IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Real-Time Narration Tool

Sai Roge [1], Yash Chaugule [2], Kabir Bokade [3]

[1,2,3] Student, Department of Electronics and Telecommunication Engineering

K.J. Somaiya Institute of Technology Sion, India

Prof. Vandana Maurya [4]

Asst. Professor, B.K. Birla College Kalyan, India

Prof. Sandeep Mishra [5]

Asst. Professor, K.J. Somaiya Institute of Technology Sion, India

I. ABSTRACT

Imagine, if everything that you're looking for is already within earshot, yet one can never even get a look at it. The Real-Time Narration Tool (RTNT) is the first in the world to convert visual information directly into real time audio narration. Be it your basketball game at the tail end being described as a buzzer beater, or just how pretty is that sunset out there, or be you in the middle of a car chase with someone when live video streaming comes into effect; RTNT affords humankind access to information they hitherto lacked via their ear drums. Such an earth shaking feat is existing via harnessing services across any technological trifecta. RTNT runs on portions of computer vision, natural language processing, and text-to-speech generation. It first admits that there is a live video feed and within milliseconds determines scenes such as GPT technology and responds with guided narratives to action and intention. With Eleven Labs text-to-speech, the oral output is so natural that one cannot tell if a computer is speaking or a human is rendering a story. Besides just testing this tool on many successful challenges, I tested it in practical situations, worldwide. It worked wonders—with implications for amazing realities—in arenas and auditoriums, gameplay, and classrooms. It has an innate sense of purpose for functionality where interfacing with arenas that are usually only seen is made accessible to many different types of doers. Even interfacing mixed realities such as virtual realities opens up blended proceedings of access and actionable response. The possibilities are endless. Future versions include multilingual offerings for global audiences and even more AI-smart narratives that learn more than merely responding to directives. This RTNT will one day service the universe and the realms of accessibility, education, and entertainment—not merely as another narrative experience that so many take for granted—but as something that will change how people experience life in their worlds with no one left behind.

Keywords - Real-Time Narration; Accessibility; Natural Language Processing; Audio Descriptions; Voice Synthesis

II. INTRODUCTION

This basic ability to perceive through multiple senses is the basis for much human interaction. Dealing with images in the context of information access for one with visual disabilities can be a troublesome process. Textual descriptions or, at the very least, audio warnings do scant justice to the depth of the experience. The Real- Time Narration Tool is a prototype meant to ameliorate these problems. It provides an auditory description of the images in real time in order to enhance engagement and understanding. Despite such advances in technology, many visually impaired people have yet to have full access to visual media-such as sports events, educational materials, or entertainment-where this alienation brings into view the need for innovative solutions for incorporating accessibility. Current assistive technologies provide little immediacy and lack the context that makes dynamic visual experiences accessible to all users.

III. LITERATURE SURVEY

The literature surrounding audio narration and accessibility technologies reveals significant advancements and persistent challenges in providing real-time, context-sensitive descriptions for visually impaired individuals.

Zhang. et al, Understanding the Limitations of Large Language Models –

The authors considered the limitation of large-scale language models and noted specifically the need for interpretability and transparency of the generated outcome. In fact, this research brings out the proper models which should be developed based on higher clarity concerning their decision-making procedure. This becomes quite important for applications such as RTNT that work under NLP for coherent description generation.

OpenAI, GPT-4 Technical Report –

The report clearly describes the architecture and abilities most recently known to have phased out the previous models. Nonetheless, there is not so much exploration pertaining to issues of massive social significance, ethical implications, and built-in biases in large-scale language models. It emphasizes the need to devise tools to empower advanced NLP techniques yet recognizing the ethical frameworks so that this access is equitable to all users.

ElevenLabs, Advancements in Neural Text-to-Speech Technologies –

This covers the state-of-the-art technology that is involved in neural text-to-speech and high-fidelity audio synthesis. Whereas the quality and naturalness in voice have experienced an evolution through TTS, the multilingual feature and good representation of regional accent as part of speech synthesis are not put forth efficiently. This stands as a great opportunity for RTNT in further bettering the user-friendliness options towards catering to the linguistic needs of varied cultural user profiles.

Moreover, there are many audio description services that are based on pre-recorded scripts which lack flexible real-time applications. Conventional methods represent a mere shadow of measuring the dynamism that live events and interactive environments, those where the content changes infinitely, can adopt. The aim of Real- Time Narration Tool is to bridge such a gap through providing instant audio descriptions that adjust time into real- time visual input which aims to enhance user involvement and understanding of content.

IV. METHODOLOGY

The development phase is the most critical phase of development of the Real-Time Narration Tool, in which many different technologies are used together to finally arrive at an operating prototype. In this section, the focal part deals with development, which encompasses input acquisition and processing, description generation, and system architecture for audio synthesis.

A. Technology Stack

The RTNT utilizes a combination of programming languages and libraries to achieve its functionality: **Programming Language**: Python is chosen for its versatility and extensive support for libraries related to artificial intelligence, machine learning, and audio processing.

