GREET - A VIDEO CONFERENCING PLATFORM

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING(CSE)

SUBMITTED BY

17 October 2023

KIET Group of Institutions, Delhi-NCR,
Ghaziabad (UP)
Department of Computer Science and Engineering

Name of Student	AMAN BHATT
University Roll No.	2100290100023
Class Roll No.	23
Branch	CSE
Name of Student	KANISHK CHAUDHARY
University Roll No.	2100290100078
Class Roll No.	12
Branch	CSE
Name of Student	SARTHAK SINGHAL
University Roll No.	2100290100148
Class Roll No.	16
Branch	CSE
Batch	2025
Proposed Topic	GREET - A VIDEO CONFERENCING PLATFORM
Submitted by	17 October 2023

Table of Content

Introduction

Existing video conferencing platforms lacks features like - multilingual communication, emotion recognition, sign language interpretation, and post-meeting summarization in a single application.Therefore, there is a need for a comprehensive platform that integrates these features to enable effortless multilingual conversations, enhance emotional understanding, facilitate communication for individuals who use sign language, and provide efficient post-meeting summary.

Technology Used:

The project employs a combination of medical imaging, machine learning, and image recognition technologies:

1. **Language Translation:**
   a) Integration with machine translation API’s like Google Translate or Microsoft Translator.
2. **Emotion Recognition:**
   a) Open CV
   b) Deep learning frameworks like Tensor Flow or PyTorch
3. **Sign Language Translation:**
   a) Computer vision libraries like Open CV or Tensor Flow
4. **Automatic Summarization:**
   a) NLP

Field of Project:

Developing a comprehensive video conferencing application with multilingual support, emotion recognition, sign language translation, and auto summarization is a multifaceted endeavour. This project involves software development, natural language processing, computer vision, user experience design, security, scalability, AI, server infrastructure, and customer support. It demand

a diverse and skilled team working together to deliver a sophisticated, inclusive, and privacy-conscious solution for users.

Special Technical Terms:

1. **Natural Language Processing (NLP):**

A branch of artificial intelligence that focuses on the interaction between computers and human language, including tasks like language translation and sentiment analysis.

2. Summarization Algorithm:
A set of rules or a computational model that automatically condenses longer text or spoken content into shorter, more concise summaries.

3. WebRTC (Web Real-Time Communication):
A free, open-source project that provides web browsers and mobile applications with real-time communication capabilities.

Rationale

The development of this multifunctional video conferencing application addresses the evolving needs of a diverse and interconnected world. By combining multilingual support, emotion recognition, sign language translation, and auto summarization, the application aims to revolutionize the way people communicate, fostering inclusivity, empathy, and efficiency in online interactions.

Objectives

- Develop a multilingual video conferencing platform with real-time language translation.
- Implement an emotion recognition system for enhanced emotional understanding during video conferences.
- Create a sign language translation feature for individuals who rely on sign language.
- Design an automatic summarizer to generate concise meeting summaries.
- Ensure a user-friendly interface for easy toggling and access to features.

Literature Review

They present MeetDot, videoconferencing system with live translation captions overlaid on screen. [1] The system aims to facilitate conversation between people who speak different languages, thereby reducing communication barriers between multilingual participants. Currently, their system supports speech and captions in 4 languages and combines automatic speech recognition

(ASR) and machine translation (MT) in a cascade. They use the re translation strategy to translate the streamed speech, resulting in caption flicker. Addition ally, their system has very strict latency requirements to have acceptable call quality. They implement several features to enhance user experience and reduce their cognitive load, such as smooth scrolling captions and reducing caption flicker. The modular architecture allows them to integrate different ASR and MT services in their backend. Their system provides an in targeted evaluation suite to optimize key intrinsic evaluation metrics such as accuracy, latency and erasure. Finally, they present an innovative cross-lingual word-guessing game as an extrinsic evaluation metric to measure end-to end system performance. They plan to make our system open source for research purposes.They describe MeetDot, a videoconferencing system with live translation captions, along with its components: UI, ASR, MT, and captioning. They implement an evaluation suite that allows them to accurately compute metrics from the sequence of captions that users would see. They also describe a cross-lingual word game for A/B testing different captioning algorithms and conditions. Their future work includes improved ASR/MT, extrinsic testing, and an open-source release. Their overall goal is to provide a platform for developing translation captions that are accurate and “right behind you.”

