Grenze International Journal of Engineering and Technology, June Issue

A Speech Controlled Visual Storytelling App-Assistant for E-learning

Prof. N.S. Shirsat[1], Vaishnavi Vishnu Jagdale[2], Rajkumar Madhav Pawar[3], Prerana Chandrakant Jadhav[4] and Shivang Jalandhar Gupta[5]
[1]Assistant Professor, Department of Information Technology ,PVG's College of Engineering and Technology & G.K. Pate (Wani) Institute of Management, Pune, India
Email: nss_it@pvgcoet.ac.in
[2-5]PVG's College of Engineering and Technology & G.K. Pate (Wani) Institute of Management, Pune, India
Email:[vaishnavijagdale2002, rajm.pawar7038, preranacj2001, shivang17d]@gmail.com

***Abstract*—E-learning refers to the use of electronic technologies to deliver educational content and facilitate learning outside of traditional classroom settings. E-learning offers flexibility in terms of time, location, and pace of learning, making it accessible to a diverse range of learners. The main problem identified with these tools is the limited engagement and accessibility for children and very less interaction between students and teachers as well as kids and parents. Traditional E-learning tools are lacking interactive elements that promote active engagement and participation from learners. This article exhibits a comprehensive system where users can provide speech input to the system, which then retrieves relevant images from a dataset based on natural language processing (NLP).The use of NLP ensures that the flow of the story is maintained seamlessly, as the system intelligently positions images at appropriate points in the narrative. Leveraging computer vision technology, individual images stored in the dataset are accurately placed within the context of the story, enhancing interactivity and engagement. By combining speech recognition, NLP, and computer vision for image positioning, this solution offers a more interactive and immersive learning experience, particularly in the context of narrative-driven content like Panchatantra stories.**

***Index Terms*— Speech recognition, Natural language processing, computer vision, image positioning, coreference resolution.**

I. INTRODUCTION

Fig.1 shows the blueprint of an application. Digital storytelling [1], at the crossroads of traditional oral narration and interactive multimedia technologies, represents a dynamic and transformative approach to conveying narratives. This innovative method leverages the power of technology, specifically Automatic Speech Recognition, Natural Language Processing (NLP), and Computer Vision, to seamlessly blend spoken language with contextually relevant visual narratives. Digital storytelling provides educators with a unique opportunity to engage learners by tapping into their creativity through immersive, visually compelling narratives.

In the realm of human-computer interaction, the ability to comprehend and respond to spoken language has emerged as a pivotal capability. Leveraging speech Recognition technology, this research aims to transcribe spoken input into textual descriptions that serve as the foundation for an immersive visual narrative. The exhibited system consists of three main modules- Automatic speech recognition, Image Retrieval and Image

Grenze ID: 01.GIJET.10.2.608_1

Positioning. Automatic speech recognition (ASR) is a technology that enables the transcription and interpretation of spoken language into text. Image retrieval using Natural Language Processing (NLP) involves leveraging linguistic cues from text data to search and retrieve relevant images from a dataset. By focusing on the semantic context, the system can retrieve images that vividly capture the essence of the spoken scenario, thereby transforming traditional learning experiences into interactive and engaging ones. Furthermore, this research explores the application of hidden states within the NLP module to facilitate contextual analysis, allowing for a deeper understanding of the user's narrative.

