ISSN: 2581-4311
DOI: doi.org/10.30967/IJCRSET/Chengjie-Lu/177

URF PUBLISHERS
connect with research world

International Journal of Current Research in Science, Engineering & Technology

https://urfpublishers.com/journal/ijcrset

Vol: 8 & Iss: 1

Research Article

A Dual-Path Attention Fourier Convolutional Network for Human Motion Prediction

Chengjie Lu*

School of Electronics and Information Engineering, Shenzhen University, Shenzhen, China

Citation: Lu C. A Dual-Path Attention Fourier Convolutional Network for Human Motion Prediction. *Int J Cur Res Sci Eng Tech* 2025; 8(1), 275-279. DOI: doi.org/10.30967/IJCRSET/Chengjie-Lu/177

Received: 09 April, 2025; **Accepted:** 15 April, 2025; **Published:** 17 April, 2025

***Corresponding author:** Chengjie Lu, School of Electronics and Information Engineering, Shenzhen University, Shenzhen, China

ABSTRACT

Human Motion Prediction is developed to project human motion in the future frames. With lots of papers tend to predict the future motion via Recurrent Neural Network, Multi-Layer Perceptron, or Graph Convolution Network, many complicated motions have been predicted more accurately. However, most existing methods have met drawbacks in long-term predictions. To make the long-term prediction better, we propose a network called DAFCN, using Fast Fourier Convolution to optimize the result on long-term human motion prediction. Our approach also makes advantages of the Motion Attention Model to abstract the short-term useful information. Furthermore, our approach attempts to make prediction in short-term and long-term paths, which will be mixed in the output layers. Currently, our approach is experimented on Human3.6M, which demonstrate a better result on prediction. The code is available at: https://github.com/Kiramei/DAFCN".

Keywords: Deep Learning; Human Motion Prediction; Pattern Recognition; Fourier Analysis; Graph Convolution Network; Attention Mechanism

Introduction

In recent years, the field of human motion prediction has garnered significant attention due to its wide range of applications, including motion detection, human-robot interaction[1], and motion tracking[2]. Deep learning techniques, particularly those employed in the domain of Computer Vision, have been extensively applied in this area of research.

The prediction of human motion based on skeleton data presents a complex challenge that can be deconstructed into several integral components. The spatial information is encapsulated within skeleton points, which encode the three-dimensional axis information of the human body. Concurrently, the temporal dynamics of the motion are captured through motion sequences, which provide a descriptive representation of the movement. Notably, conventional methods utilizing Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs)[3] have demonstrated considerable success in temporal prediction tasks. Building upon these foundations, researchers have recognized the potential of Graph Convolutional Networks (GCN) in human motion prediction[4], as GCN facilitates effective connectivity between corresponding skeleton points. Consequently, a multitude of GCN variants have emerged, offering improved capabilities in abstracting temporal and spatial features. Additionally, researchers have explored the prediction of human motion in the frequency domain, leveraging methods based on Discrete Cosine Transform (DCT)[4] to transform temporal information into frequency-based representations. Furthermore, the utilization of attention models[5], renowned for their ability to discern salient information through a key-value-

query mechanism, has provided notable advancements in the field of human motion prediction.

Despite these notable advancements, existing approaches have not fully explored the potential of GCN in the frequency domain, nor have they considered the modification of the merging layer to address potential limitations in feature concatenation and concentration. Achieving a delicate balance in the extraction of temporal and spatial information is of paramount importance for both long-term and short-term motion prediction tasks.

To address these challenges, we propose a novel model called DAFCN, which exhibits enhanced performance in both long-term and short-term motion prediction. Extensive evaluations conducted on the widely used Human3.6M[6] dataset have demonstrated its superior predictive capabilities, characterized by reduced prediction errors and improved accuracy in capturing the intricate dynamics of human motion. The DAFCN model embodies a comprehensive understanding of the interplay between temporal and spatial information, providing a promising avenue for advancing the field of human motion prediction.

