Modelling of Automated Computational Model towards Mapping Natural Text Description to Realistic Human Motion Video Description

Shradha Kannan*, Ms. Vathsala M K**

**Student, Department of Information Science and Engineering, Ramaiah Institute of Technology, Bangalore, India*
***Assistant Professor, Department of Information Science and Engineering, Ramaiah Institute of Technology, Bangalore, India*

Abstract - *Automation of every process has become a technological trend with high computing platforms and advance learning models, and many industries and tech companies are ready to turn their complex and traditional processes towards automation. This paper has presented the modeling of the multi-domain mapping model as a significant contribution in the field of creative applications and the natural language processing research domain. However, multi-domain processing and feature space mapping is quite a challenging task in a field of computer graphics. The proposed study constructed an automation model, which takes a natural text from the user as input and constructs the realistic visual scene of human motion activities according to the given condition in the input data. The proposed automation model deals with the distinctive features of the data set in different formation of files, which contains information about human motion activities. The proposed system is implemented on a numerical computing platform. The study outcome demonstrates the performance of the proposed system with an accuracy rate of 89% in the matching of outcome with the condition given in natural text description. The proposed study also generates high quality realistic visual graphics of human motion activity.*

Key Words: NLP, Nueral Network, LSTM, GRU, Visual Graphic Animation, Human Motion Activity

1. INTRODUCTION

Advances in Machine learning (ML), artificial intelligence (AI), and higher computing platforms have unlocked new possibilities in every sector of human life [1-2]. ML and AI are areas of digital and intelligence technology that involve innovative ideas, high capability of processing both linear and nonlinear data, and automating the time consuming and complicated tasks [3]. ML and AI have become crucial for data-driven applications due to their incredible ability to capture hidden patterns, knowledge mining, and optimization in business and industrial processes. Therefore, these technologies are essential units of automation because they can improvise the functionality of traditional business processes through modern learning algorithms, thereby increasing productivity and overall outcome, which is desirable [4]. In recent years, deep learning methods have become the most popular in the field of automation because of their self-learning and robust prediction ability. In the field of creative and captioning applications, recent developments in computers and maturity in and hardware tools are promoting the intelligence system. Therefore, many industries and business companies are interested in adopting ML and AI in their workplaces for automating their tasks. Advanced learning models and architectures have also opened up in the context of natural language processing (NLP) [5-6]. Much research has been done towards NLP in the application of text generation or text captioning from images and videos based on the learning approach and recurrent networks. The mapping of video to text description has been studied extensively over the years. However, text-based video generation has not been extensively studied. Previous works in the field of NLP has much focused on generating text subtitles from the video [7-8]. However, the mapping features space in the domain of producing visual graphics from the natural description is quite a challenging issue for most of the existing methods. An important attention in visual graphic construction from the linguistic description is that both image and moving animation must be determined according to the condition described in the input natural text. The adoption of natural description to image construction mechanisms directly creates a situation in which moving animation is not uttered by the natural description. In this context, the problem of moving animation construction is mainly associated with to approach of prediction. In the prediction mechanism, the prime objective is to learn a nonlinear feature of the given frames to guess the best possible or subsequent frames [8]. Various other research studies have also presented their approach for mapping natural text description feature space to visual motion graphic feature space [9-10]. However, many issues, such as computational complexity, low-resolution, accuracy, are associated with generated output in the existing system.

The proposed study attempts to explore the applicability of ML concepts in the research domain of realistic human motion animation generation from the natural text sequences. The study introduces the modeling of a computational-efficient human motion prediction model to generate the most relevant visual scenes. The study adopts human motion activity dataset-(MOCAP) provides by KIT. The study also models two different neural networks i.e., LSTM and GRU, which specific forms of Recurrent neural network for the mapping of natural text given by the user as input and model construct relevant, realistic visual animation of human movement activity. The scope of the study justifies with outcome and model performance analysis.

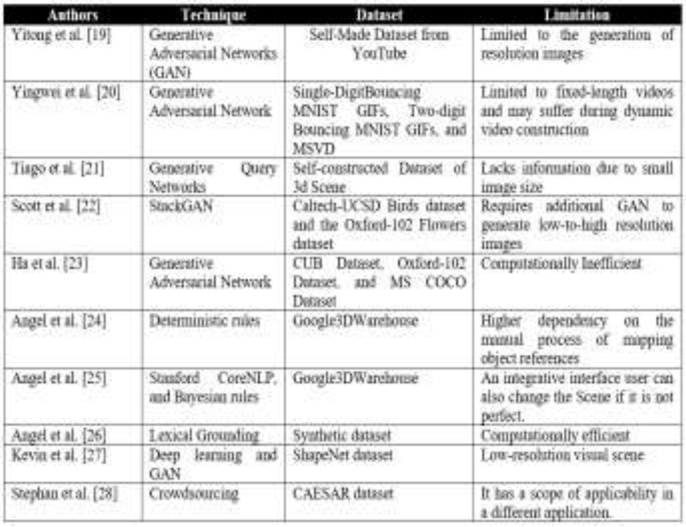
The rest of the section of this paper is planned in the following manner. Review of existing literatures is conducted in section II. Description and design of proposed system is discussed in section III. Implementation strategies followed by algorithm steps is discussed in section IV. Result analysis is presented in section V. Finally, overall contribution of proposed study is concluded in section VI.