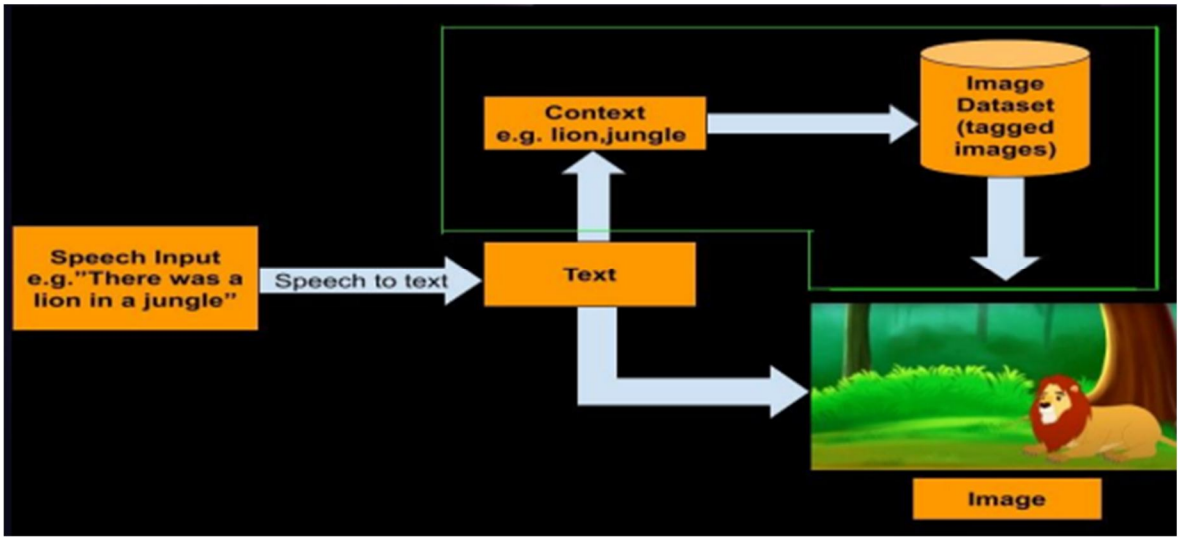

Fig.1. Blueprint of the application

Image positioning using computer vision (CV) and prepositions involves analyzing the context of a sentence to determine the spatial relationships between objects described in the text and positioning images accordingly. This article demonstrates the efficiency of this integrated approach in transforming static learning experiences into dynamic and interactive journeys. By presenting a coherent visual narrative driven by spoken language, we aim to capture the learner's attention, stimulate their imagination, and fundamentally redefine the boundaries of educational engagement.

II. Literature Survey

In proposing the Speech-Controlled Visual Storytelling App-Assistant for E-learning, various approaches have been explored for the implementation of submodules.

Y. Li [1] proposed StoryGAN, a model that converts multi-sentence paragraphs into sequential images, enhancing narrative understanding. The model employs a sequential conditional GAN framework with a deep Context Encoder, achieving superior image quality and contextual consistency metrics. However, StoryGAN faces challenges due to its substantial computational demands, potentially limiting accessibility and scalability. Additionally, the creation of new datasets introduces potential biases, affecting the model's generalizability.

G. Zeng [2] addressed children's distance learning with Pororogan, a GAN-based model demonstrating deep semantic understanding. The proposed method effectively distinguishes between common and complex actions, generating dynamic image sequences to enhance the learning experience. Despite these strengths, the model faces challenges in handling highly complex or abstract children's stories. The computational demands for generating dynamic image sequences may limit real-time applications or accessibility on resource-constrained devices. Furthermore, the coherency metric used for evaluation may not fully capture the nuances of children's story visualization, suggesting a need for further research to refine assessment methods.

Y. Song [3] proposed CP-CSV, a framework addressing the challenge of maintaining global consistency in story visualization. CP-CSV effectively tackles these challenges by incorporating three critical modules: a story and context encoder for learning story and sentence representations, a figure-ground segmentation module (an auxiliary task) to provide information for preserving character and story consistency, and a figure-ground aware generation module for generating image sequences while incorporating figure-ground information. Furthermore, the model introduces a novel metric called "Fréchet Story Distance" (FSD) to evaluate the performance of story visualization. Through extensive experiments, the authors demonstrate that CP-CSV excels in maintaining

character details and achieving high consistency among different frames, while FSD serves as an improved metric for assessing story visualization performance. However, the model faces limitations, including significant computational demands, reliance on diverse training data availability, potential lack of human interpretability, and challenges in real-time applications due to computational complexity.

