feature space to represent emotional affect is discussed by Pollick [30]. Hsu used a linear time-invariant model to characterize movement style for style translation of human motion [15] .

More recent works have used machine learning methods for identifying emotional affect. Holden, Du and Wang use auto-encoder architectures to learn these features, along with broader style features [14, 10, 36]. Loghmani uses RNN's [24]. Karumuri used a 3-layer

CNN to classify emotional affect/style [16]. Xia uses a mixtures of auto-regressive models to capture relationships between styles of movement [39] . Kim uses independent component analysis to model the emotional affect or style of different parts of the body [18] . However, the ResNet architecture has not been applied to style classification. This work attempts a simple classification task of predicting the style labels on clips of motion.

One challenge with this task is that when we look at an animated character, depending how subtle the movement is, it can sometimes be difficult even for a human to identify the style of motion or emotional affect. One person may think someone is "excited" and someone else could categorize them as a high-energy person who is just "happy". This ambiguity, to even the human eye, can make it difficult to validate the results of style classification. Some studies have focused on the basic question of whether emotional affect was able to be reliably identified by humans. Walbott and Crane studied this topic, and concluded that emotional affect during walking was recognizable by humans and was able to be elicited in a laboratory setting [35, 7]. However, they did not achieve full accuracy of human classification of affect. Some studies have also attempted to predict permanent personality traits of individuals from motion capture. Wang assesses the impact of hand motion on virtual character personality [37]. Smith focuses on the impact of animated gesture performance on personality perceptions and showed that people's judgments of character personality mainly fall in a 2D subspace rather than independently impacting the full set of traits in the standard Big Five model of personality [34]. Carmurri showed that free dancing movement style can be indicative of personality traits in the Big Five such as extroversion and empathy [4]. This motivates the comparison of results from this work to a human study that uses the same data.

This work also investigates how well different representations of motion capture data perform when used for machine learning applications. Pavllo found that quaternion representations of joint angles in motion capture data were able to achieve more accurate results in motion prediction problems using recurrent neural networks (RNNs) [29]. This work tests

the accuracy of Euler Angle vs Quaternion representations of joint angles as well as accuracy of raw and scaled positional representations.

1.2 Methods

1.2.1 Motion Capture Data Sets

As mentioned previously, there are two separate data sets used for two separate identification tasks. The first data set from [25] has clips from ten dancers trained at George Mason University School of Dance. Each dancer was recorded performing the same choreography (set of movements) during ten different training runs and the data is labeled by individual dancer ID. The task performed on this data set is the identification of dancer ID. The complete choreography is about 1 minute long, so the total data set is about 100 minutes of motion capture data. A frame from this data set is shown below in 1.1.

Figure 1.1: This image shows skeletons from motion capture data overlaid with one another in the Dance choreography data set. There are ten Dancers total who each perform the same choreograph. Small differences in their timing and poses can be observed from this frame.

The second data set from [39] consists of actors performing several different types of actions (walk, fast walk, running, jumping, punching, fast punching, kicking, transitions) in a variety of labeled styles including "angry", "depressed", "old", "childlike", "proud", "sexy", and "strutting". The task performed for this data set is identifying the style of movement.

1.2.2 Data Pre-processing

Several different representations of motion capture data exist and are tested for accuracy and biases including positional representations, Euler angle representations, and Quaternion representations. As mentioned previously raw motion capture data can consist of one point for each joint in the skeletal model with x, y, and z coordinates in space over time for each joint. Whereas joint angles only convey information about the angle of each joint relative to each other and do not have information about distance traveled in space directly. Both Euler Angles and Quaternions can be used to represent joint angles, however, Euler angles inherently can have discontinuities in the representations whereas Quaternions have less discontinuities. All types of representations undergo the same pre-processing.

To format the data in each representation for ingestion into the neural network, clips were broken down into quarter second sub-clips, and re-formatted into image format shown below in Figure1.2. The image was constructed as in [16] so that each row of the image corresponds to the joint angle of a particular joint in the skeleton, and each column is the time progression of the angle of the joint. The roll, pitch, and yaw of the joint angle are represented by the red, green, and blue channels of the image (rgba was used for quaternion data). The resulting images track in detail the movement of each joint.