2. RELATED WORK

Several research works have been carried towards addressing problems associated with the NLP domain for text generation form the video description. However, research work towards video generation from natural language description is less studied and explored. The research work in a similar direction is considered in the study of Pan et al. [11]. In this, the authors introduce the Hierarchical Recurrent Neural-(HRN) encoder focusing on the learning of temporal details in the video description for text representation. Another work for video captioning is considered by Venugopalan et al. [12], where the authors have focused on the problem of producing text in open domain videos and presented sequence-to-sequence-(S2S) framework using LSTM to construct natural description about the events in the input visual scene. A different approach for paragraph captioning is conducted in the research work of Yu et al. [13]. Here, the authors used the HRN Network and attention mechanism to consider both temporal and spatial relationships between video descriptions to generate single-line text for the short event in video and multi-line paragraph description for a long event. A recent work towards video generation from natural text description is considered by Plappert et al. [14], where the authors have introduced a bidirectional mapping model based on the RNN model and Sequential learning mechanism for text feature space to human motion visual scene description. A multi RNN autoencoder based Bidirectional mapping model is also presented in the work of Yamada et al. [15] for visualization of robotic motion activities from the linguistic descriptions.

Ballas et al. [16] provided a prediction model based on GRU for Spatio-temporal feature learning for the video description generation. An instigation study is carried out by Austin et al. [17]. In this, the authors have investigated how large feature mining from large text descriptions can improvise the performance of video generation in LSTM based prediction model. An approach based on a combined multilayered neural network and dropout autoencoder can be seen in the study of Ghosh et al. [18] for Spatio-temporal modeling for synthesizing for realistic motion sequences for long-term predictions. Various other research works [19-28] are carried in the same direction as the research problem as highlighted in table 1.

Table 1 Summary of related work



Authors	Technique	Dataset	Limitation
Yitong et al. [19]	Generative Adversarial Networks (GAN)	Self-Made Dataset from YouTube	Limited to the generation of resolution images
Yingwei et al. [20]	Generative Adversarial Network	Single-DigitBouncing MNIST GIFs, Two-digit Bouncing MNIST GIFs, and MSVD	Limited to fixed-length videos and may suffer during dynamic video construction
Tiago et al. [21]	Generative Query Networks	Self-constructed Dataset of 3d Scene	Lacks information due to small image size
Scott et al. [22]	StackGAN	Caltech-UCSD Birds dataset and the Oxford-102 Flowers dataset	Requires additional GAN to generate low-to-high resolution images
Ha et al. [23]	Generative Adversarial Network	CUB Dataset, Oxford-102 Dataset, and MS COCO Dataset	Computationally Inefficient
Angel et al. [24]	Deterministic rules	Google3DWarehouse	Higher dependency on the manual process of mapping object references
Angel et al. [25]	Stanford CoreNLP, and Bayesian rules	Google3DWarehouse	An integrative interface user can also change the Scene if it is not perfect.
Angel et al. [26]	Lexical Grounding	Synthetic dataset	Computationally efficient
Kevin et al. [27]	Deep learning and GAN	ShapeNet dataset	Low-resolution visual scene
Stephan et al. [28]	Crowdsourcing	CAESAR dataset	It has a scope of applicability in a different application.

3. PROPOSED SYSTEM

The proposed system introduces modeling of an automated system for mapping of natural language description to realistic visual animation generation. The design of the proposed system is implemented analytically that takes natural text features from users, and after execution via neural networks, it provides outcome as motion vector feature, which after processing with the visual graphic application, generates final outcome as realistic human motion activity visual animation. The proposed system contains several functional and operation blocks followed by integration of two neural networks viz. i) LSTM-() and GRU-(). The modeling of the proposed system is depicted in figure 1. The prime objective of the proposed automated system is to predict human motion activity from text-conditioned given by the user. The proposed system contains four vital components, namely MOCAP Dataset, NLP- (Natural Language Processing module), Sequence2Sequence (LSTM and GRU), and Physic Processing. The human motion activity dataset is considered in the proposed study to train both neural networks to learn text features and motion features. Here, the training operation is carried out independently for both neural networks to perform sequence classification and feature learning to generate a motion vector. After, execution of training operation, both the neural network is then integrated with another feature processing module called physics processing, which is mainly subjected to generate full-motion vector based on the understanding of linguistic knowledge.