Libraries:

a. OpenCV: Tools that capture image frames from video sources, detect objects, and allow image analysis.

b. OpenAI GPT API: API for visual input, producing a natural language description from the input through advanced NLP techniques, capable of broad coherent, contextually relevant narratives.

c. ElevenLabs Text-to-Speech: This engine converts the generated text descriptions into high- quality audio output. It possesses voice cloning features which allow the creation of a personalized narrator.

B. System Architecture

RTNT, or real-time narration technology, consists of several interconnected components working hand in hand to provide narration on a real-time basis:

a. Camera Input:
   The camera is relatively straightforward, keeping the camera fed a real-time video impression of the environment. This can either be a webcam or simply an exterior camera attached to the computer.
   i. The camera continuously streams video. It is an important requirement for the processing of dynamic scenes.

b. Image Processing Module:
   i. This module processes the video frame by frame using OpenCV. A plethora of functions are performed, such as object detection and motion tracking, which help detect relevant objects in each frame.
   ii. The accuracy of this module is crucial, as it directly impacts the quality of the generated descriptions.

c. Scene Description Generation Module:
   i. The information of the detected objects is then forwarded to the OpenAI GPT API. This module then develops a natural language description based on the identified objects and their context within the scene.
   ii. For example, the camera detects a “boy” and “mobile” , the description would be “a young boy clicking a selfie.”

d. Text-to-Speech Engine:
   i. The textual description gets passed on to the “Elevenlabs” which is a TTS engine.This engine turns the description to human-like voice using synthetic voices.
   ii. Voice cloning capabilities can potentially be utilized to add a personal touch to the narration. Voice cloning capabilities can be utilized to create a more personalized narration experience.

e. Audio Output Device:
   i. The speech gets passed through speakers which makes users to listen to narration.

C. Input Acquisition and Processing

a. RTNT uses OpenCV to continuously capture frames and the corresponding activities in the frame.
b. The image-processing module takes frames from the video feed and analyzes the frames in search of key elements.
c. Detected objects are then categorized and prepared for description generation.

D. Description Generation

a. NLP models summarize the scene by generating natural language descriptions based on the detected objects and their relationships.
b. OpenAI GPT API is being used to generate the textual description of the input data.
c. The API analyzes the data and outputs coherent description of what is happening in real time.

E. Audio Synthesis

a. ElevenLabs TTS engine converts the description to audio format.
b. The TTS engine synthesizes speech based on the textual output from the NLP module.

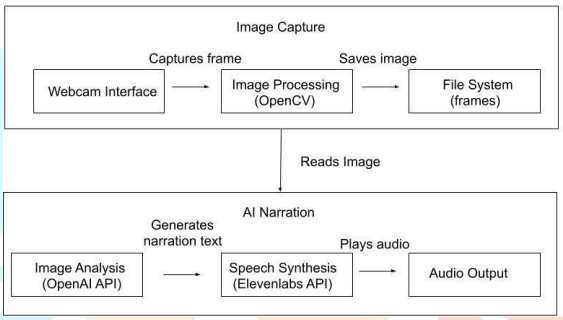

Figure-1 : System architecture

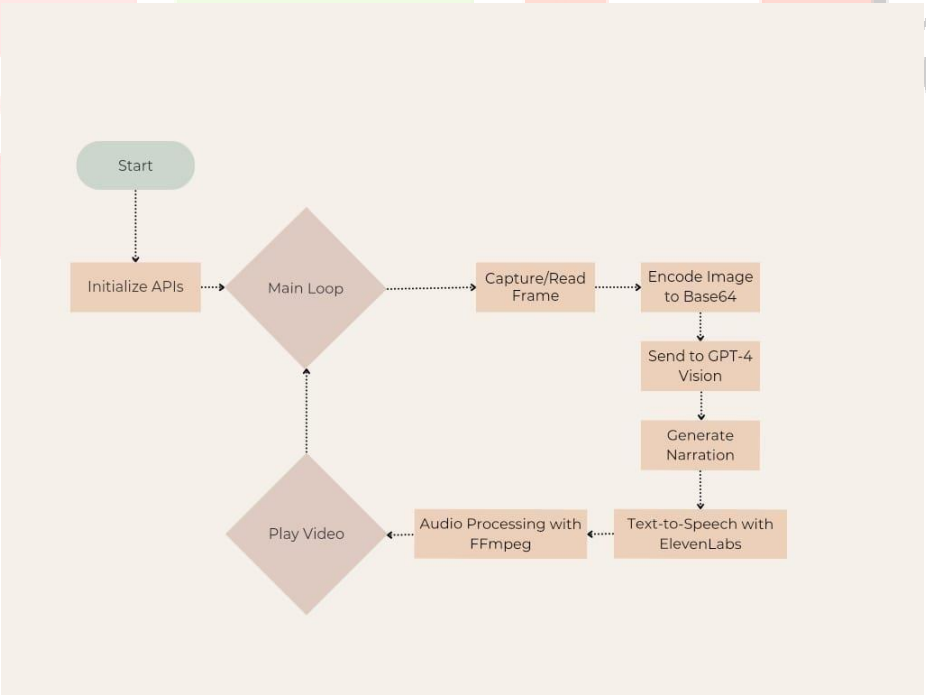

Figure-2 : Flowchart of RTNT

V. TESTING

A carefully planned participant recruitment strategy will be implemented to ensure that the Real-Time Narration Tool (RTNT) effectively meets user needs. The aim is to collect valuable feedback from those who stand to gain the most from this technology.