1.2.3 Neural Network Architecture and Data Pipeline

The network architecture for our neural network is shown in Figure 1.3 below [12] [32]. The architecture is composed of ResNet blocks. This architecture was chosen due to its success

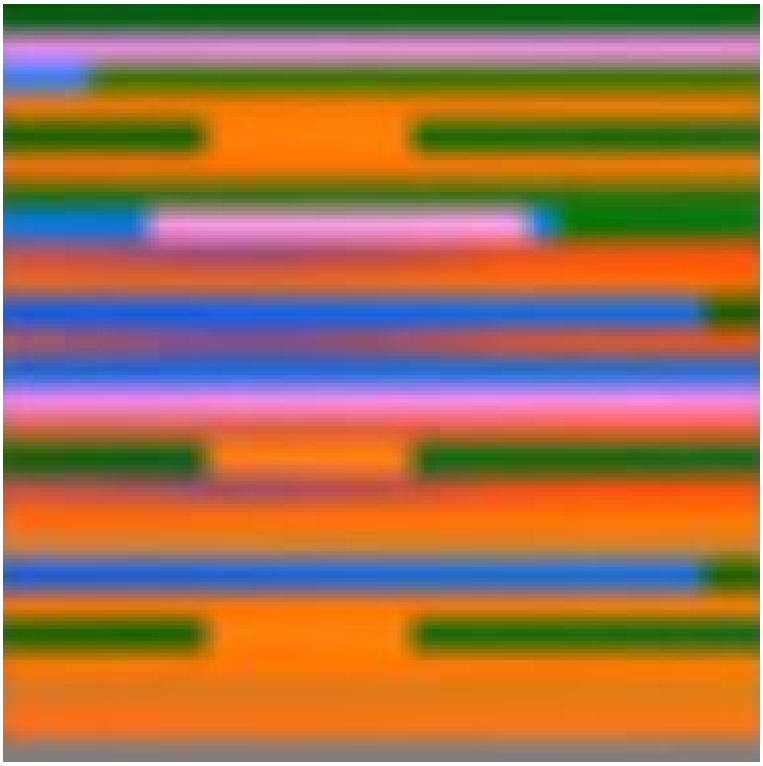


Figure 1.2: Example of motion capture sub-clip that is input to neural network for classification. For each pixel of the image, the row corresponds to the joint of the skeleton (e.g., left shoulder, right elbow, etc.) and the column corresponds to the progression of the joint angle over time. The color of the pixel is created by assigning the Roll, Pitch, and Yaw angle values to the R, G, and B channels of the color image. The change in color of some rows illustrates a movement of some joints, either synchronously or in a particular succession.

in image classification. In 2016, the ResNet model outperformed non-residual deep CNN models in image classification and segmentation tasks. This architecture is used due to it's potential for good performance compared to earlier methods that used deep or shallow CNN architectures. The residual network adds an identity shortcut that allows the network to create an initial "fit" for a task and then to also fit smaller and smaller residual functions to come up with a more accurate overall fit for the task at hand. This architecture is shown in figure 1.3.