III. Proposed System Overview

Fig. 2 shows the architecture of proposed system. The approach is divided into three parts and:

A. Automatic Speech Recognition

1. Audio will be captured from the microphone.
2. A message will be displayed to inform the user that the program is actively listening.
3. The captured audio will be converted to text
    a. If successful, the recognized text will be printed.
    b. The text will be split into sentences based on periods ('.') and any leading or trailing whitespace will be removed from each sentence.
    c. Non-empty sentences will be added to the list.

B. Image Path Retrieval

1. Iterating over each sentence in the input list of sentences is done
    a. The sentence is processed to obtain a parsed document and these documents are stored to prevent garbage collection.
    b. Coreference spans i.e. which pronoun is used for which noun is extracted from the processed document.
2. For each sentence:
    a. Pronouns are replaced with their corresponding noun heads based on the identified coreference clusters.
    b. Sentences are updated with the replaced nouns
3. For each updated sentence:
    a. Nouns, adjectives, and verbs from the updated sentence are extracted. Identified characters from the nouns and actions from adjectives and verbs
    b. Identified characters and associated tags are compared with the image dataset and image paths are retrieved corresponding to the identified characters and tags.
    c. For each sentence noun is replaced by corresponding image path and then sentence is trimmed and only image paths and preposition is kept

C. Image Positioning

1. Corresponding image from the specified path is loaded and resized to match the specified size.
2. Relative positions of objects based on prepositions and object types are calculated.
3. Canvas is created by resizing the background image.
4. Iterated over the objects in the specified rendering order and placed each object image onto the canvas at its specified position.

D. User Interface

1. User interface contains mic to give input and text box to provide text input based on user’s choice
2. Images will be displayed on the screen

E. Dataset

1. One list contains information about characters or entities within the context of the application. Each dictionary in the list represents a distinct character
2. One dictionary provides more detailed information about specific objects or characters, possibly used for rendering or visualization.
3. Another list specifies the order in which the objects or characters should be rendered
4. All this information is contained in a python file to make it easy to import this data to other python files.

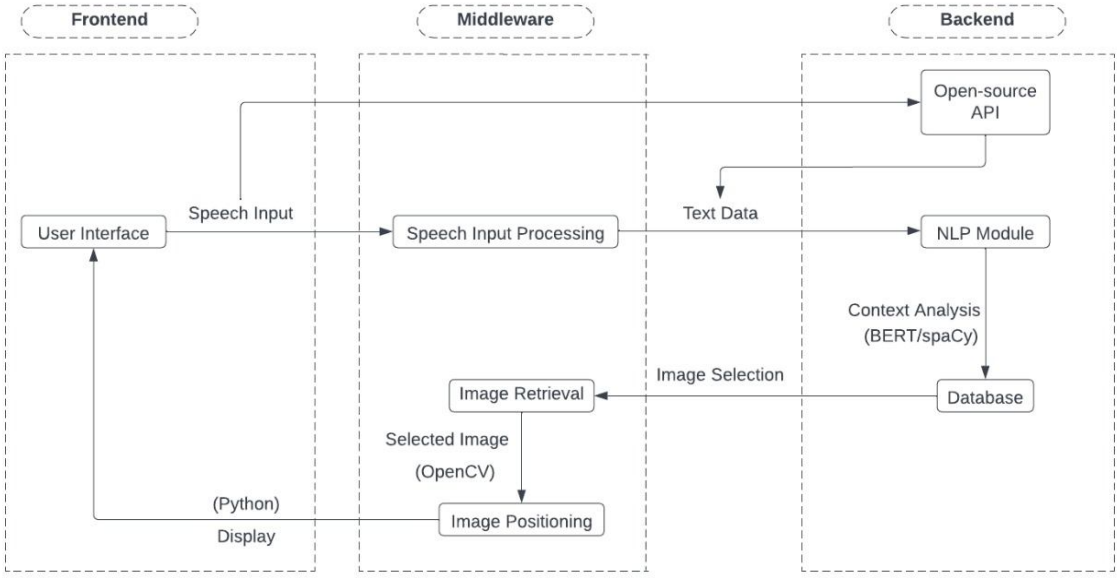

Fig. 2. Proposed system architecture

IV. Data Preparation

A. Need to Create Dataset

The datasets available on internet sources mostly comprise the entire scene of the story but our requirement was as follows:

- Knowledge of the characters
- Image files representing each character
- Tags or descriptors for each character
- Knowledge of the required size of each object or character image

B. Data Acquisition

- Images representing each character or entity were gathered for inclusion in the dataset and the background of each image was removed and converted into .png format.
- it was determined whether to include any tags or descriptors for each character/entity. These attributes, such as "sleeping," "angry," or "scared," were considered based on the context of the application.
- the size of each character image was determined. This involved measuring the dimensions of each image file or deciding on a standard size for all images.

C. *Dataset Creation*

- Once all the necessary information was gathered, it was organized into a structured format. This involved creating dictionaries or JSON objects for each character/entity, with keys for "character," "tags," and "image_path."
- Dictionaries or JSON objects were created for each object or character image, including keys for "image_path" and "size."
- The rendering order of the objects or characters was decided upon and organized into a list accordingly.
- Subsequently, all the individual dictionaries and lists were combined into a single dataset and stored in a python file.

D. Metadata

- Total number of images: 15+ (for single story)

- Software used to remove background of each image: romove.bg
- Sample images from our dataset (Fig. 3 shows sample images from dataset)

Fig. 3. Character images from the dataset

V. System Implementation

The proposed system demands NLP (Natural Language Processing) for understanding the sentence provided by the user and computer vision to position the image appropriately.

A. Image Path Retrieval

Image path retrieval is mainly done into two phases.
Each input sentence is processed using spaCy's (NLP library) coreference resolution capabilities. It replaces the default coreference resolution components with spaCy's transformer-based coreference resolution .It means for each sentence which pronoun is used for which noun is identified. Regular expressions are used to replace each pronoun occurrence with its corresponding head noun in the updated sentence. Again based on the noun, adjective and verb image paths are retrieved and the sentence is updated by replacing each noun with it's corresponding image path. Updated sentence is cleaned by keeping only image path and preposition and all other parts of the sentence are removed.
After cleaning the sentence the tuple is made for each pair of noun image path and preposition between them, if pairing is not possible then 'on' is kept as default preposition and one invisible image with very less size is paired with remaining image path and such tuples are made. For each sentence list of such tuples is made and this list will be input to image positioning module.

B. Image Positioning

The list for each sentence is returned by image path retrieval module. For each sentence, relative positions of the identified objects are identified based on prepositions like "on," "in," "under," etc. Using predefined rules, it is decided where each object should be placed in relation to others. The positions are determined on the basis of trial and error. After determining the positions, Object images are resized based on the specified sizes. Resized images are placed onto a canvas, ensuring that they appeared in the correct positions relative to each other. The scene is rendered by placing the objects in a specified order. It is ensured that objects are placed correctly, considering their relative positions and any overlap.
The composed image is saved as a png file and it is rendered on the user interface.

VI. Results

A. Image Path Retrieval

Fig. 4 shows the original sentence and image paths retrieved from that sentence.