Sign language is used by deaf and hard hearing people to exchange information between their own community and with other people. [2] Computer recognition of sign language deals from sign gesture acquisition and continues till text/speech generation. Sign gestures can be classified as static and dynamic. However static gesture recognition is simpler than dynamic gesture recognition but both recognition systems are important to the human community. The sign language recognition steps are described in this survey. The data acquisition, data preprocessing and transformation, feature extraction, classification and results obtained are examined. Some future directions for research in this area also suggested.
After thorough analysis, the following are conclusions for future research in sign language recognition:

- Current systems are mainly focused on static signs/ manual signs/ alphabets/ numerals.
- Standard dataset not available for all countries/subcontinents / languages.
- A need for large vocabulary database is the demand for current scenario.
- Focus should be on continuous or dynamic signs and nonverbal type of communication.
- Sign language recognition systems should adopt data acquisition in any situation (not restricted to laboratory data).
- Systems should be able to distinguish face, hand (right/left) and other parts of body simultaneously.
- Systems should perform recognition task in a convenient and faster manner.

Automatic emotion recognition based on facial expression is an interesting research field, which has presented and applied in several areas such as safety, health and in human machine interfaces. [3] Researchers in this field are interested in developing. techniques to interpret, code facial expressions and extract these features to have a better prediction by computer. With the remarkable success of deep learning, the different types of architectures of this technique are exploited to

achieve a better performance. The purpose of this paper is to make a study on recent works on automatic facial emotion recognition FER via deep. learning. We underline on these contributions treated, the architecture and the databases used, and we present the progress made by comparing the proposed methods and the results obtained. The interest of this paper is to serve and guide researchers by review. recent works and providing insights to make improvements to this field.
This paper presented recent research on FER, allowed us to know the latest developments in this area. We have described different architectures of CNN and CNN-LSTM recently proposed by different researchers and presented some different database containing spontaneous images collected from the real world and others formed in laboratories, to have and achieve an accurate detection of human emotions. We also present a discussion that shows the high rate obtained by researchers that is what highlight that machines today will be more capable of interpreting emotions, which implies that the interaction human machine becomes more and more natural. FER are one of the most important ways of providing information about the emotional state, but they are always limited by learning only the six-basic emotion plus neutral. It conflicts with what is present in everyday life, which has emotions that are more complex. This will push researchers in the future work to build larger databases and create powerful deep learning architectures to recognize all basic and secondary emotions. Moreover, today emotion recognition has passed from unimodal analysis to complex system multimodal. Panticet Roth Krantz show that multimodality is one of the conditions for having an ideal detection of human emotion. Researchers are now pushing their research to create and offer powerful multimodal deep learning architectures and databases, for example the fusion of audio and visual studied by Zhang and Ringeval for audio-visual and physiological modalities.

The disclosure of audio-visual meeting recordings is a new challenging domain studied by several large-scale research projects in Europe and the US. [4] Automatic meeting summarization is one of the functionalities studied. In this paper we report the results of a feasibility study on a subtask, namely the summarization of meeting transcripts. A Maximum Entropy based extractive summarization system using a mix of 15 features improved the performance of a baseline system selecting all utterances longer than 10 words with 20% (F-measure). However, stronger contextual awareness seems to be necessary to reduce the precision of the summarizer. The study required the creation of reference extractive summaries, which is documented in the paper.
The study gave some insight into the structure of meetings, showing.Some interesting features that could be used in further research. The approach toclassify each segment individually, without looking at the context is obviously too.naive. Still, our results can function as a reference baseline for comparison with future results.To continue work on this matter, a new approach using lexical chains is investigated.Lexical chains are capable of pinning down topic hotspots in a document and connecting the most important sentences. The use of lexical chaining can be implemented as a whole new method, or as enhancement on the feature set of our current summarization system, e.g. by producing better (context based) estimates of which tokens are topical.