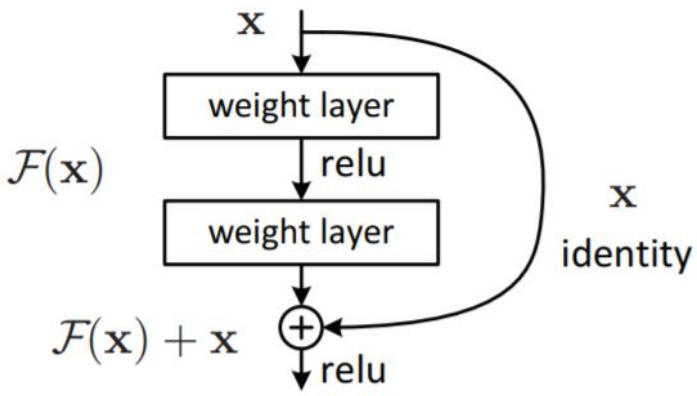


Figure 1.3: Residual learning block

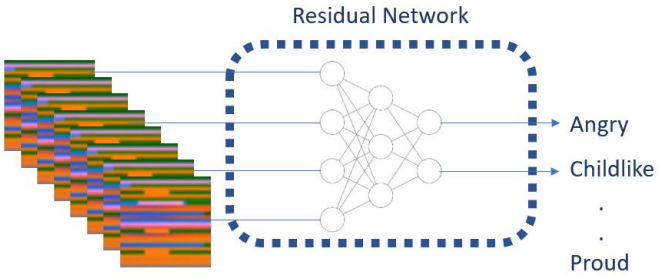


Figure 1.4: A flow diagram showing data that is fed into the residual network which predicts style labels.

1.3 Results

This section outlines the results of experiments with each data set. The Dance data experiment tries to identify individual Dancers, and the Style data experiment aims to identify style of motion when performing a variety of tasks such as jumping, walking, and running. Each experiment uses the same pre-processing methods.

1.3.1 Dancer Identification Results

This section compares results for identifying individual dancers based on motion capture data. There are several options for representation of motion capture data, and different representations are compared here to test if one results in less accuracy or increased biases compared to others. The representations tested include Euler Angles, Quaternions, and po-

sitional formatted data. Euler Angles and Quaternions represent joint angles of a person's skeletal model, and positional data represents x,y,z coordinates in space. One concern with Euler Angle representation is discontinuous jumps that can occur inherently with this representation. Quaternions also represent joint angles, but are able to represent variation in joint angle continuously, avoiding the discontinuities with Euler Angles. Thus, the network was expected to be able to learn better from a Quaternion Representation.[29] were able to achieve better motion prediction results using Quaternion representations of joint angles. Positional representations differ from joint angle representations in that the joint angle representations do not include information about the Dancer's position in space, only information about the movement of their joints is preserved. The results for positional data thus have more information for the network to learn from and are expected to be more accurate. One concern with the positional representation is that the network would be able to learn to differentiate individuals based on their differences in height and limb length. For this reason, a representation was created that mapped the positional coordinates for each dancer onto the same skeleton model for all Dancers. The skeletal model was created by averaging all the joint offsets for each Dancer to find the average height and limb length skeletal model. This was done to minimize as much as possible any errors with the projection such as foot sliding.

Table 1.1 shows the comparison of accuracy for positional formatted motion capture data (x,y,and z over time), Uniform skeleton model positional formatted motion capture data, Euler Angle representation of joint angles (roll, pitch, and yaw), and Quaternion representation of joint angles. From the table, it appears that the positional data representation outperforms angular representations when the positional skeleton model for each dancer is not scaled to a uniform size, and that the Quaternion representation does perform 1 percent better than the Euler Angle representation of joint angles. The F1 Score of 1.0 for the unscaled positional representation also suggested that the presence of differences in the skeletal proportions of each dancer had a strong effect on predictions. Figure 1.6 shows the learning

Representation	Average Accuracy (F1 Score)
Positional	1.0
Positional Scaled	0.82
Euler Angle	0.85
Quaternion	0.86

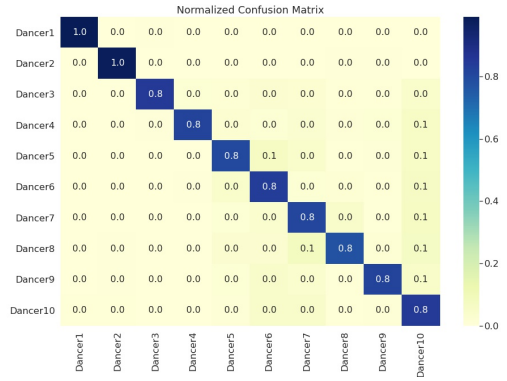
Table 1.1: Comparison of F1 Score for Dancer Identification task using different representations.

curves for each representation. The "train loss" and "val loss" are the curves for training and validation loss, and the "train acc" and "val acc" show curves for training accuracy and validation accuracy.

