Gesture Generation from Urdu Text based on Deep Learning Approach

Hussan Fatima [*], Sammat Fareed[**]

[*] FACULTY OF ENGINEERING & COMPUTING, NATIONAL UNIVERSITY OF MODERN LANGUAGES ISLAMABAD.
[**] Alhamd Islamic University Islamabad Campus

***Abstract*:** When we interact with any machine-like robot or any other machine, gestures are important alongside speech. For years, efforts have been made to give computers human-like capabilities through research, but most gesture generation studies have been constrained by dependencies on speech input, the English Language, and specific speakers. This research aims to address these constraints through creating a gesture-generation model that generates high-quality gestures in return of Urdu textual input without requiring speaker assistance. We have developed our gesture model using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) algorithms, training it on a custom-created dataset. The implementation results demonstrate the model's effectiveness, achieving a Percentage of Corrected Key points (PCK) value of 75% and a mean absolute error rate of 0.3. These results validate that the model is successful in producing reliable gestures for AI interactions in the Urdu language. This research not only tackles the technical challenges of gesture generation but also advances the larger objective of enhancing AI's ability to interpret and respond to human nonverbal cues across different languages and cultural contexts.

Keywords:
Gesture Recognition; Natural Language Processing (NLP); Deep Learning; Urdu Text Analysis; CNN-LSTM Architecture.

1. INTRODUCTION

In addition to providing context for spoken words, non-verbal cues improve communication by helping to explain difficult ideas. Gestures provide the communication specifics that words are unable to express. Any entity's visual representation improves the context's clarity. These nonverbal cues include hand gestures like waving and pointing that are typically employed by anyone speaking. Gesture teaching approaches are used by special education systems to instruct their students. Regarding Human-Computer Interaction (HCI), the same phenomenon is true [1]. Creating user-friendly interfaces for consumers to utilize contemporary technology is the primary focus of HCI. Technology that is usable by people with special needs or disabilities is encouraged to be created, and certain parameters such as aesthetics, emotions, satisfaction, and engagement are taken into consideration when designing interfaces that are not only practical but also enjoyable and meaningful for users.

It aims to dismantle barriers and give everyone an equal opportunity to use computers, ensuring that everyone can benefit from and interact with them effectively. Deep Learning is essential to enable machines to mimic certain human talents [2, 3]. While many machine and deep learning models are being created to generate gestures from speech, the quality of the gestures is still not optimal. The same gestures can be made using text instead of speaking, according to a recent study [2].

This work uses a sequential Deep Learning Algorithm to provide a novel gesture creation model. With the use of a hybrid deep learning methodology centered on CNN and LSTM, this study creates a series of movements against text, with each gesture representing a different word. As opposed to this, we employ a standard text that isn't speaker-specific. We generate gestures using the sequential LSTM and extract features using a convolutional neural network. Researchers focusing on the human-computer interface are giving nonverbal communication modes more importance [4,5].

Although similar voices are combined with non-verbal clues in artificial movements, humans find humanoid robots to be more

lifelike. This enhances artificial agents' social communication abilities [10]. The elements of communicative skills, such as hand, arm, and facial expression motions, offer the best opportunities for improving these agents' communication capacities [6, 7].

While poses and sign language are both parts of an ongoing cycle and can be thought of as the two extremities of continuity, pantomime and "Italianate" are the two frontiers that fill the intermediate zone with words. Speech, sometimes referred to as spoken words, can aid in content comprehension and provide context for speech keywords. This has produced positive outcomes in interactions between humans and virtual agents [8].

When building and mapping humanoid robots, this method should be applied to give them a more lifelike appearance. In particular, humanoid robots are expected to behave and speak like people since they resemble people [9]. They are also expected to engage with people by claiming to be able to "communicate like humans" by moving their hands, arms, and faces. These impulsive hand and voice actions can be mimicked to improve artificial intelligence. These recently discovered models that generate an artificial gesture use RNNs, GANs, and several additional models that show good results on sequential data, such as LSTM. The voices and motions of the two creatures are to be modeled by artificial robots using these deep learning models and techniques. What we have contributed is:

• Provide a creative framework for text to produce motions.
• Make a comparison to demonstrate the suggested model's efficacy.

A portion of the relevant literature is included in section 3, where we describe our process for creating the speaker-free gestures dataset. In Section 4, the suggested Text-to-Gesture Model is explained.

Additional experimental results and discussions are included in Section 6. Lastly, sections 7 and 8 provide some limitations and suggestions for further development, respectively.