Related Work

Recurrent neural network

RNNs have demonstrated remarkable success in sequence-to-sequence prediction tasks, thus finding widespread application in the domain of human motion prediction. For instance, Fragkiadaki, et al.[7] proposed the Encoder-Recurrent-Decoder (ERD) model, which incorporates a non-linear multi-layer feed-forward network for encoding and decoding motion before and after the recurrent layers. To mitigate error accumulation, curriculum learning was employed during training. In a similar vein, Jain, et al.[8] introduced the Structural RNN model, which relies on a manually-designed spatio-temporal graph to encode motion history. However, the fixed structure of this graph limits the model's flexibility in capturing long-range spatial relationships between different limbs. To enhance motion estimation, Toshev, et al.[9] presented a residual-based model that predicts velocities instead of poses. Interestingly, it was observed that a simple zero-velocity baseline, involving constantly predicting the last observed pose, outperformed prior approaches. While this yielded improved performance compared to previous pose-based methods, the predictions generated by the RNN still exhibited discontinuities between observed and predicted poses. To address this issue, Ruiz, et al.[10] treated human motion prediction as a tensor inpainting problem and utilized a generative adversarial network for long-term prediction. However, the use of an adversarial classifier complicates training, posing challenges for its deployment on new datasets.

The development of deep neural networks has ushered in exciting advancements in motion prediction. Several studies have employed RNNs to model the temporal correlations of human motion. However, these frame-by-frame methods exhibit limitations in long-term motion prediction due to the inherent problem of error accumulation, while RNN-based networks suffer from first-frame discontinuity. In response, researchers have endeavored to improve prediction results by employing sequence-to-sequence residual models, generative adversarial learning, and imitation learning. In contrast to the frame-by-frame framework, sequence-to-sequence methods effectively mitigate cumulative errors in long-term prediction. Notably, convolution-based approaches treat the historical sequence as a whole and extract motion features in spatial or temporal dimensions, while attention-based mechanisms utilize attention models to capture joint-to-joint and frame-to-frame dependencies.

Graph convolution network

Graph Convolutional Networks are a type of deep learning model specifically designed for graph-structured data. They extend traditional convolutional neural networks (CNNs) to handle irregular and non-Euclidean data domains, such as social networks, molecular structures, and recommendation systems.

The key idea behind GCNs is to generalize the convolution operation from regular grids (e.g., images) to graph structures. GCNs leverage the neighborhood information of each node in the graph to perform node feature aggregation and extraction. By iteratively aggregating features from neighboring nodes, GCNs can capture local and global structural patterns in the graph.

Kipf and Welling[1] proposed a scalable and efficient GCN model for semi-supervised learning on graph-structured data. Their GCN formulation is based on the concept of graph Laplacian, which represents the smoothness of signals on a graph. The model applies a localized first-order approximation of the spectral graph convolution operation, which enables efficient training and inference. The authors demonstrated the effectiveness of GCNs on various benchmark datasets for node classification tasks.

Since its introduction, GCN has become a popular research topic in the field of graph representation learning and has inspired numerous extensions and applications. Several variants of GCNs have been proposed to address different challenges, such as graph pooling, handling directed or attributed graphs, and incorporating temporal dynamics.

Discrete cosine transforms

The Discrete Cosine Transform (DCT) is a mathematical technique used to convert a signal or data sequence from the time or spatial domain into the frequency domain. It is widely employed in signal processing, particularly in image and video compression.

The DCT operates by representing a signal as a linear combination of cosine functions with different frequencies. Through this transformation, the DCT extracts the frequency components present in the signal. The resulting DCT coefficients represent the signal's energy distribution across different frequencies.

In the context of image and video compression, the DCT is applied to blocks of pixel values. By dividing an image or video frame into smaller blocks and applying the DCT to each block, the temporal information within the blocks is transformed into frequency components. The DCT coefficients capture the relative importance of different frequencies in each block.