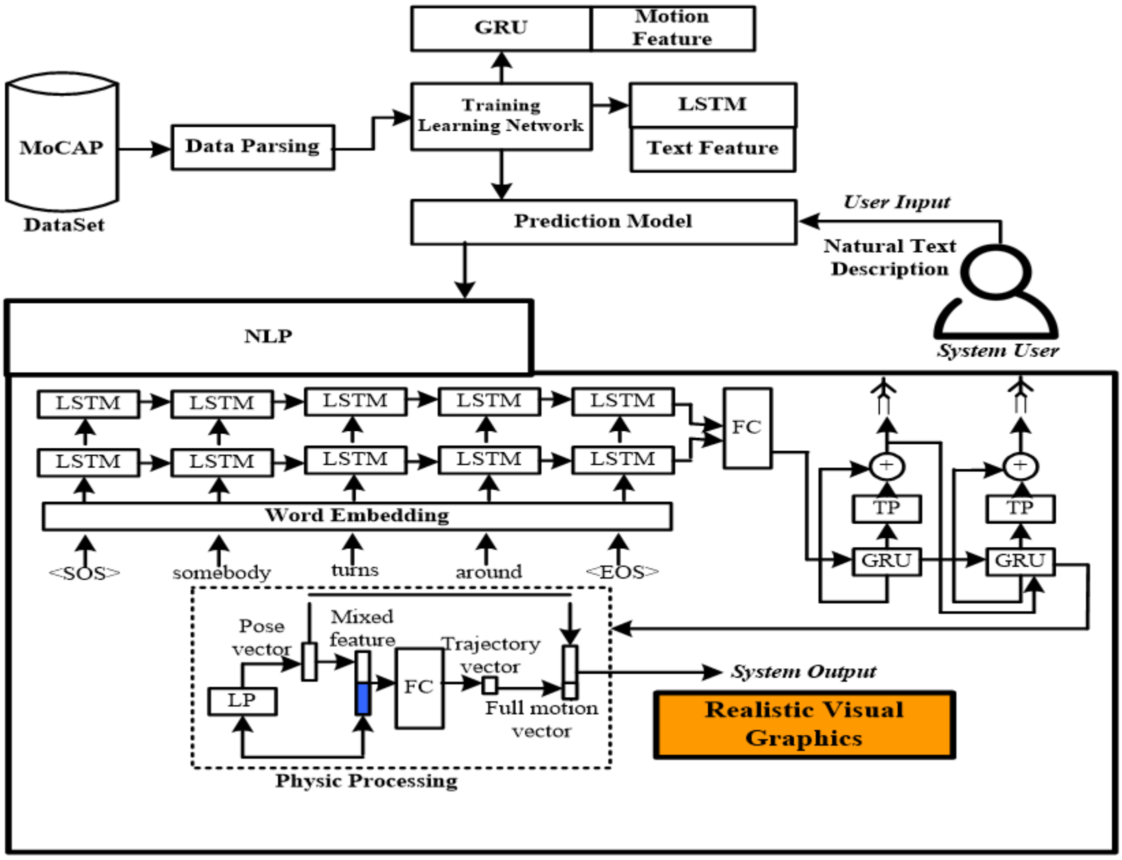


Figure 1 Architecture of proposed automated system

The system user provides an input to the proposed system in the form of natural text. Since the computational machine cannot execute or read the natural language directly, therefore, the input text given by the system user is subjected to the work embedding process. The word embedding model uses a lexicon-based word2vector filter, which converts natural language description into word vector-(machine-readable and understandable format). The system then performs a sequence classification of linguistic description from the word vector using NLP and LSTM. In this, initially, the NLP unit learns the significant features i.e., verb and adjective, from the linguistic description, which further subjected to LSTM processing. The LSTM component then performs sequence classification and extracts semantic features in the form of frame feature and language feature. The output obtained from LSTM with frame data then goes to another learning module GRU as output, which, after processing, provides a motion feature. In this, the GRU performs processing of data obtained from LSTM and regulates it by passing on information as it propagates forward through the sequence chain. After successful execution, the GRU then generates a motion feature based on the spatial relationship between frame feature and text feature given by the user. Both LSTM and GRU have memory and neural network layers, also known as gates. These layers basically adjust the data information in the sequence processing flow to determine the information related to adding the next content. The LSTM unit predicts the behavior of the training set by identifying its pattern. After the processing is completed, it classifies an invisible sample and predicts the possibility of outputting the best possible next word. Then, the results obtained from the LSTM module are then subjected to GRU, which processes the text features and frame features and generates motion features. The motion feature obtained from the sequence processing and classification module (LSTM and GRU) are then processed with an important module called physics processing. In this, the obtained motion feature is then mapped with pose trajectories and other mixed features like speed and position of nodes over time to generate motion vector, which after rendering with the visual graphic application, constructs realistic animation of human motion activity. Description of dataset adopted in the proposed system is given as follows:

Dataset: The study considers MOCAP-(motion capture) collected from the internet source provided by the KIT. The MOCAP dataset is a collection of large human motion activity data which was captured using experimental set-up using different sensors fixed on the human body. The proposed system first preprocesses the dataset to extract and read significant files having different formats. Each file is having an important role in the generation of motion vector as it contains signification features related to human pose and motion activity. The details of the dataset file and its varying formats are illustrated in figure 2

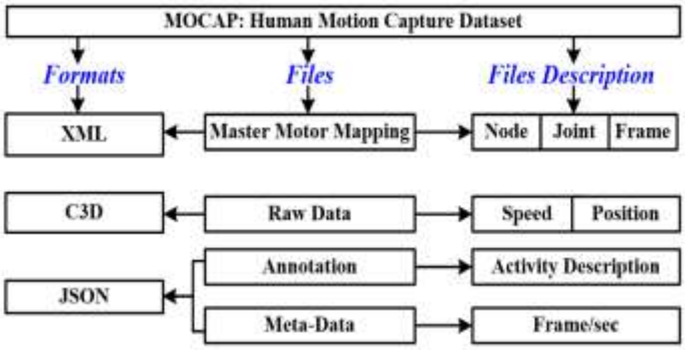


Figure 2 description of dataset adopted in the study

The MOCAP dataset holds four different files, namely MMM-(Master motor Mapping file), RawData, Annotation, and Metadata, with three different file formats. Here each file in the dataset is with different formats, namely XML, C3D, and JSON, respectively. The MMM file holds information about human activity features like node, joint, and frame. The node is basically a point in three-dimensional space, and the joint is a connecting line between the nodes. The frame is information that holds changing locations of the nodes. The raw data file in the C3D format holds information associated with raw features like speed and position of motion activity. The annotation file in JSON format keeps the information about the description of motion activity, and another file MetaData in the same file format contains information about the variable motion activity position over time i.e., Frame per second.

i) NLP (Natural Language Processing)

NLP is a field of computational intelligence that enables machines to read and understand the natural text by performing a semantic analysis that captures word features based on spatial relationships between the original texts.

ii) Word Embedding Model

The neural network-based prediction model cannot understand the natural language described in a human-readable format. Therefore, the word-embedding model is used in the initial layer of neural networks to convert natural text into a machine-readable format. Here, the word2vector mechanism is used in the word embedding process. It is a united set of language feature learning and linguistic modeling in which words in the vocabulary are mapped to real-number vectors. Here, the words in the input text will be converted nonzero and non-fuzzy real number vectors, which will be the neural network's input for further processing.

iii) LSTM- (Long Short-Term Memory)

LSTM is a special form of Recurrent Neural Network (RNN) that processes natural text and keeps forwarding relevant information in order of sequence. LSTM has a cell, which acts as a memory and carries relevant information throughout the sequence processing. LSTM has another unit called a gate, which is of multiple types of viz. i) forget gate, ii) input gate, and iii) output gate. The forward gate agrees to retain the information relevant to the information in earlier steps. The input gate decides which kind of information to be added from the recent steps. The output gate is responsible for checking or estimating the next probable hidden state.

iv) GRU (Gated Recurrent Unit)

The GRU is also a specific form of RNN, but it is more advanced and robust than LSTM. The GRU has two gates, namely i) reset gate and ii). Update gate. The function of the update gate is similar to the forget gate of the LSTM model and the input gate of the LSTM model. It ensures which information needs to be discarded and which information needs to be considered. The reset gate determines what existing information needs to be overlooked. GRU also has the functionality of tensor operations, and due to this, training is faster in GRU compared to LSTM.

v) Physic Processing

The physic processing refers to mapping of motion feature and its behavior with other related mixed features to motion vector. In this, a function parse_MMM() function provided by SCImoX library is applied to generate motion vector as final outcome.