1.1 RELATED WORK

The main conclusions from the various techniques for creating gestures are given in this section.

References	Input	Techniques	Results	Limitations
[2]	Images of Object	Adversial Autoencoders, Self-Supervised Learning	Mean=38.56	Use only 6D pose offline detection.
[13]	Video and Images	A universal deep learning model with an attention-based graph with several branches.	The accuracy of the model was 94.1, 92.0, and 97.01%.	Using gestures and 3D hand skeleton data, a sign language-based communication system can be created.
[25]	2D images and videos	Auto encoder	MMRAEs efficiently decrease the size of the network while simultaneously increasing accuracy overall.	Avoid concentrating on limited modeling-based MMRAEs 172 by taking advantage of feature correlation both inside and between classes.
[26]	Audio	audio-gesture recordings from a database with a GAN model and KNN technique	This method performs better than the state-of-the-art in terms of audio-synchronicity and naturalness	One of the search-based algorithms' limitations is that it may require more computing time than its purely learning-based counterparts' single-pass inference techniques.

3. PROPOSED METHODOLOGY

Depending on their personalities, speech patterns, and situations, speakers use different gestures. In this study, spoken Urdu text and

2-D key-point coordinates of gestures made by a single speaker were combined to build a speaker-specific gesture dataset. Videos of a speaker speaking in front of a stationary camera are included in the dataset, and gestures made by the speaker are recorded using the OpenPose posture detection system [13]. Our Gesture Generation Model was created using the 49 2-D key points that the system had extracted regarding movements of the hand, arm, wrist, and shoulder. The Google Cloud Speech-to-Text API was used to convert the spoken content into text so that motions could be produced from Urdu text [14]. We separated the transcription into individual words and labeled any non-verbal frames with "." After that, vector representations of the textual data were created using FastText embeddings. The CNN and LSTM-based gesture generating model was trained using these vectors. The CNN was in charge of taking textual information and passing them to the LSTM layers so they could produce gestures that matched. In order to guarantee precise gesture portrayal, important points were pre-processed and matched the text. The collection is made up of movements taken from 15 videos that correspond to about 300 words. The model was trained using these words and their related movements. Standard measures, such as PCK (Percentage of Correct Keypoints) and MAE (Mean Absolute Error), were used in the model's evaluation to gauge how accurate the generated gestures were.

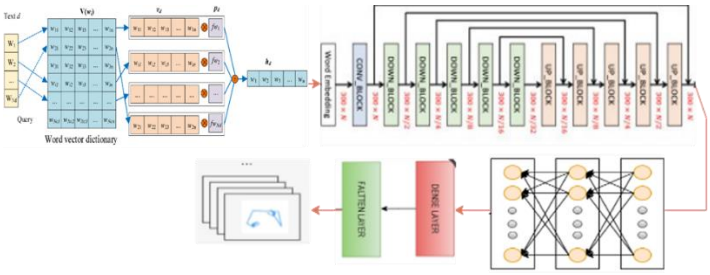

Figure 2: Proposed Text-to-Gesture Model [10]

crucial data regarding hands, arms, wrists, and shoulders. When looking for ideal performance, setting a rigorous threshold might not cause confusion because every joint is near to every other joint. The value of PCK vs the entire amount of time spent in training is plotted on the graph to demonstrate the effectiveness of the model. Correct Key-Points = No. of Key Points In between Threshold

$$PCK = \frac{No.of\ Corrected\ Key-Points}{Total\ no.of\ Key-Points} \times 100$$

$$MAE = \left(\frac{1}{n}\right) \times \sum |y_{actual} - y_{predicted}|$$

$$Sd = \sqrt{\sum_{i=}^{N} \left(\frac{1}{n}\right) \times \left(y_{actual} - y_{predicted}|\right)^2}$$