The energy compaction property of the DCT allows for efficient representation of the signal by concentrating most of the signal's energy in a small number of significant DCT coefficients[11]. This property makes the DCT well-suited for compression purposes, as the less significant coefficients can be quantized or discarded without significant loss in perceptual quality.

Attention

Attention is a mechanism widely used in deep learning models to selectively focus on specific parts of an input sequence, allowing for enhanced information processing and abstraction. It has been successfully applied in various fields, including natural language processing, computer vision, and human motion prediction.

At its core, attention involves assigning importance weights or scores to different elements within an input sequence based on their relevance or significance to the current context. These elements can be individual tokens in a sentence, pixels in an image, or joints in a motion sequence. The attention mechanism allows the model to dynamically calculate these weights and use them to selectively attend to the most relevant elements.

Attention mechanisms have proven to be valuable in the field of human motion prediction, enabling models to effectively capture relevant temporal dependencies and abstract valuable information from complex motion sequences. By incorporating attention into human motion prediction models, these models can selectively focus on crucial joints or frames, enhancing their ability to predict future human movements accurately[12].

In human motion prediction, attention mechanisms are typically employed to model the dependencies between different joints or frames within a motion sequence. By assigning attention weights to each joint or frame based on its relevance to the prediction task, attention mechanisms enable the model to emphasize the most informative elements while downplaying the less relevant ones.

The mechanism works by calculating attention weights that reflect the importance of each joint or frame in the context of predicting future motion. These weights are then used to aggregate the information from different joints or frames, allowing the model to attend to the most salient parts of the motion sequence for accurate prediction.

Fast Fourier convolution

Deep neural networks have driven significant advancements in various research domains, including human motion prediction. Shchekotov I et al.[13] presents a novel convolutional unit, named fast Fourier convolution (FFC), specifically designed to enhance human motion prediction in deep neural networks. The FFC unit addresses two critical requirements: the utilization of large receptive fields and the fusion of multi-scale information.

Receptive fields play a crucial role in capturing spatial dependencies within motion sequences. While many networks employ stacked convolutions with small receptive fields, context-sensitive tasks like human pose estimation benefit from larger receptive fields. To efficiently implement non-local receptive fields and fuse multi-scale information, the paper leverages the spectral transform theory, employing the Fourier transform as the basis for FFC[14].

The FFC unit consists of operations with varying receptive fields, including non-local operations achieved through Fourier transform. These operations are applied to disjoint subsets of feature channels, and the resulting feature maps are aggregated to generate the final output. Importantly, FFC can seamlessly replace vanilla convolutions in existing mainstream CNN architectures without additional computational burden.

Experimental evaluations on human motion prediction tasks, such as action recognition and human key point detection, demonstrate the superior performance of FFC compared to previous models. FFC achieves enhanced prediction accuracy while maintaining computational efficiency comparable to vanilla convolutions. The results highlight the potential of FFC to significantly advance the field of human motion prediction by incorporating non-local receptive fields into deep neural networks.

Method

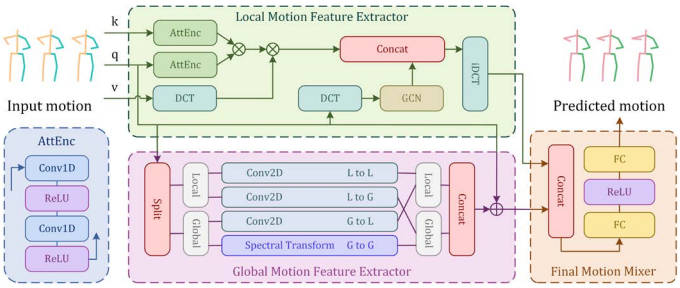

Figure 1: The whole structure of DAFCN: Proposed model contains two main feature extractors for grab the motion features, which are the Local Motion Feature Extractor (LMFE) and the Global Motion Feature Extractor (GMFE), where in the method description, part of the former one is denoted as Attention model and part of the latter one is denoted as FFC. LMFE adopts the attention mechanism for temporal feature extraction and GCN for spatial feature extraction, while DCT enhanced the perception of the model in the frequency field. On the other hand, GMFE decomposed the input motion into local part and global part, where the core lies in the usage of Spectral Transform, which let FFC takes effect to obtain more motion feature in the frequency field. Finally, our method adopts a simple MLP layer for motion output, which is demonstrated as Final Motion Mixer.