vi) Use Case Diagram

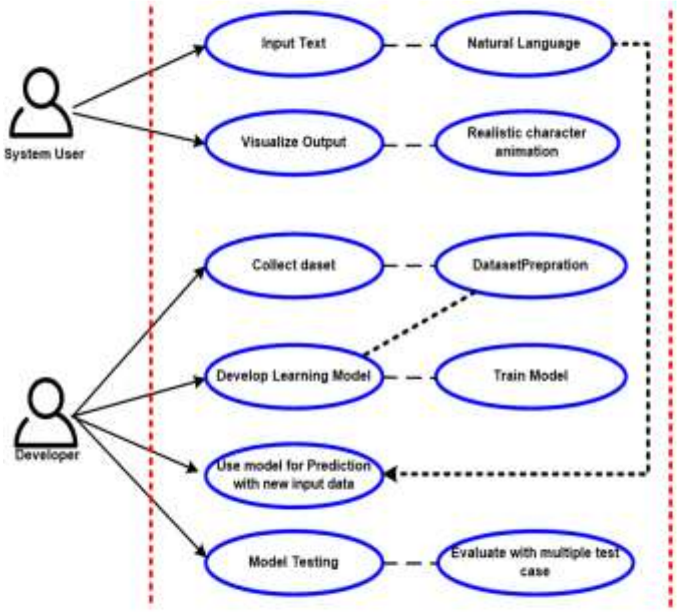


Figure 3 Use Case diagram of proposed model

In figure 3 mentioned above, a set of distinct operations is described, which shows the relationship between the user and the developer and allow users to provide input to the system in natural language and visualize the output in the form of realistic characters in visual graphics. The role of the developer is multiple, from data analysis to model testing. Once the model is trained, predictions are performed using new data provided by the user input as natural text. Developers or testers can use different test cases to test all the functions and modules of the system until they are completely satisfied.

vii) Sequence Diagram

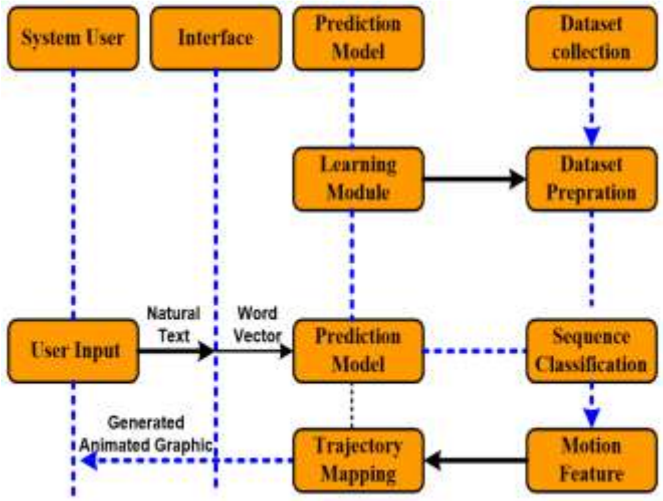


Figure 4 Sequence diagram of proposed system

The sequence diagram in figure 4 exhibits the operation flow in order from one object to another. The system user provides input in the command-line interface and able to get expected output in another interface. Developers collect the dataset and preprocess it for training purposes. The prediction model takes the input from the user and converts it into a word vector using the word embedded model at the bottom layer of the neural network model and performs internal operation the final outcome in the form of visual animation of human motion activity.

4. IMPLEMENTATION STRATEGY

This section discussed the computational strategies adopted in the implementation of the proposed system design. The implementation and design of the proposed system are carried on a numerical computing tool that takes input as natural language description in the format of text given by and after execution; it provides realistic visual graphic animation of human movement activities. The significant steps followed in the implementation design are described below.

Algorithm-1 Natural Description to Visual Graphic

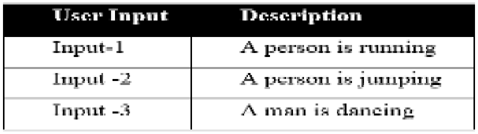
```
Input: Natural Description-(T)
Output: Visual Graphics of human motion Activity-($V_G$)
Start
    1. Initialize T, D-(Dataset)
    2. $D_f \rightarrow f_1(D)$
    3. $D_{set} \rightarrow$ Word2Vector: $f_2$ (D)
        a. NLP: fs(Statement)
        b. Get $\leftarrow$ Verb, Frame
    4. Model $\rightarrow f_3$(LSTM, GRU)
    5. LSTM $\leftarrow f_4(D_{set})$
        a. Input = $f_5$(shape: [1])
        b. L1 = $f_5$(size, connect:$\rightarrow$input)
        c. L2 = $f_5$(size, connect:$\rightarrow$ L1)
        d. Output = $f_5$(shape: [1,3], connect:$\rightarrow$ L2)
    6. GRU $\leftarrow f_4$(LSTM:Output)
        a. Input = $f_5$(shape: [1,3,shape(frames)[0]])
        b. L1 = $f_5$(size, connect:$\rightarrow$input)
        c. L2 = $f_5$(size,connect=L1)
        d. Output = $f_5$ (shape(frames),connect$\rightarrow$L2)
        e. $M_{feat}$ = LSTM(Data),GRU(Data)
    7. Train $\leftarrow f_6$(GRU, LSTM)
    8. Execute Model
        a. $D_{read} \rightarrow f_5(D_{path})$
        b. Input$\leftarrow$T
        c. Word2Vector $\rightarrow f_2$ (T)
        d. NLP $\rightarrow f_s$(T)
        e. LSTM $\rightarrow [T_{feat}, F_{feat}]$
        f. GRU $\rightarrow [M_{feat}]$
        g. $M_{vec} \rightarrow f_7(M_{feat})$
        h. $\mathbf{V_G \leftarrow f_8(M_{vec})}$
End
```