Feasibility Study

A feasibility study for a Video Conferencing Application with advanced features indicates promising potential. There is a market demand for such an application, and it can have a positive economic impact. However, it comes with high development costs and technical challenges. The feasibility study suggests that with a well-planned approach and risk mitigation strategies, this project can be viable and successful in meeting the demands of a competitive market.

Methodology/Planning of Work

The methodology for creating an advanced Video Conferencing Application involves defining objectives, market research, stakeholder engagement, technology assessment, meticulous development, integration, testing, and optimization. User training, legal compliance, and strategic marketing ensure a comprehensive approach. This structured process aims to meet market demands and user expectations effectively.

Facilities Required

To create a robust video conferencing application that caters to a global audience with multilingual capabilities, emotion recognition, sign language interpretation, and auto summarization, a multifaceted approach is required. The application must offer real-time translation services, seamlessly breaking down language barriers. Advanced machine learning models are vital for emotion recognition in video feeds, enhancing non-verbal communication. Incorporating sign language interpretation functionality ensures inclusivity for users who are deaf or hard of hearing. Auto summarization relies on sophisticated natural language processing algorithms to condense meeting content effectively. Distributed data centers across the globe are essential for low-latency, high-availability services, while stringent security measures are necessary to safeguard sensitive information. Efficient bandwidth management and cross-platform compatibility ensure a smooth experience across various network conditions and devices. Scalability, user-friendly design, feedback mechanisms, content management, accessibility features, and compliance with international standards round out the comprehensive facilities required for such a feature-rich application, which necessitates a skilled team and ongoing maintenance for long-term success.

Software Requirements:

1. **Media Processing**: Incorporate video streaming and audio processing libraries for real-time

communication.

2. **Machine Learning Frameworks**: Utilize frameworks like TensorFlow and OpenCV for emotion recognition.

3. **Natural Language Processing (NLP)**: Implement NLP libraries for text analysis and language translation.

4. **Speech-to-Text Services**: Integrate services for converting spoken language into text for auto summarization.

5. **Sign Language Recognition**: Include software or libraries for real-time sign language interpretation.

Hardware Requirements:

1. **High-Performance Computing Resources**: Training deep learning models, especially CNNs, can be computationally intensive. Access to powerful GPUs (Graphics Processing Units) or TPUs (Tensor Processing Units) can significantly speed up the training process.
2. **Sufficient RAM**: Deep learning tasks often require a significant amount of memory, especially when dealing with large datasets. Having enough RAM to accommodate your data and models is important.
3. **Storage**: Medical images can consume a lot of storage space. Make sure you have ample storage to store your datasets and model checkpoints.
4. **Multi-Core Processor**: While a powerful GPU is essential, a multi-core CPU can also aid in tasks that aren't GPU-accelerated.
5. **External Storage or Cloud Services**: If you're dealing with a massive dataset and your local hardware resources are limited, you might consider using cloud services like AWS, Google Cloud, or Azure for storage and computation.
6. **Medical Imaging Hardware:** If you're involved in collecting your own medical imaging data, you might need access to medical imaging devices like MRI machines, echocardiogram machines, etc.
7. **Backup and Redundancy**: Since your data is critical, having a backup system in place is essential to prevent data loss.

Expected Outcomes

The development of this multifunctional video conferencing application addresses the evolving needs of a diverse and interconnected world. By combining multilingual support, emotion recognition, sign language translation, and auto summarization, the application aims to revolutionize the way people communicate, fostering inclusivity, empathy, and efficiency in online interactions.

References

[1] Meetdot Videoconferencing with Live Translation.

[2] Sign Language Recognition: State of The Art.

[3] Facial Emotion Recognition Using Deep Learning: Review and Insights.

[4] Automatic Summarization Of Meeting Data: A Feasibility Study.