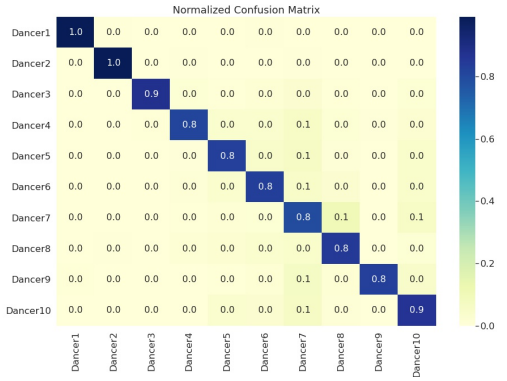
The Confusion Matrix for each representation is shown in Figure 1.5 to show whether some Dancers are easier to distinguish than others. "Dancer 10" appears to be confused the most for other Dancers, especially "Dancer 6" and "Dancer 7". It should be noted that the "Dancer 10" data was missing one training run of data. For confusion matrix results that are not normalized, refer to the appendix.

Dancer Identification with Single Frame vs Multiple Frames

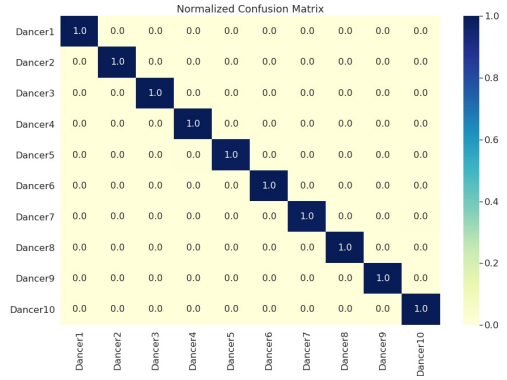
Table 1.2 below shows that with the positional data, it's possible to predict based on single frames/poses, whereas the angular representation is not able to predict based on a single pose. One possible reason for this could be because the angular representation only shows relative pose information and is more independent of height and physical proportions of each individual. If it is predicting on a single pose, this likely means it is using the size of the dancer as factor in it's predictions. Figures 1.7 and 1.8 show the learning curves for each case. When supplied pose data without absolute positional data, the network is not able to learn and classify dancers correctly. This suggests the positional representation includes biases from the dimensions of the dancers. The confusion matrix for each case is included in the appendix.



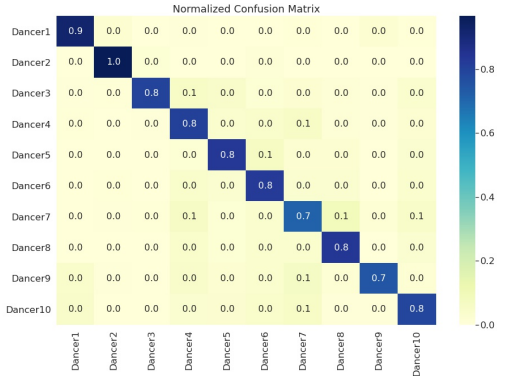
(a) *EulerAngleRepresentation*



(b) *QuaternionRepresentation*



(c) *PositionalRepresentation*

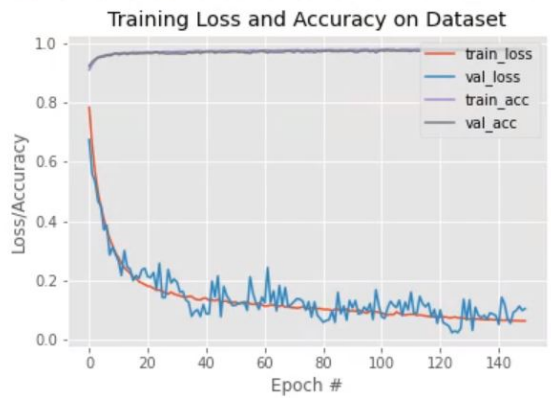


(d) *ProjectedPositionalRepresentation*

Figure 1.5: Normalized Confusion Matricies for identifying Dancers with different motion capture data representations.