Building upon previous research, our objective is to enhance the integration of short-term and long-term prediction. To achieve this, we have devised a method that involves dividing the source into two distinct parts, which is portraited in **(Figure 1)**. The first part utilizes an Attention model to extract valuable information from the preceding human motion sequence. Conversely, the second part employs a FFC approach to capture global and indistinct information.

In our approach, we incorporate window convolution, where the given sequence length is denoted as N, and the length of the sequence to be predicted is represented as T. we set the window size as $T \times 2$, resulting in the acquisition of $N - T \times 2 + 1$, windows for prediction purposes.

To enhance the predictive capabilities of our method, we integrate the DCT. By transforming the axis information into the frequency domain, the GCN becomes more adept at abstracting the bone structure graph, thereby yielding improved prediction outcomes.

Furthermore, we emphasize the significance of FFC as a robust long-term informer. Previous studies have highlighted the cross-multiplication mechanism inherent in FFC, which leverages Fourier transform on a single channel. This novel approach enables the extraction of global information that contributes to enhanced prediction accuracy.

Once all the individual components are completed, they are concatenated and fed into a MLP layer. Employing an MLP

layer for data fusion proves to be an excellent choice, given its remarkable ability to integrate diverse data sources effectively.

In summary, our proposed methodology combines an Attention model for short-term information extraction, FFC for long-term information capture, window convolution for prediction across multiple windows, DCT for transforming axis information into the frequency domain, GCN for abstracting bone structure graphs, and an MLP layer for comprehensive data mixing. By integrating these components, our approach aims to improve the accuracy and robustness of human motion prediction.

Experiment

Human3.6M is used as our training and testing dataset, which is a widely used benchmark dataset for motion prediction. It consists of recordings of seven actors performing 15 actions, with each human pose represented as a 32-joint skeleton. The 3D coordinates of the joints are computed using forward kinematics on a standard skeleton.

To align with previous works, we remove the global rotation, translation, constant angles, and 3D coordinates of each human pose. The motion sequences are down-sampled to 25 frames per second. Our method is evaluated on subject 5 (S5) from the Human 3.6M dataset. Instead of testing on a limited number of random sub-sequences per action, as shown to introduce high variance, we report results on 256 sub-sequences per action when using 3D coordinates.

Let us now introduce the loss functions we use to train our model on either 3D coordinates or joint angles. For 3D joint coordinates prediction, we make use of the Mean Per Joint Position Error (MPJPE) proposed in Ionescu C et al. 's work[15]. In particular, for one training sample, this yields the loss

$$\mathcal{L} = \frac{1}{J(M+T)} \sum_{t=1}^{M+T} \sum_{j=1}^{J} \|\hat{p}_{t,j} - p_{t,j}\|^2,$$

where $\hat{p}_{t,j} \in \mathbb{R}^3$ represents the 3D coordinates of the j_{th} joint of the t_{th} human pose in $X^{\mathcal{N}-M+1:\mathcal{N}+T}$, and $p_{t,j} \in \mathbb{R}^3$ is the corresponding ground truth[16].

ℓ_1

$$\mathcal{L} = \frac{1}{K(M+T)} \sum_{t=1}^{M+T} \sum_{k=1}^{K} |\hat{x}_{t,k} - x_{t,k}|,$$

$\hat{x}_{t,k}$ $X^{\mathcal{N}-M+1:\mathcal{N}+T}$ **(Table 1)**.