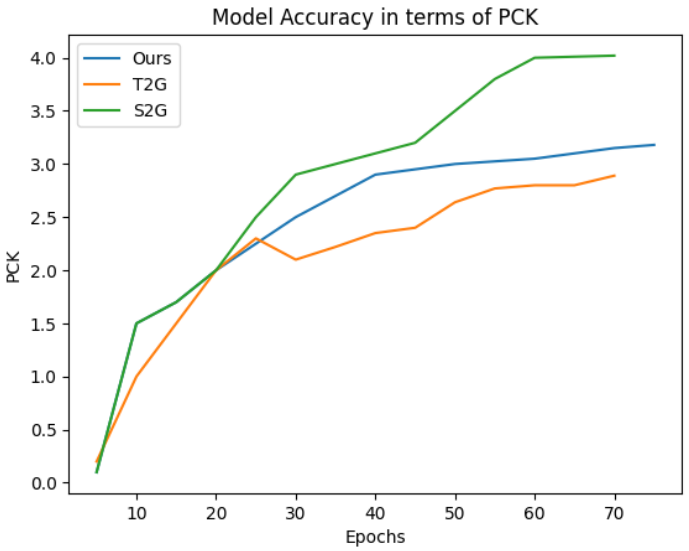

Figure 3 PCK

4. RESULTS AND ANALYSIS

Two performance evaluation matrices were utilized in the assessment of the suggested Text Gesture Model, (PCK): the percentage of rectified key points and the mean absolute error (MAE). The model was implemented in the PyTorch [82] environment. Figure 4.2 shows the highest percentage that we were able to obtain for important points spanning numerous epochs that were accurately predicted. Our dataset contained

The model's effectiveness is indicated by the graph, which demonstrates a rise in PCK value in relation to the quantity of training epochs. The suggested model's achievement is contrasted with two existing gesture models [2,13] in Fig. 3, which depicts the three-layered sequential LSTM model's performance. It is shown that PCK has increased in relation to the corresponding

number of epochs. Figure 3 shows the clearing of the Fall of Absolute Error with the relative Training Epochs.

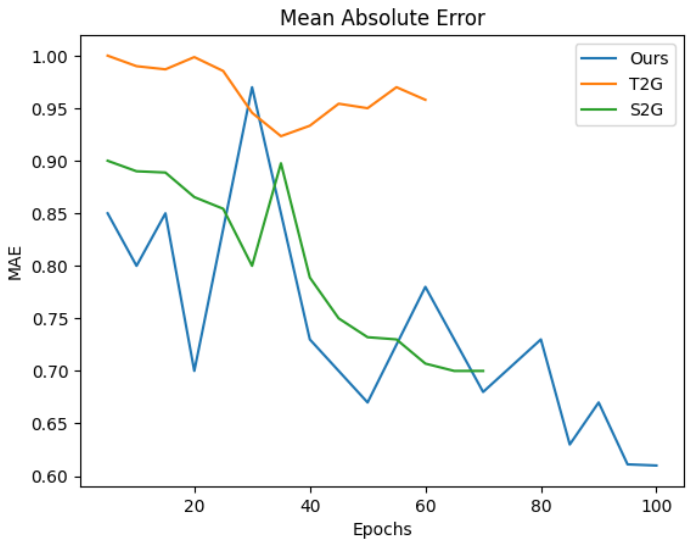

Figure 4 MAE

Table 1 presents the quantitative evaluations of the proposed model and a comparison of it with earlier methodologies. The model's efficacy is demonstrated by the highest accuracy it attained.

Table 1 Evaluation

	PCK	MAE	Threshold	Language	Speaker Specific
Our hybrid model	0.65	0.323	α = 0.1, 0.2	Urdu	No
T2G	0.288	0.958	α= 0.1, 0.2	English	Yes
S2G	0.4	0.707	α = 0.1, 0.2	English	Yes

5. LIMITATIONS AND FUTURE WORK

While the gesture creation system shown in this research yields high-quality gestures, it has a number of shortcomings. Among the few gesture variables in the dataset used to train the model were key spots on the wrists, arms, hands, and shoulders. There were just Urdu language terms in the submitted text. These limitations might motivate further study avenues to be looked at.

The field of study is not restricted in any way. Artificial intelligence is the only area in computer science that provides scholars and researchers with a wide range of research topics. A discussion of a few of them is given below:

After the model has been trained, the gesture model reported in this paper generates gestures utilizing an additional dataset. This opens up a new line of inquiry, and gestures can be produced in response to real-time input by developing and improving models. Many efforts have been undertaken in this area since this behavior is now being investigated.

The suggested approach might result in gestures that are in opposition to the Urdu-based language-based text input mode. By providing sample terms from languages other than Urdu, this effort can be made better. The character dataset from that language can be used to train the model in order to accomplish this. Keep in mind that gestures used in various languages can differ from one another.

6. CONFLICT OF INTEREST

The writers declare no conflict of interest

7. ACKNOWLEDGEMENT

We would like to thank everyone who help us to compile our work

8. REFERENCES

[1] Omid Alemi, William Li, and Philippe Pasquier. Affect-expressive movement generation with factored conditional restricted boltzmann machines. In 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pages 442–448. IEEE, 2015.