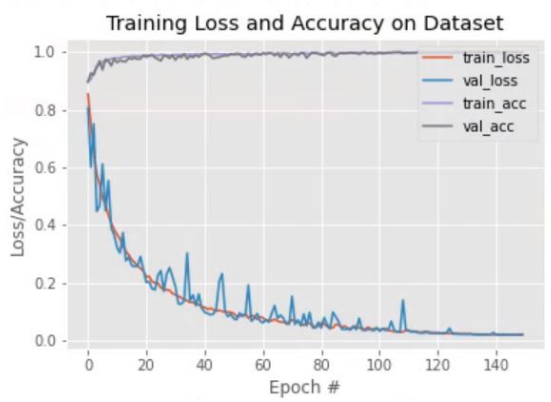
Representation	**Average Accuracy (F1 Score)**
Positional	0.98
Euler Angle	0.13

Table 1.2: F1 Score for Dancer Identification task using a single frame for predictions.

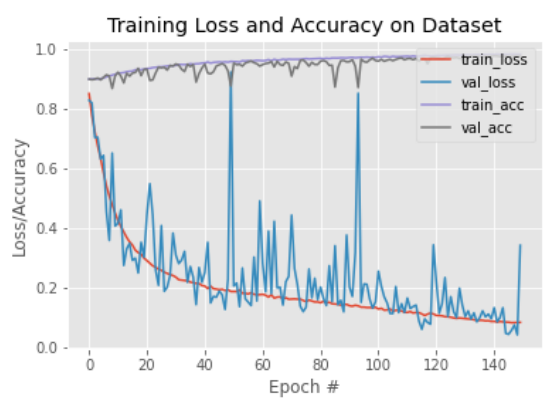


(a) *EulerAngleRepresentation*

(b) *QuaternionRepresentation*



(c) *PositionalRepresentation*



(d) *ProjectedPositionalRepresentation*

Figure 1.6: Learning Curves for Identifying Dancers with Different motion capture data representations.

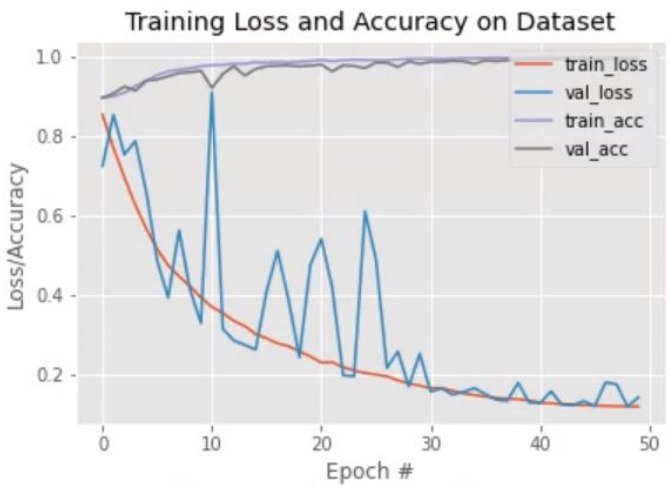


Figure 1.7: Learning Curve for Positional Data Prediction on Single Frame

1.3.2 Classification of Emotional Affect

This section shows results of training the network to identify style of a motion capture clips. As mentioned before, there are 6 styles in this data set, “angry”, “childlike”, “depressed”, “neutral”, “old”, “proud”, “sexy”, and “strutting”. The database and styles were created by [39]. Since this task can also be tough for humans to do, the results are compared to results from a human study in [33] which asked humans to identify style of motion for the same data set. A Quaternion representation was used for this data set. The overall average accuracy of our model in discerning style from the data set in [39] is .63. Accuracy for each style is shown in Figure 1.9. Relatively higher accuracy was achieved for the “old” and “strutting” styles, and lower accuracy was achieved for childlike” and “sexy” styles of movement. The overall avg F1 score for this model was lower than the F1 score for the dancer identification data set. This could be due to the small training data size of this data set which has 79829 total frames with 6 classes of styles (including a “neutral” style). The dancer identification data set was composed of approximately 384,000 frames.

Compared to the human subject study from [33], results are similar for many affect types. However, our network did a better job at identifying “Strutting”, and the human subjects did a better job at identifying ”Depressed”. Overall, the performance seems comparable to

Figure 1.8: Learning Curve for Prediction on Single Frame of Euler Angle Representation Data

human ability.

