2019 IEEE Bombay Section Signature Conference (IBSSC)

Smart e-stick for Visually Impaired using Video Intelligence API

Priyanka Ambawane
Department of Computer Engineering
Maharashtra Institute Of Technology
Pune, India
dearpriyankasa@gmail.com

Devshree Bharatia
Department of Computer Engineering
Maharashtra Institute Of Technology
Pune, India
devshreebharatia1791997@gmail.com

Piyush Rane
Department of Computer Engineering
Dhole Patil College Of Engineering
Pune, India
piyushrane17@gmail.com

***Abstract*—The visually impaired need to confront several problems while performing their daily activities. Independent navigation becomes one of the major problems in their social life. This issue further deteriorates when they travel across new unknown environments. This paper provides an efficient solution for the visually impaired in the form of a hardware automated stick based on Google's Cloud Video Intelligence API. This system uses real time video processing to analyze the obstacles or objects coming in the path of the blind and provides feedback in the form of voice messages. Hence the system facilitates real time navigation in both indoor and outdoor environments. Also the cost effectiveness and the versatility of the system make it more adaptive and easy to use.**

***Index Terms*—Video Intelligence API (Application Programming Interface),smart blind stick, label detection, object detection and recognition, text recognition, shot change detection, Google's text-to-speech engine(gTTS).**

I. Introduction

Visual impairment can limit people's ability to perform everyday tasks and can also affect their ability to interact with the surrounding world. Blindness, the most severe form of visual impairment, can reduce people's ability to move independently. Also moving through a new and an unknown environment becomes a real challenge when one cannot rely on their own eyes. Thus the visually impaired people face a major problem while navigating independently in unknown environments. They have to discover entrances, know the present area, be aware of environmental attributes like footstep sounds and track the entire journey until the goal is reached. In such a scenario, a system that helps to navigate through routes and avoid obstacles would provide an enormous advantage to accomplish this task. Considering the problems of the blind and to improve their social life an innovative low cost walking aid called smart e-stick has been proposed in this paper. The smart e-stick is a simple camera driven automated system for the aid of the visually impaired. The system will provide real-time video monitoring via the camera and real-time video processing using a microcontroller. The video will record the state of the existing surroundings (both indoor and outdoor) and the contents of the surroundings will be analyzed to provide voice feedback to the visually impaired. This video analysis will be done using Google's Cloud Video Intelligence API. The entire system will be implemented via a rechargeable circuit. Hence the smart e-stick system will help to identify objects as well as read text and accordingly provide voice output to the visually impaired. All this processing will be done dynamically, hence providing a sense of vision to the blind.

II. Existing System

The existing smart e-stick systems use variety of sensors for facilitating obstacle detection and navigation of the visually impaired. Some of these sensors include ultrasonic [3], IR (Infrared) [4], RFID (Radio Frequency IDentifier) tags [1], LDR (Light Detection Resistor) etc. The use of these sensors reduces the compactness of the stick and also affects the power consumption requirements. This excessive dependency on large number of sensors and their integrated functioning can prove to be a major drawback of the existing systems. Moreover the obstacle detection mechanism used by the automated e-sticks today can be used only for static objects and not for moving objects. This further poses a limitation and fails to provide a sense of vision to the visually impaired. With further advancements in technology, image processing is being used for object detection and localization [2] for further high-level navigation of the visually impaired. But this technique also involves processing of a static image captured by the camera at that instance of time. Hence such systems are not suitable for dynamic use at run time. Further a limitation of this system is that it is difficult for a visually impaired person to determine when to click the image for processing. Also multiple cameras will be required to sense the state of the environment in multiple directions. This further increases the processing complexity for the microcontroller being used. Hence the walking canes available today are highly inefficient and propose the need for the development of integrated, fully-functional smart systems.

III. Proposed System

Very few of the smart e-sticks available today facilitate real-time route navigation of the visually impaired. But, none of these frameworks facilitate both indoor and open air applications. In order to help the visually challenged people and improvise the existing systems, the automated smart e-stick has been proposed. The proposed system uses a simple