Table 1: MPJPE Comparison of Prediction Result on H3.6M.

Model	80ms	160ms	320ms	400ms	560ms	720ms	880ms	1000ms
conv Seq2Seq	13.5	27	52	63.1	82.4	98.8	112.4	120.7
HRI	10.4	22.6	47.1	58.3	73	91.8	101.1	112
DAFCN	9.7	22.1	46.8	57.9	77.3	91.5	103.6	111.3

In accordance with the configuration used in our reference baselines, which are convSeq2Seq and HRI, we present our results in Table 1 for both short-term prediction (less than 500ms) and long-term prediction (greater than 500ms). For the Human3.6M dataset, our model is trained by utilizing the previous 50 frames to forecast the subsequent 10 frames. To generate poses further into the future, we recursively employ the predicted poses as input to the model.

Based on the results presented in Table I, it is evident that our model consistently achieves superior performance compared to the two aforementioned models in both short-term and long-term prediction tasks. The results clearly demonstrate the effectiveness and superiority of our model over the competing approaches.

Conclusion

In this paper, we propose an approach called DAFCN for human motion prediction addresses the limitations of existing methods in long-term predictions by utilizing Fast Fourier Convolution and incorporating the Motion Attention Model to capture short-term relevant information. The hybrid prediction strategy, which combines short-term and long-term paths in the output layers, further enhances performance. Experimental evaluations on the Human3.6M dataset demonstrate the superiority of our model in both short-term and long-term prediction tasks, offering valuable insights for advancing the field of human motion prediction and enabling more accurate predictions in real-world scenarios.

References

1. Koppula HS, Saxena A. Anticipating human activities for reactive robotic response. Proc IEEE Int Conf Intell Robots Syst (IROS) 2013:2071.
2. Gong H, Sim J, Likhachev M, Shi J. Multi-hypothesis motion planning for visual object tracking. Proc IEEE Int Conf Comput Vis (ICCV) 2011:619-626.
3. Jain A, Zamir AR, Savarese S, Saxena A. Structural-RNN: Deep learning on spatio-temporal graphs. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR) 2016:5308-5317.
4. Mao W, Liu M, Salzmann M, Li H. Learning trajectory dependencies for human motion prediction. Proc IEEE Int Conf Comput Vis (ICCV) 2019:9489-9497.
5. Mao W, Liu M, Salzmann M. History repeats itself: Human motion prediction via motion attention. Proc Eur Conf Comput Vis (ECCV) 2020:474-489.
6. Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell (TPAMI) 2014;36(7):1325-1339.
7. Fragkiadaki K, Levine S, Felsen P, Malik J. Recurrent network models for human dynamics. Proc IEEE Int Conf Comput Vis (ICCV) 2015:4346-4354.
8. Jain A, Zamir AR, Savarese S, Saxena A. Structural-RNN: Deep learning on spatio-temporal graphs. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR) 2016:5308-5317.
9. Toshev A, Szegedy C. Deep pose: Human pose estimation via deep neural networks. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR) 2014:1653-1660.
10. Ghosh P, Song J, Aksan E, Hilliges O. Learning human motion models for long-term predictions. Proc Int Conf 3D Vision (3DV) 2017:458-466.
11. Ahmad T, Iqbal M, Rahman S. Using discrete cosine transform based features for human action recognition. J Image Graph 2015;3:96-101.
12. Chu X, Yang W, Ouyang W, Wang X. Multi-context attention for human pose estimation. Proc IEEE Conf Comput Vis Pattern Recognit 2017:1831-1840.
13. Shchekotov I, Komkov S, Pavlov D, et al. FFC-SE: Fast Fourier convolution for speech enhancement 2022.
14. Chi L, Jiang B, Mu Y. Fast Fourier convolution. Adv Neural Inf Process Syst 2020;33:4479-4488.

15. Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell (TPAMI) 2014;36(7):1325-1339.

16. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks 2016.