1.4 Conclusion

This work focused on measuring the ability of a ResNet architecture to identify individuals during dance motion and to identify style of movement during non-dance motions. The same neural network was used for both tasks. For the dance motion data set, an investigation was also performed as to whether joint angle representation makes a difference in accuracy of the model, and whether using positional data introduces biases due to differing heights of dancers. It was found that the Quaternion representation was 1 percent more accurate than an Euler angle joint angle representation, potentially due to the absence of discontinuities in the quaternion representation. It was also found that using positional representations without scaling for differences in height may introduce biases due to the differing skeletal models of each dancer with unique skeletal proportions. An attempt was made to reduce these biases by mapping the positions for each dancer onto an average skeleton to remove differences in frame. However, this resulted in a slightly less accurate model than the joint angle representation, and is thought to have introduced foot sliding errors. This could be an

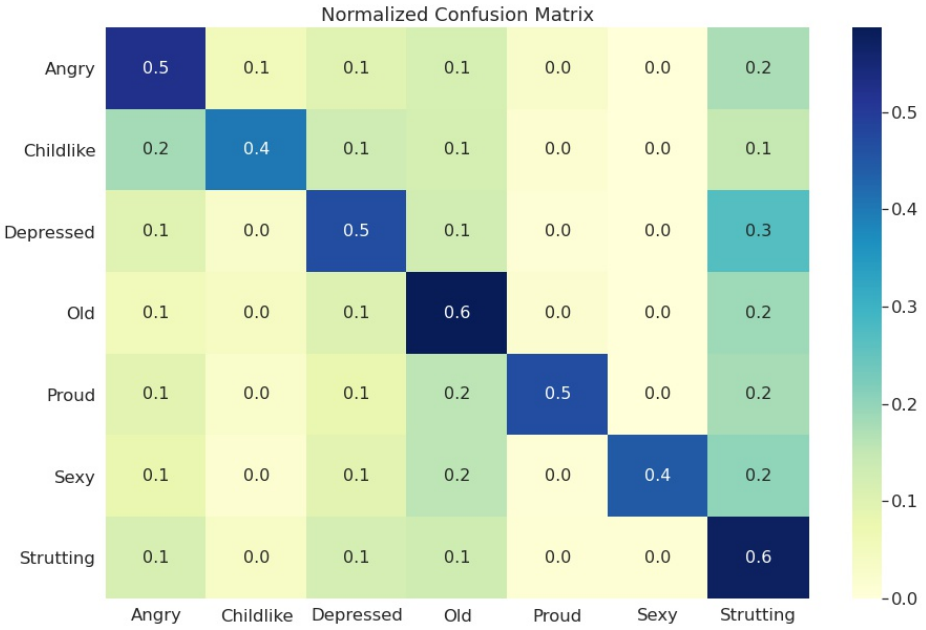


Figure 1.9: Normalized Confusion Matrix for Style Identification

area for future research.

The same network was trained to identify style from a data set containing a variety of actions such as running and jumping in 6 different style categories. The accuracy of this model is compared to a study that asked humans to identify the same styles on the data set. It was found that the model performed better than human subjects in some categories and worse than the human subjects in other categories.

The network may have performed better on the task of Dancer Identification than on the task of style classification due to the larger training data size of the Dancer Identification data set.

Future work could compare accuracy of using ResNet blocks in auto-encoder networks for style encoding in motion retrieval to regular deep neural network auto-encoders.

Figure 1.10: Learning Curve for Style Identification

Human Subject Study on Classification of Emotional Affect (Xia Dataset)

Affect	Identified (Human Study)	Misidentified (Human Study)	% Correct (Human Study)	% Correct Our Network
angry	34	31	.52	.5
childlike	32	23	.58	.4
depressed	47	8	.85	.5
neutral	-	-	-	-
old	38	17	.69	.6
proud	22	33	.40	.5
sexy	7	48	.13	.4
strutting	24	31	.44	.6

Table 1.3: Comparison of human subject performance to our network performance for the style identification task using the data set from Xia et al. The number of questions for each style in the human study is included as number identified correctly and incorrectly.