978-1-5386-7401-7/19/$31.00 ©2019 IEEE

camera driven automated e-stick for the aid of the visually impaired. A simple camera module is used for recording real-time video from the surrounding environment. This video recorded can be uploaded on the cloud after small intervals of time and specific, predefined objects in the surrounding can be identified. Google's Video Intelligence API is used for extracting and analyzing only relevant information within the entire video shot-by-shot or per frame. This video recording, uploading and processing will be done parallely on the local system and cloud-based servers, hence increasing the efficiency of the entire system. The proposed system can not only recognize objects but text can also be easily read by the visually impaired in the form of newspapers, magazines, notice boards, documents etc. It provides the above as voice output to the user via speakers or earphones. Hence the first advantage of the proposed system is that it completely eliminates the dependency on sensors like LDR, Ultrasonic, and IR etc. Also the stick is suitable for detecting both static and moving objects in both indoor and outdoor environments. This real time video processing helps to provide a sense of vision to the blind. Also the recorded videos are not permanently stored on the local system. The same clip (of predefined time interval) of the video will be replaced every time a new clip is recorded. Thus this further reduces the memory requirements. Also, the entire circuit will be powered by a rechargeable battery module. Hence the proposed system is much more efficient and much smarter that the conventional stick used by the visually impaired today.

IV. Video Intelligence Working

Video Intelligence is an API service provided by Google to analyze video content and provide optimized results. The videos will be first uploaded on the cloud and stored using the cloud storage. Using video intelligence important and relevant content of the video will be extracted. This meta-data generated will then be stored on the cloud and can be used for various applications. In the system proposed by this paper, this meta-data will be used to provide advanced navigation functionalities for the visually impaired. This basic video intelligence functionality is shown in the figure 2.

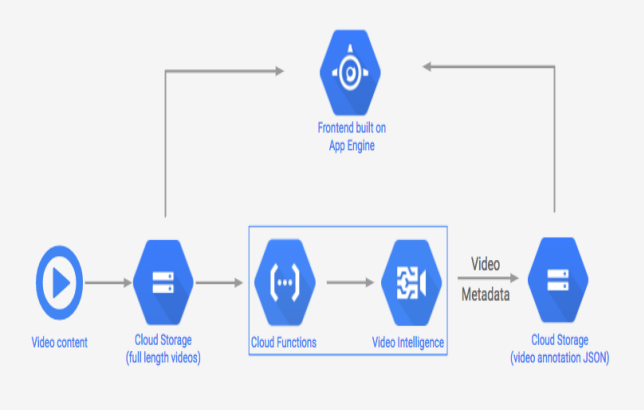

Fig. 1. Video Intelligence Functionality

V. System Architecture

The system architecture represents all the individual modules of the system.The hardware modules include electronic components such as the central processing unit/microcontroller, camera, the speaker and the rechargeable power supply unit.The software unit consists of the various APIs used: Google's text-to-speech translation API and also Google's Video Intelligence API for label,object,text and shot change detection.

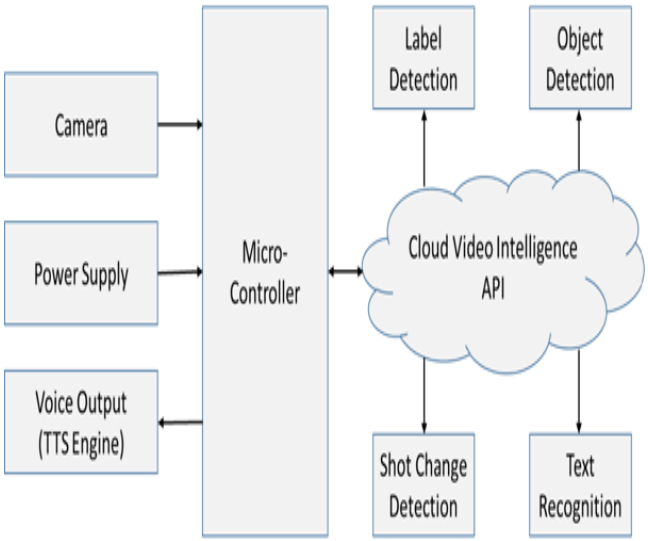

Fig. 2. System Architecture

VI. Functionalities of smart e-stick using video intelligence API

A. Feature 1: Label and Object Detection

This feature helps to provide both high-level as well as low-level description regarding the video to the visually impaired. At high-level it provides a general description about the video contents and at granular level it provides a detailed description about every scene in the video.Several entities and actions from the video can be detected and labelled. The labels can be obtained according to their confidence values.

Step 1: Initially the JSON object is constructed so that a request to use the Video Intelligence API service can be made.The object is then used to call the annotate _video() method.The path to the Google cloud storage location where the video file is stored will be provided in the form of URI(Uniform Resource Information) along with the feature to be performed i.e. LABEL _DETECTION.