B. Image Positioning

Fig. 5 shows how images get positioned and composed on the canvas

Fig. 4 Image Path Retrieval Output

Fig. 5. Output image1

C. User Interface

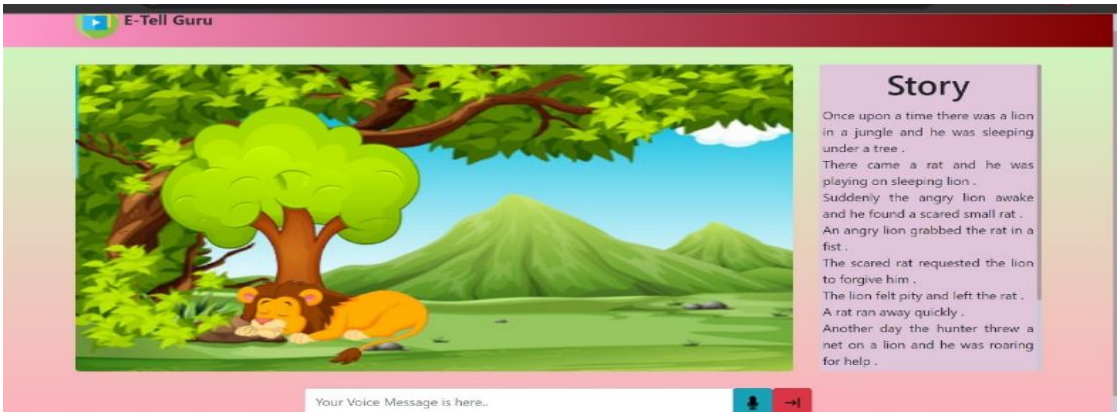

Fig. 7. User Interface

VII. CONCLUSION

This article exhibits a speech controlled visual storytelling app-assistant for e-learning consisting of two modules one is to retrieve the image paths based on the characters in the sentence and other is to position the images appropriately so that it look appropriate. This system has been designed keeping in view the flow of the story. Image retrieval is used instead of image generation to maintain the speed and flow of the story which is making this system engaging for the kids. Cartoon images are used to attract kids and to make the system enjoyable for them. Speech input is used to enhance the interaction between a kid and storyteller who may be a teacher or parent. An integration of natural language processing and computer vision techniques to interpret textual

descriptions and to compose corresponding visual scenes is key achievement of the system. A Seamless rendering and display of composed scenes, enhancing user interaction and understanding of the generated content. The exhibited system can be used in educational institutions for explaining complex concepts of history, geography and science as well as in corporate world to convey ideas in an efficient manner. The system can be scaled to support training programs by generating visual scenarios that simulate real-world workplace situations. Employees can practice problem-solving, decision-making, and interpersonal skills in a safe and controlled environment.

ACKNOWLEDGEMENT

We appreciate the resources and support provided by PVG's COET and GKPIOM Pune and I/C principal of the college Dr. M.R. Tarambale.Special thanks to head of the department of information technology Dr. S.A. Mahajan. Our gratitude also goes to the reviewers of this project Prof. M.R. Apsangi and Prof. N.R. Sonawane of for their insightful feedback.

REFERENCES

[1] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. E. Carlson, and J. Gao, "Storygan: A sequential conditional GAN for story visualization," in IEEE CVPR, 2019.
[2] G. Zeng, Z. Li, and Y. Zhang, "Pororogan: An improved story visualization model on pororo-sv dataset," in ACM CSAI, 2019.
[3] Y. Song, Z. R. Tam, H. Chen, H. Lu, and H. Shuai, "Character-preserving coherent story visualization," in ECCV, 2020.
[4] S. Kotagiri, A. Venkataramana and G. Kiran, "Blind Aid: State of the art for Scene Text Detector and Text to Speech," 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, 2022, pp. 1-5.
[5] M. N. Munjal and S. Bhatia, "A Novel Technique for Effective Image Gallery Search using Content Based Image Retrieval System," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019.
[6] Z. Chen, "Neural Language Models in Natural Language Processing," 2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI), Zakopane, Poland, 2023, pp. 521-524.
[7] I. Carvalho, H. G. Oliveira and C. Silva, "The Importance of Context for Sentiment Analysis in Dialogues," in IEEE Access, vol. 11, pp. 86088-86103, 2023.
[8] Duygu Altınok, Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem, Packt Publishing, 2021.
[9] J. Praveen Gujjar, H. R. Prasanna Kumar and M. S. Guru Prasad, "Advanced NLP Framework for Text Processing," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023.
[10] B. Wang, F. Liu, X. Yang, M. Hu and D. Li, "Span extraction and contrastive learning for coreference resolution," 2022 International Conference on Intelligent Manufacturing, Advanced Sensing and Big Data (IMASBD), Guilin, China, 2022, pp. 98-103.