Bibliography

[1] A. Aristidou, Q. Zeng, E. Stavrakis, K. Yin, D. Cohen-Or, Y. Chrysanthou, and B. Chen. Emotion control of unstructured dance movements. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 1–10, 2017.

[2] M. Balazia and P. Sojka. Gait recognition from motion capture data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1s):1–18, 2018.

[3] S. Bhowmik, A. Ghosh, J. Debsinha, R. Kajal, and A. Professor. A literature survey on human identification by gait. *Imp J Interdisc Res*, 2(7):1340–1343, 2016.

[4] A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International journal of human-computer studies*, 59(1-2):213–225, 2003.

[5] E. Carlson, P. Saari, B. Burger, and P. Toiviainen. Dance to your own drum: Identification of musical genre and individual dancer from motion capture using machine learning. *Journal of New Music Research*, 49(2):162–177, 2020.

[6] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[7] E. Crane and M. Gross. Motion capture and emotion: Affect detection in whole body movement. In *International Conference on Affective Computing and Intelligent Interaction*, pages 95–101. Springer, 2007.

[8] J. E. Cutting and L. T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society*, 9(5):353–356, 1977.

[9] Z. Deng, Q. Gu, and Q. Li. Perceptually consistent example-based human motion retrieval. In *Proceedings of the 2009 symposium on Interactive 3D graphics and games*, pages 191–198, 2009.

[10] H. Du, E. Herrmann, J. Sprenger, N. Cheema, S. Hosseini, K. Fischer, and P. Slusallek. Stylistic locomotion modeling with conditional variational autoencoder. In *Eurographics (Short Papers)*, pages 9–12, 2019.

[11] K. Forbes and E. Fiume. An efficient search algorithm for motion data using weighted pca. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 67–76, 2005.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] D. Holden, J. Saito, T. Komura, and T. Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, pages 1–4. 2015.

[15] E. Hsu, K. Pulli, and J. Popović. Style translation for human motion. In *ACM SIGGRAPH 2005 Papers*, pages 1082–1089. 2005.

[16] S. Karumuri. From Motions to Emotions: Classification of Affect from Dance Movements using Deep Learning. `https://dl.acm.org/doi/pdf/10.1145/3290607.3312910`, 2019.

[17] E. Keogh, T. Palpanas, V. B. Zordan, D. Gunopulos, and M. Cardle. Indexing large human-motion databases. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 780–791, 2004.

[18] Y. Kim and M. Neff. Component-based locomotion composition. University of California, 2012.

[19] L. Kovar and M. Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Transactions on Graphics (ToG)*, 23(3):559–568, 2004.

[20] I. Kviatkovsky, I. Shimshoni, and E. Rivlin. Person identification from action styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 84–92, 2015.

[21] R. Laban and L. Ullmann. The mastery of movement. ERIC, 1971.

[22] F. Liu, Y. Zhuang, F. Wu, and Y. Pan. 3d motion retrieval with motion index tree. *Computer Vision and Image Understanding*, 92(2-3):265–284, 2003.

[23] G. Liu, J. Zhang, W. Wang, and L. McMillan. A system for analyzing and indexing human-motion databases. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 924–926, 2005.

[24] M. R. Loghmani, S. Rovetta, and G. Venture. Emotional intelligence in robots: Recognizing human emotions from daily-life gestures. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1677–1684. IEEE, 2017.

[25] E. McKenna. *Examining the Spatial and Temporal Properties of Unconstrained Motor Skill Learning*. PhD thesis, George Mason University, 2019.

[26] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 39(4):82–1, 2020.

[27] J. Meng, J. Yuan, M. Hans, and Y. Wu. Mining motifs from human motion. In *Eurographics (Short Papers)*, pages 71–74, 2008.

[28] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017.

[29] D. Pavllo, D. Grangier, and M. Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018.

[30] F. E. Pollick. The features people use to recognize human movement style. In *International gesture workshop*, pages 10–19. Springer, 2003.

[31] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[32] G. Shi. Implementing a ResNet model from scratch. `https://towardsdatascience.com/implementing-a-resnet-model-from-scratch-971be7193718`, 2019.