[2] Eiichi Asakawa, Naoshi Kaneko, Dai Hasegawa, and Shinichi Shirakawa Evaluation of text-to-gesture generation model using convolutional neural network. Neural Networks, 151:365–375, 2022.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multiperson 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7291–7299, 2017.

[4] Justine Cassell, Hannes H¨ogni Vilhj´almsson, and Timothy Bickmore. Beat:the behavior expression animation toolkit. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pages 477–486, 2001.

[5] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae). Geoscientific model development discussions, 7(1):1525–1534, 2014.

[6] Yi-Liang Chen and Feng Lin. Safety control of discrete event systems using finite state machines with parameters. In Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148), volume 2, pages 975–980. IEEE, 2001.

[7] Chung-Cheng Chiu and Stacy Marsella. How to train your avatar: A data driven approach to gesture generation. In International Workshop on Intelligent Virtual Agents, pages 127–140. Springer, 2011.

[8] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. Predicting co-verbal gestures: A deep and temporal modeling approach. In Intelli- gent Virtual Agents: 15th International Conference, IVA 2015, Delft, The

Netherlands, August 26-28, 2015, Proceedings 15, pages 152–166. Springer, 2015.

[9] Leon O Chua and Tamas Roska. The cnn paradigm. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 40(3):147–156, 1993.

[10] Richard Dosselmann and Xue Dong Yang. A comprehensive assessment of the structural similarity index. Signal, Image and Video Processing, 5:81–91, 2011.8

[11] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. In Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games, pages 1–10, 2019.

[12] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Expressgesture: Expressive gesture generation from speech through database matching. Computer Animation and Virtual Worlds, 32(3-4):e2016, 2021.

[13] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens,and Jitendra Malik. Learning individual styles of conversational gesture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3497–3506, 2019.

[14] Ahuja, C., et al. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. 2020. Springer

[15] Marc Erich Latoschik, Martin Frohlich, Bernhard Jung, and Ipke Wachsmuth. Utilize speech and gestures to realize natural interaction in a virtual environment. In IECON'98. Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society (Cat. No. 98CH36200), volume 4, pages 2028–2033. IEEE, 1998.

[16] Soonhye Park and J Steve Oliver. Revisiting the conceptualisation of pedagogical content knowledge (pck): Pck as a conceptual tool to nderstand teachers as professionals. Research in science Education, 38:261–284, 2008.

[17] Manuel Rebol, Christian G¨uti, and Krzysztof Pietroszek. Passing a non-verbal turing test: Evaluating gesture animations generated from speech. In 2021 IEEE Virtual Reality and 3D User Interfaces (VR), pages 573–581. IEEE, 2021.

[18] Bowen Wu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-gan and unrolled-gan. Electronics, 10(3):228, 2021.

[19] Zabala, U., et al. Learning to gesticulate by observation using a deep generative approach. in Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings. 2019. Springer.

[20] Zhang, Z. (2018, June). Improved Adam optimizer for deep neural networks. In 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS) (pp. 1-2). Ieee.

[21] Klein, B., L. Wolf, and Y. Afek. A dynamic convolutional layer for short range weather prediction. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[22] Guo, D., et al. Hierarchical LSTM for sign language translation. in Proceedings of the AAAI conference on artificial intelligence. 2018.

[23] Gibet, S., et al., Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges. Universal Access in the Information Society, 2016. 15: p. 525-539.

[24] Kusters, A. and S. Sahasrabudhe, Language ideologies on the difference between gesture and sign. Language & Communication, 2018. 60: p. 44-63

[25] Saunders, B., N.C. Camgoz, and R. Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[26] Camgoz, N.C., et al. Sign language transformers: Joint end-to-end sign language recognition and translation. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[27] Stoll, S., et al., Text2Sign: towards sign language production using neural machine translation and generative adversarial networks. International Journal of Computer Vision, 2020. 128(4): p. 891-908.

[28] Camgoz, N.C., et al. Multi-channel transformers for multi-articulatory sign language translation. in Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. 2020. Springer.

[29] Zelinka, J. and J. Kanis. Neural sign language synthesis: Words are our glosses. in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.

AUTHORS

First Author: Hussan Fatima, Masters in Computer Science, Faculty of Engineering & Computing. Institution: National University of Modern Languages Islamabad.

Second Author: Sammat Fareed, Bachelors in Computer Science. Institution: Alhamd Islamic University Islamabad Campus

Correspondence Author: Hussan Fatima, Masters in Computer Science, Faculty of Engineering & Computing, National University of Modern Languages Islamabad.