The algorithmic steps mentioned above are responsible for executing a model for mapping natural language description feature space to visual graphic human motion feature space. The execution of the first step initializes the system variable T, D. The variable T is assigned to take input from the system user, and variable D is assigned for the Dataset (line 1). In the second step of the algorithm, the system uses a function *f1*( ) for the preprocessing of the dataset. In this process, all dataset folder-(Df) with varying file formats are extracted and prepared for the proposed model training process (line 2). In the next step of the proposed system algorithm, all the sentences are converted into a vector of a real number using function *f2*( ), which basically refers to word2vector operation. In the same process, NLP is function *fs*( ) called spacy is applied overall word vector to extract text feature in the form of the verb, and at the same time, the frame files from the dataset are also read by the NLP function for the training process (line 3). The next significant step of the algorithm is subjected to the design of neural networks (LSTM and GRU), where function *f3*( ) is defined to parse all necessary packages and modules for the modeling of neural networks (line-4). The execution of the next step in the proposed algorithm is basically associated with modeling of LSTM, where another function *f5*( ) is applied to define layers of LSTM. After then, the modeling of GRU is carried using the same function with different sizes of layers. Here, the data obtained from the output layer of LSTM is further subjected to a new input for the GRU. It should be noted that the model proposed is implemented based on the integration of LSTM and GRU. But before the modeling process is carried out, the system uses a function *f*4( ), which stores all the features of the dataset in a sequential list, which was initially obtained by the NLP function (line 5). In a similar way, the modeling of GRU is also carried using the same function *f5*( ) for defining neural network layers (line 6). Then the system calls function *f6*( ) for executing GRU and LSTM training operation. It should also be noted that training operation on both the neural network model is given independently (line7). After the successful execution of training operation, the system then executes its next step for model validation and performance analysis (Line 8). In this step, the system takes input from the system user in the form of natural text-(T). The model then converts the natural text description into word vector and obtains significant sematic features from the sentence followed by LSTM based sequence processing and classification, which after execution, generates text features-(Tfeat) and frame feature-(Ffeat). The obtained features are then fed to GRU as input, which after processing, provides motion feature-(Mfea). The motion features are then fed to another module for physic processing. In this, the system uses third party function $f7(\ )$ called as parse_MMM and provides a motion vector(Mvec) which after processing with visual graphic application generates realistic visual graphics of human motion activity as an end result, which is according to the condition described in text description given by the system user.

5. RESULT ANALYSIS

This section discusses the outcome obtained and performance of the proposed system. The outcome is analyzed with different input text provided by the system user as shown in table 2.

Table 2 Description of User Input



User Input	Description
Input-1	A person is running
Input -2	A person is jumping
Input -3	A man is dancing

The visual outcome based on above inputs is depicted in following figures as follows:

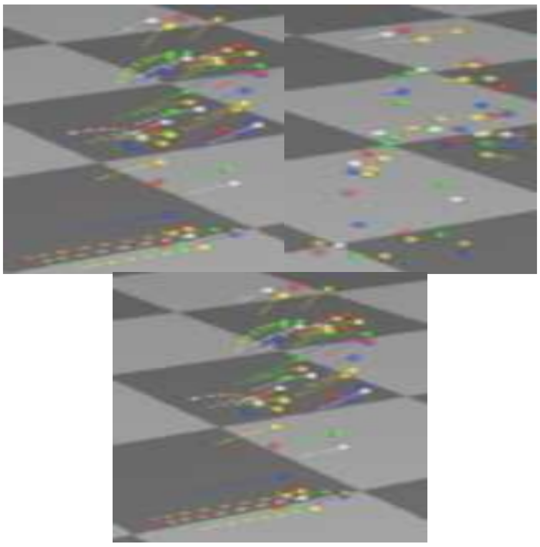


Figure 5 Visual outcome for input: a person is running

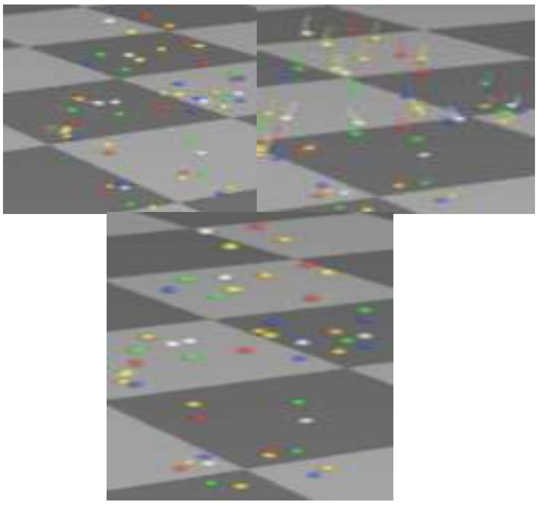


Figure 6 Visual outcome for input: a person is jumping

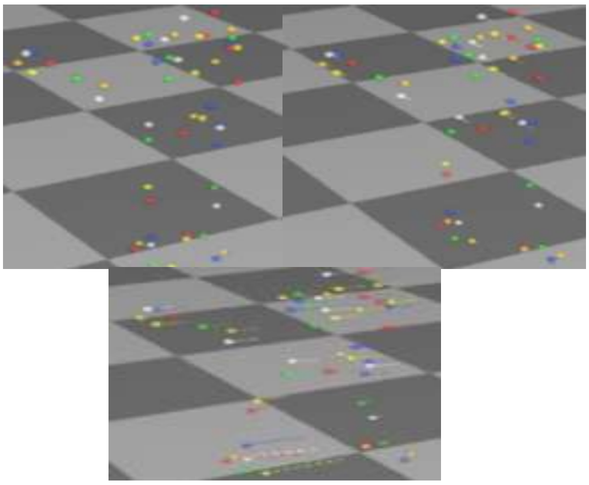
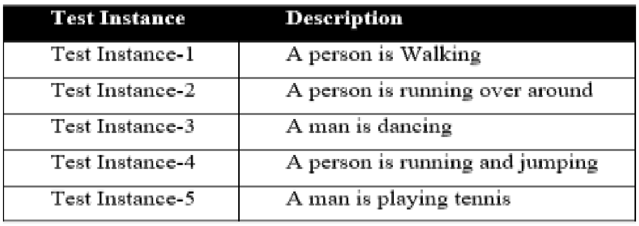


Figure 7 Visual outcome for input: a person is dancing

The visual outcome depicted in this paper may be associated with blurriness because the proposed system generates a running video sequence of human motion activity, and the figures demonstrated here are just snippets of the outcome obtained. In addition, the proposed system provides video sequences in the resolution dimension. Therefore, the introduced model archives better performance in terms of visual quality. The proposed system also assessed in terms of prediction accuracy, where five different test instances are analyzed to check the closer correlation among the generated outcome and condition described in the user input natural text. Table 3 exhibits the details of different test instances carried for performance evaluations.

Table 3 Description of Test Instances



Test Instance	Description
Test Instance-1	A person is Walking
Test Instance-2	A person is running over around
Test Instance-3	A man is dancing
Test Instance-4	A person is running and jumping
Test Instance-5	A man is playing tennis

Table 3 highlights the description of different test instances with different user input text description.

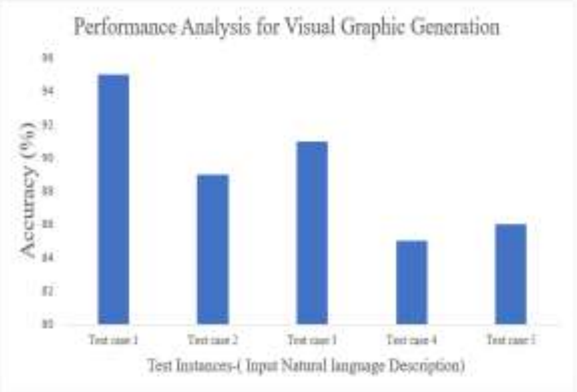


Figure 8 Analysis of Prediction model with different test instance of user Input

The above figure 8 exhibits the performance of proposed prediction model in terms of closer mapping of natural description and visual graphic outcome. The analysis shows that the proposed system has achieved approximate 89% of accuracy rate in mapping of natural text to animation generation.

6. CONCLUSION

This paper has proposed a multi-domain computational model for mapping natural text feature space to visual animation feature space. The modeling of the proposed model is basically performed using two neural network models, which are a specific form of RNN. Human motion activity dataset-(MOCAP) for feature learning and both models are trained independently. The outcome revealed that the proposed model provides high-resolution animation graphics with less processing time. The proposed work presents a significant contribution to the computer graphics research field generated from the natural language description. In the future work, the proposed system can be further improvised or integrated with additional functionality where the system takes input directly from the human voice to generate visual mobile animation based on the given condition in the voice description.

REFERENCES

[1] Wang, Mowei, Yong Cui, Xin Wang, Shihan Xiao, and Junchen Jiang. "Machine learning for networking: Workflow, advances and opportunities." *Ieee Network* 32, no. 2 (2017): 92-99.

[2] A. A. Gebremariam, M. Usman and M. Qaraqe, "Applications of Artificial Intelligence and Machine Learning in the Area of SDN and NFV: A Survey," 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), Istanbul, Turkey, 2019, pp. 545-549, doi: 10.1109/SSD.2019.8893244.