978-1-5386-7401-7/19/$31.00 ©2019 IEEE

```
video_client =
videointelligence.VideoIntelligenceServiceClient()

features =
[videointelligence.enums.Feature.LABEL_DETECTION]

operation = video_client.annotate_video(path,
features=features)
```

Fig. 3. Request Construction for the API

The structure of the JSON object used is as follows:

JSON representation

```
{
  "inputUri": string,
  "inputContent": string,
  "features": [
    enum(Feature)
  ],
  "videoContext": {
    object(VideoContext)
  },
  "outputUri": string,
  "locationId": string
}
```

Fig. 4. Structure of JSON object

Step 2: Using the existing operation request for our existing operation, we periodically check the state of that operation. Once our operation has indicated that the operation is done, we can parse the response.

```
result = operation.result(timeout=90)
print('\nFinished processing.')
```

Fig. 5. Checking the operation

Step 3: Once the operation has been completed, the results will be present in segmentLabelAnnotations. We then loop through all the labels in segmentLabelAnnotations and extract the following information:

- Label Description: A textual description of the label using segment _label.description
- Label Category: A list of entity categories using category _entity.description
- Confidence Score: A number associated with every returned label which determines its accuracy/relevancy. Confidence scores range from 0 (no confidence) to 1 (very high confidence).
- Timestamp: A time segment with the time offset for the entity's appearance from the beginning of the video.

```
segment_labels =
result.annotation_results[0].segment_label_annotations
for i, segment_label in enumerate(segment_labels):
    print('Video label description: {}'.format(
        segment_label.entity.description))
    for category_entity in segment_label.category_entities:
        print('\tLabel category description: {}'.format(
            category_entity.description))

    for i, segment in enumerate(segment_label.segments):
        start_time = (segment.segment.start_time_offset.seconds +
                      segment.segment.start_time_offset.nanos / 1e9)
        end_time = (segment.segment.end_time_offset.seconds +
                    segment.segment.end_time_offset.nanos / 1e9)
        positions = '{}s to {}s'.format(start_time, end_time)
        confidence = segment.confidence
        print('\tSegment {}: {}'.format(i, positions))
        print('\tConfidence: {}'.format(confidence))
```

Fig. 6. Parsing the response

Step 4: The final result will be as follows:

```
Video label description: urban area
        Label category description: city
        Segment 0: 0.0s to 38.752016s
        Confidence: 0.946980476379


Video label description: traffic
        Segment 0: 0.0s to 38.752016s
        Confidence: 0.94105899334


Video label description: vehicle
        Segment 0: 0.0s to 38.752016s
        Confidence: 0.919958174229
```

Fig. 7. Final Output Format

The label annotation can be done in two different modes:

- SHOT _MODE: By default this mode is used where segment-level annotation will be performed.If segments are not specified, the API will treat the video as a single segment. If more specific results are to be obtained, video has to be divided into segments and they can be passed to the videoContext field in the annotate request.

- FRAME _MODE: Here, frame-level annotations will be processed. This is a costly option, as it analyses all the frames in the video and annotates each of them, but it may be a suitable option depending on your specific use case.

In the smart e-stick, the video will be uploaded by the microprocessor either in the form of frames or segments depending upon the mode to be used. Hence the video recording, video uploading and the annotation tasks will be carried out parallely thus optimizing the overall performance of the system.

This label and object detection feature thus becomes helpful in providing a sense of vision to the blind regarding their surroundings. It helps to label entities such as traffic signals, road-side signs, danger signs, zebra crossing etc for advanced navigation of the visually impaired. According to the confidence values, the most prominent output having confidence score say greater than or equal to 0.75 can be provided to the blind using voice commands.

B. Feature 2: Shot Change Detection

Any scene changes within the video being recorded can be immediately identified and notified to the visually impaired as and when required. This will further provide a closer, in-depth analysis of the changing environmental conditions in the surrounding to the blind person.

By default the Video Intelligence API examines a video or video segments by frame. That is, each complete picture in the series that forms the video. The Video Intelligence API can also be used to annotate a video with video segments that are selected based on content transition (scenes) as opposed to the individual frames.

For example, a golf video following two players across the golf course with some panning to the woods for background may produce two shots: ”players” and ”woods,” giving the developer access to the most relevant video segments showing the players for highlights.

To detect shot changes in a video, call the annotate _method() and specify SHOT _CHANGE _DETECTION in the features field.

C. Feature 3: Text Detection and Recognition (OCR)

Text can be detected and extracted from the recorded video or video segments by calling the annotate _method() and specify TEXT _DETECTION in the features field.This feature will help the visually impaired to read both machine-printed as well as handwritten text from newspapers, notice boards, road-side signs etc. The text can then be provided as output using voice commands.