[33] H. J. Smith. Efficient Neural Networks for Real-time Motion Style Transfer. `https://dl.acm.org/doi/pdf/10.1145/3340254`, 2019.

[34] H. J. Smith and M. Neff. Understanding the impact of animated gesture performance on personality perceptions. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[35] H. G. Wallbott. Bodily expression of emotion. *European journal of social psychology*, 28(6):879–896, 1998.

[36] Y. Wang and M. Neff. Deep signatures for indexing and retrieval in large motion databases. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pages 37–45, 2015.

[37] Y. Wang, J. E. F. Tree, M. Walker, and M. Neff. Assessing the impact of hand motion on virtual character personality. *ACM Transactions on Applied Perception (TAP)*, 13(2):1–23, 2016.

[38] S. Wu, Z. Wang, and S. Xia. Indexing and retrieval of human motion data by a hierarchical tree. In *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*, pages 207–214, 2009.

[39] S. Xia, C. Wang, J. Chai, and J. Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)*, 34(4):1–10, 2015.

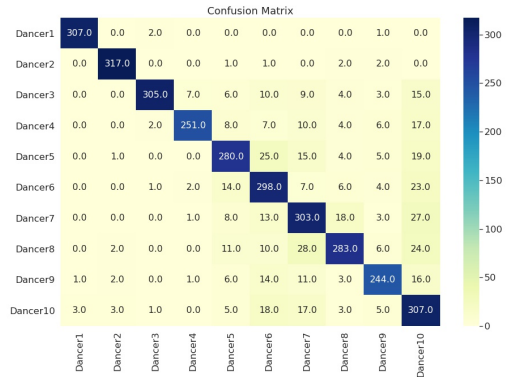
[40] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[41] X. Xiao and W. Wan. Human pose estimation via improved resnet-50. 2017.

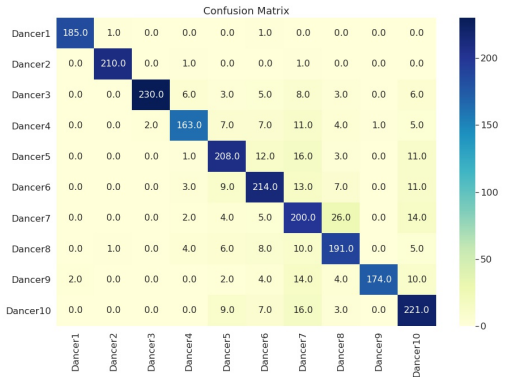
Appendix A

Appendix Title

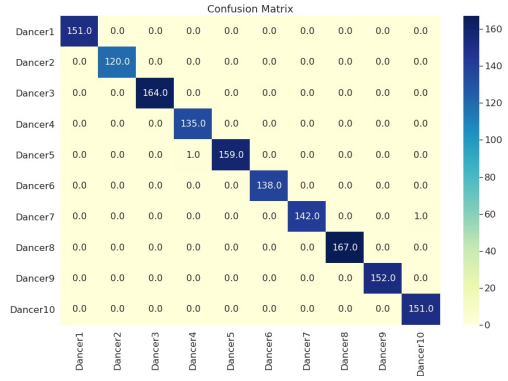
A.1 Additional Figures



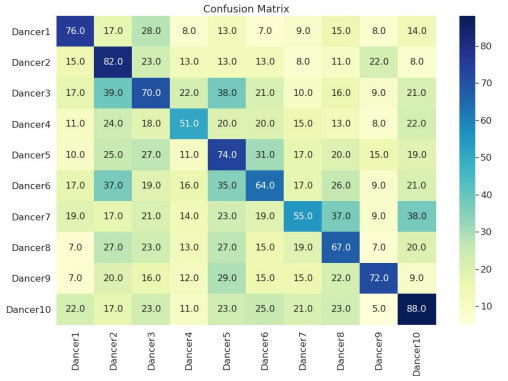
(a) *EulerAngleRepresentation*



(b) *QuaternionRepresentation*



(c) *PositionalRepresentation*



(d) *ProjectedPositionalRepresentation*

Figure A.1: Non-normalized Confusion Matricies for identifying Dancers with different motion capture data representations.