[3] L. Tuggener et al., "Automated Machine Learning in Practice: State of the Art and Recent Results," 2019 6th Swiss Conference on Data Science (SDS), Bern, Switzerland, 2019, pp. 31-36, doi: 10.1109/SDS.2019.00-11.

[4] F. Musumeci et al., "An Overview on Application of Machine Learning Techniques in Optical Networks," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383-1408, Second quarter 2019, doi: 10.1109/COMST.2018.2880039.

[5] Wei, Wei, Jinsong Wu, and Chunsheng Zhu. "Special issue on deep learning for natural language processing." (2020): 1-3.

[6] Fiorucci, Marco, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. "Machine Learning for Cultural Heritage: A Survey." Pattern Recognition Letters 133 (2020): 102-108.

[7] Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell,T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In IEEE CVPR

[8] Vondrick, C., and Torralba, A. 2017. Generating the future with adversarial transformers. In CVPR

[9] Habibie, Ikhsanul, Daniel Holden, Jonathan Schwarz, Joe Yearsley, Taku Komura, Jun Saito, Ikuo Kusajima, Xi Zhao, Myung-Geol Choi, and Ruizhen Hu. "A Recurrent Variational Autoencoder for Human Motion Synthesis." In BMVC. 2017.

[10] Holden, Daniel, Jun Saito, and Taku Komura. "A deep learning framework for character motion synthesis and editing." ACM Transactions on Graphics (TOG) 35, no. 4 (2016): 1-11.

[11] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1029–1038, 2016.

[12] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor
Darrell, and Kate Saenko. Sequence to Sequence–Video to Text. In Proceedings of the IEEE
International Conference on Computer Vision, pages 4534–4542, 2015.

[13] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video Paragraph Captioning using Hierarchical Recurrent Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4584–4593, 2016

[14] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a Bidirectional Mapping Between Human Whole-Body Motion and Natural Language using Deep Recurrent Neural Networks. Robotics and Autonomous Systems, 109:13–26, 2018.

[15] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. Paired Recurrent Autoencoders for Bidirectional Translation Between Robot Actions and Linguistic Descriptions. IEEE Robotics and Automation Letters, 3(4):3441–3448, 2018.

[16] Subhashini Venugopalan, Lisa Anne Hendricks, Raymond Mooney, and Kate Saenko. Improving LSTM-based Video Description with Linguistic Knowledge Mined from Text. In

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016.
[17] ÖmerTerlemez, Stefan Ulbrich, Christian Mandery, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Master Motor Map (MMM)—Framework and Toolkit for Capturing, Representing, and Reproducing Human Motion on Humanoid Robots. In 2014 14th IEEE-RAS International Conference on Humanoid Robots (Humanoids), pages 894–901. IEEE, 2014.
[18] Ghosh, Partha, Jie Song, Emre Aksan, and Otmar Hilliges. "Learning human motion models for long-term predictions." In 2017 International Conference on 3D Vision (3DV), pp. 458-466. IEEE, 2017. [19] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video Generation from Text. In Proceedings of AAAI-2018, 2018.
[20] Yingwei Pan, ZhaofanQiu, Ting Yao, Houqiang Li, and Tao Mei. To Create What You Tell: Generating Videos from Captions. In Proceedings of the 2017 ACM on Multimedia Conference, pages 1789–1798. ACM, 2017.
[21] Tiago Ramalho, TomášKocisk`y, Frederic Besse, SM Eslami, GáborMelis, Fabio Viola, Phil Blunsom, and Karl Moritz Hermann. Encoding Spatial Relations from Natural Language. arXiv preprint arXiv:1807.01670, 2018.
[22] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis. In Proceedings of the 33rd International Conference on Machine Learning (ICML), pages 1060–1069, 2016.
[23] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the International Conference on Computer Vision (ICCV), pages 5907–5915, 2017.
[24] Angel Chang, Manolis Savva, and Christopher Manning. Semantic parsing for text to 3d scene generation. In Proceedings of the ACL 2014 Workshop on Semantic Parsing, pages 17–21, 2014.
[25] Angel Chang, Manolis Savva, and Christopher D Manning. Learning spatial knowledge for text to 3d scene generation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2028–2038, 2014.
[26] Angel Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D Manning. Text to 3D Scene Generation with Rich Lexical Grounding. In Proceedings of ACL, 2015.
[27] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2Shape: Generating Shapes from Natural Language by Learning Joint Embeddings. arXiv preprint arXiv:1803.08495, 2018.36
[28] StephanStreuber, M Alejandra Quiros-Ramirez, Matthew Q Hill, Carina A Hahn, Silvia Zuffi, Alice O'Toole, and Michael J Black. Body Talk: Crowdshaping Realistic 3D Avatars with Words. ACM Transactions on Graphics (TOG), 35(4):54, 2016.