Target Participants -

A. Visually Impaired Individuals – The primary beneficiaries of RTNT will include individuals who suffer from visual impairments, ranging from complete blindness to those with partial vision. Participants would be recruited from non-profit organizations and other institutions to ensure the practical usage of RTNT in daily life.

B. Educators and Instructors – RTNT can be used for teaching purposes like explaining science experiments and make classes more interactive even in remotest places of world.

C. Recruitment Approach – Our program would make sure that it collects proper consent from both participants and educators while testing our tool and ensuring its practicality, user-friendly and accurate descriptions of the surroundings.

The user testing process for RTNT will be categorized into two primary phases :

A. Usability Testing in Controlled Environment

The first phase will occur in settings such as classrooms where educators can engage with our tool to concentrate on the following aspects:

a. User-friendliness – To what extent is the interface accessible to users?
b. Navigation efficiency– Are users capable of interacting with the system in an effective manner?
c. Audio quality – Is the narration delivered in a clear, coherent, and comprehensible manner?
d. Overall user satisfaction – To what degree does RTNT meet the expectations of its users?

B. Field Testing (Real-World Application)

The RTNT system will be tested and evaluated live dynamic environments which include the following:

-Every real time moment of a particular sports to check its real time accuracy
- Classroom visual materials describing educational content

Individuals will be able to check the accuracy, user friendliness and efficiency of the RTNT and how well it adapts to the input video and to evaluate the auditory experience to make a qualitative and quantitative approach towards the future iterations of the tool.

VI. EXPECTED FINDINGS

Several outcomes are expected for the initial development phase of RTNT to modify the upcoming versions of the tool. This can be categorized into following :

A. User Experience Insights

The testing process will also center on how the users perceive and interact with RTNT. Participant feedback will aid in assessing

- Ease of Use: Is the CLI interface user-friendly or need modifications to make it more accessible?
- Seamless Integration: Do users find it smooth and reliable how the visual input relates to the audio narration?
- Engagement & Satisfaction: Do participants feel that RTNT enhances their ability to engage with their surroundings?
- Entertainment & Informational Value – Are the narrations entertaining, informative, and natural or very mechanical and disconnected?

B. Application Versatility

The results would highlight RTNT's operations in different contexts:

In education, feedback may reveal that the narrations in real-time enhance understanding of more complex materials.

In live events feedback might indicate a great improvement in access for the blind, offering them a visual experience equal to that of sighted people.

VII. APPLICATIONS

The advantages will be significant, as visually impaired individuals gain greater independence and a better understanding of their surroundings, while various sectors like healthcare, education, entertainment, and agriculture also benefit..

A. Medical Applications

RTNT has vast promises in the field of health as it improves efficiency, communication, and patient care.

a. Surgical Assistance

The RTNT will provide verbal, step-by-step instructions for intricate surgical procedures, enabling the surgeon to focus on the task at hand while receiving real-time updates on the patient's clinical status. This reduces reliance on visual monitors, thereby improving response times to critical changes.

b. Remote Healthcare & Telemedicine

With the increasing popularity and use of telehealth services, RTNT can provide real-time summaries of a patient's health status, presenting that information in a clear and organized manner. This enhances communication between doctors and patients, ensuring that important details are not overlooked during virtual consultations..

c. Patient Monitoring

To track vital signs of patients and alerting respective staff and doctors to take precatious measures in order to save patients life minimizing constantly visual checks from clinical staff.

.B. Social and Educational Impact

RTNT makes the world an open platform for the visual impaired to experience more taste of the surroundings and to take actions without any worries.

a. Accessibility & Public Spaces

RTNT offers real-time descriptions enabling blind people to navigate independently in public spaces like parks, railway stations, museums, and shopping malls. Education & Learning Improvement

b. Description of complex educational visuals with RTNT such as :

- Diagrams and graphs in textbooks
- Mathematical equations on a whiteboard
- Science experiments and demonstrations in classrooms.

VIII. CONCLUSION

RTNT represents a significant advancement in helping blind individuals access their surroundings by providing immediate audio descriptions. Utilizing computer vision, natural language processing, and text-to-speech technology, this tool delivers contextual and dynamic audio feedback that aligns with real-life scenarios. Evaluations and tests indicate that the tool effectively interprets and narrates visual information, offering users clear, relevant, and engaging auditory responses. However, there are still some challenges that need to be tackled. Future improvements include adding customization features to make RTNT a user friendly and multilingual capabilities .

IX. REFERENCES

[1] Zhang, et al. (2023). Understanding the Limitations of Large Language Models

[2] OpenAI. (2023). GPT-4 Technical Report.

[3] ElevenLabs. (2023). Advancements in Neural Text-to-Speech Technologies , Blog