Optical character recognition (OCR) is an algorithm used to extract the text from each clip of the video so that its content can be further analyzed and produced as output in the voice format.The following two steps are involved in the process of Optical Character Recognition:

Step 1: Pre-processing

In this step the artefacts from the image will be eliminated and only the text will be taken into consideration. It involves removing the various graphics from the image, aligning the text properly and converting any shades of colours into black and white only. Binarization is the process used for this purpose where the text snapshot is represented in terms of 0s and 1s and the output will be in the form of connected components.

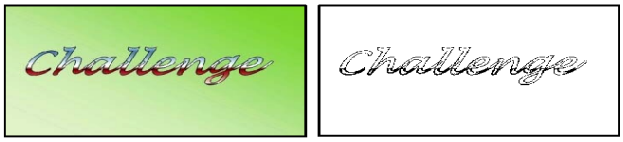

Fig. 8. Input Image and Output Image after Binarization.

Step 2: Character recognition using classification by alphabet and font

A smart approach used for character recognition is breaking down each character into constituent elements like curves and corners and then matching these physical features to actual letters. Also each scanned character will be compared pixel by pixel to a known database of fonts. The numerical probabilistic values may be obtained on the basis of goodness of fit and accordingly the text can be recognized. This algorithm is not only used for recognizing printed text but can also be used for handwritten text.

VII. Proposed System Implementation with Flowchart

The smart e-stick uses video intelligence API to provide three different types of functionalities. Hence three buttons are provided on the stick for video-level, shot-level and frame-level annotations.A separate button for text recognition using OCR will also be provided.

- Video-level: This functionality is a highly intelligent functionality which helps to provide a general description for the complete video.For this purpose it integrates the labels generated in the individual frames to obtain the overall context of the video. Caption generation algorithms can be used to generate meaningful captions for the video in integration with this functionality.
  Example: This functionality can be used when the context of the video remains the same like when the visually impaired person enters a corporate office, all the people will be working on their respective computers. In such a scenario it is better to use video-level annotation for providing efficient output to the blind.
- Shot-level: This functionality is useful when the scenes in the recorded video are continuously changing. In such a scenario video-level annotation will not add a lot of value. Example: When the visually impaired person wants to watch a movie or an advertisement, the scenes will

change after every few minutes or seconds. Hence shot-level functionality will help in providing them an in-depth analysis of the movie or advertisement.
- Frame-level:This functionality is useful when further in-depth frame-level analysis is required.

VIII. Interfacing Module: gTTS Engine

The Google Text to Speech API commonly known as the gTTS API is used to convert text to speech in python. It is a very easy to use tool which converts the text entered, into audio which can be saved as mp3 file. The output of the in-depth video intelligence analysis helps in the generation of captions and labels for the various recognized objects. It also helps in text extraction. This output is converted from text into audio using the gTTS engine and voice output will be provided to the blind using earphones or speakers.This API also provides multi-language support and hence the voice output can be generated in several local languages.

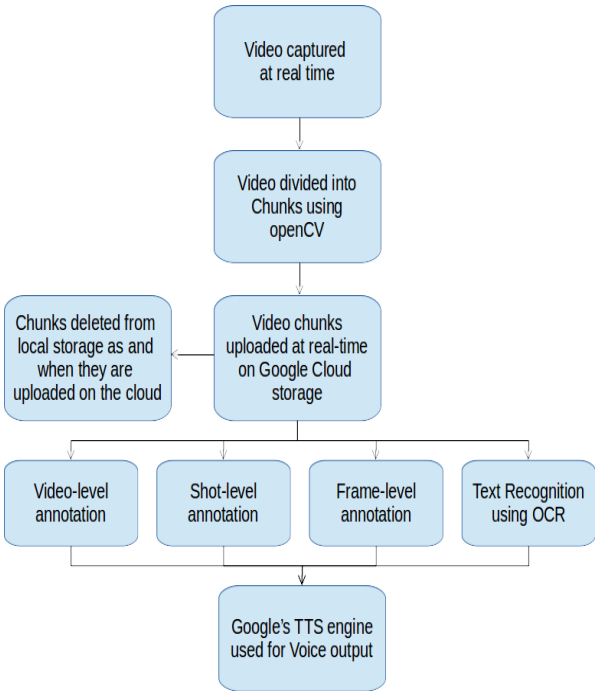

Fig. 9. Flowchart representing the Workflow

IX. Advantages

- A microcontroller driven camera module is being used in the system and no other sensors are required for the navigation of the blind. Hence this reduces the overall cost and size of the e-stick.
- The real-time video processing is being done entirely on the cloud. Hence the memory requirements of the system are reduced drastically.
- Further, this stick behaves as one of the most efficient applications of Google's Video Intelligence API. Use of Google's large database, facilitates highly accurate object and text recognition.
- Also the recorded videos are not permanently stored on the local system. The same half an hour clip of the video will be replaced every time a new clip is recorded. Thus this further reduces the memory requirements.
- Quick search using video intelligence API helps to identify objects at a faster speed. It also returns confidence levels for each entity identified, so only the relevant content can be extracted at a faster rate.
- Image attributes feature detects general attributes like dominant colours, shape etc for in-depth analysis.

X. Limitations

- As the processing is done real-time on cloud, the limitation would be its excessive dependence for high speed internet/ high speed Wi-Fi.
- Also, the charges applied by Google for the use of its API pose as a limitation.
- Further, this stick cannot be used by deaf and dumb people as the output is provided in the form of voice.
- If no internet connection is available, the smart system will not be able to function and provide object recognition.

XI. Conclusion

Hence this paper proposes a stick for the blind which is more useful and much smarter than the conventional stick used by them today. The implementation of this system can bring about great changes in the daily lives of the visually challenged. This system has the prototype that can identify the object and obstacle from the video and feeds warning back in the form of message and voice. The cost effectiveness and the versatility of this system make it more adaptive and easy to use. Also, as the real-time processing of the recorded video is done on the cloud the load of the existing local system is reduced considerably. Thus, this paper presents an approach to develop a project which would help and benefit the society using the current trending technologies.

XII. Future Scope

The smart e-stick can be further enhanced by using solar cells so that the user need not bother about recharging the e-stick. Moreover, all the smart e-sticks can be inter-connected using wireless sensor networks so that the blind can easily communicate with the other people in their community. Also caption generation algorithms for the videos captured can further improvise the voice outputs for real-time navigation purposes. In this way, the future scope of the system can be combined with our current system to improve the system utilities.

References

[1] Madhura Gharat, Rizwan Patanwala, Adithi Ganaparthi, “Audio guidance system for blind”, International Conference on Electronics, Communication and Aerospace Technology, 2017.
[2] Prof. Priya U. Thakare, Kote Shubham, Pawale Ankit, Rajguru Ajinkya, Shelke Om, “Smart Assistance System for the Visually Impaired”, International Journal of Scientific and Research Publications, Volume 7, Issue 12, December 2017.

[3] Zeeshan Saquib, Vishakha Murari, Suhas N Bhargav, "BlinDar: An Invisible Eye for the Blind People", IEEE International Conference On Recent Trends In Electronics Information Communication Technology, India, May 2017.
[4] Ayat Nada, Samia Mashelly, Mahmoud A. Fakhr, and Ahmed F. Seddik, "Effective Fast Response Smart Stick for Blind People", Research Gate, Conference Paper, April 2015.
[5] D.Sekar, S.Sivakumar, P.Thiyagarajan, R. Premkumar, M. Vivek kumar, "Ultrasonic and Voice Based Walking Stick for Blind People", Internation Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 4, Issue 3, March 2016.
[6] Mouhamad D. Mashat, Abdulaziz A. Albani, "Intelligent Blind Cane System", UKSim-AMSS 19th International Conference on Modelling Simulation, 2017.
[7] https://cloud.google.com/video-intelligence/ (Last accessed: 10/04/2019)
[8] https://www.youtube.com/watch?v=mDAoLO4G4CQ (Last accessed: 05/04/2019)
[9] https://cloud.google.com/video-intelligence/docs/label-tutorial (Last accessed: 05/04/2019) [10] https://www.youtube.com/watch?v=2EMpylXjAmI (Last accessed: 07/04/2019)
[10] Sunil Kanzariya, Prof. Vishal Vora, "Real Time Video Monitoring System Using Raspberry Pi", National Conference on Emerging Trends in Computer, Electrical Electronics, 2015.
[11] Lun Zhang, Stan Z. Li, Xiaotong Yuan and Shiming Xiang, "Real-time Object Classification in Video Surveillance Based on Appearance Learning", IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[12] Dr.M. Senthamil Selvi, Mrs. J. Angel Ida Chellam, "Smart Video Surveillance: Object Detection, Tracking and Classification", International Journal of Innovations Advancement in Computer Science, March 2018.
[13] https://pythonspot.com/speech-recognition-using-google-speech-api/ (Last accessed: 02/04/2019)
[14] Luis Valentn , Sergio A. Serrano, Reinier Oves Garca, Anibal Andrade, Miguel A. Palacios-Alonso, L. Enrique Sucar, "A Cloud-Based Architecture for Smart Video Surveillance", 2nd International Conference on Smart Data and Smart Cities, October 2017.

978-1-5386-7401-7/19/$31.00 ©2019 IEEE