AID FOR THE VISUALLY IMPAIRED

[1]Aishwarya S. Padmanabhan, [2]Anjali Unnikrishnan, [3]Anupam Sharma, [4]Apoorva Shetty
Undergraduate Student
Department of Computer Science Engineering
B.M.S. College of Engineering

Abstract— This project aims at using principles of Image Processing and Artificial Intelligence to detect and identify objects. The module must convey what the identified object is to the visually impaired user using an Image to Text, and Text to Speech synthesizer pipelined together. The three phases of the module are Detection, Identification, and Communication.

Index Terms— Image Processing, Visually Impaired, Computer Vision, GPS module, Auditory and Pressure Sensors, Arduino Board

I. INTRODUCTION

Several projects in the field of Image Processing work towards aiding those with a visual impairment, like the Indoor Navigational Aid System[1], Smart Stick for Blind[2], etc. Our module aims at emulating certain aspects of these projects to create an object detection and identification module that will aid the visually impaired in basic day-to-day activities.

To provide basic aid for the visually impaired, there are several questions that need to be answered for example: How will the module communicate with the Visually Impaired, What methodology or algorithm is best suited for real time detection, How large must the data set be to accommodate the database for identification, etc. This paper aims at answering basic fundamental questions for developing our module, and to look at previous projects and papers and try to implement the best tried and tested method to create our module. It also tries to minimise error of previously completed similar projects.

The research for this project can be divided into three distinct sections:

Object Detection
Object Identification
Communication with the User

II. OBJECT DETECTION

The first step towards achieving the goal of our project is to detect objects placed in front of the module. There are multiple ways to go about this, using cameras, sensors – pressure or auditory.

Sensor Based Assistance System for Visually Impaired[3] focuses on obstacle detection using sensors and arduino board on a regular long cane. The signal reflected from an obstacle is collected by sensors, which use the arduino board to process the signals. According to the processed data, a message is invoked from the flash memory. Distance of obstacle is calculated using ultrasonic sensors and different materials are recognised using IR transmitter along with Light Dependent Resistor (LDR).

Smart Stick for Blind[2] consists of proximity sensors, ultrasonic sensors, a GPS module, and stereo cameras. For calculating distance, they have used ultrasonic or laser beams array, which are reflected by the objects and the system calculates the distance using the time difference between emitted and received light. The Artificial Vision unit of this project uses stereo cameras and processing unit. The unit extracts data about static and dynamic objects using image processing algorithm. The hardware of the system is based on a pair of stereo cameras mounted on a helmet connected via connectors to a portable computer. The system speed achieves 25 frames per second due to the special Express card that allows external power supply. Their system is able to detect objects like humans, cars, building, trees that are at a distance between 1 and 15m

Darshan: Electronics Guidance For The Navigation Of Visually Impaired[4] - In this paper they have used ultrasonic sensors and USB camera. The system detects the obstacle using ultrasonic sensors and sends feedback in form of beep sound via earphone. To find humans they use a camera, if the camera cannot find the person then they use cloth reflective index to tell if a human is present or not. This project includes detecting organic matter as separate from detecting inanimate objects. However, our module just aims to detect inanimate objects as identification of every individual human would include a database that is far too large to be encompassed in available hardware.

The Object Detection Module (ODM) in the camera based assistive aid for Visually Impaired[5] uses a HC-SR04 sensor. They took the Average length of the stride of a person as 0.7 m to 0.8m. Hence, the module is programmed to start beeping whenever there is an obstacle within 100 cm of the user. This module only works for obstacles that are above knee level.

Electronic travel aids and electronic orientation aids for blind people: technical, rehabilitation and everyday life points of view, Tom Pouce[6] used optical devices to do the detections. Several near infrared beams are generated by collimated LEDs, in different directions and at different emission powers. Detectors receive light scattered by the objects.

Recognition of Moving Objects by Image Processing and its Applications to a Guide Robot[7] developed a robot with CCD camera, a PSD sensor and supersonic sensors. Their robot can detect Braille block using detection and recognition algorithm using vision sensor; and static objects using PSD sensor.

They input the image to min-max operation and then decompose the area (area enlargement) which is used to detect the road area. For detection of moving objects, luminosity data and Otsu methods are used to obtain the binary image of the difference between two succeeding images. Moving object may not be detected only with simple binary process. When there is no moving object, the static background is obtained by binary image. Furthermore, in comparison with the binary image of the difference image at the former frame, a white (luminosity 255) pixel in a common position is erased. The domain after the movement (the current binary difference image) can be extracted [8], [9]. The closer a moving object approaches, the wider the difference is in binary pixel area. Hence they set a label "risky" to those binary pixels. Largest "risky" moving object in the image is most dangerous if it is coming close to the robot and the robot has to avoid this risky moving object. Therefore, the largest risky moving object is assumed to be the OSAP (The OSAP is characterized by the three parameters: Height, Width, and the coordinates of the Judgement pixel of the OSAP. The risky degree of each pixel of detected object is

evaluated. The pixel with highest value is most risky pixel). And the robot has to expect the direction of motion of the object. Then a judgement is made on whether the object is coming closer or moving away and a safe zone is decided. The robot then moves to the safe zone.

After thorough analysis, we have decided to use a still camera to capture images that can be scaled and rotated using affine transformations. We will then use the concepts of edge or region based segmentation, depending on their performance with respect to the images. We aim to analyze the picture for depth perception, however we may use an ultrasonic sensor to improve accuracy.

III. OBJECT IDENTIFICATION

Once the object has been detected, identification can be done using a multitude of concepts or algorithms as outlined in several papers and projects.

Building Recognition System Based on Deep Learning[10] explains the use of Deep Learning algorithms in image recognition. Convolutional neural networks are successful in image recognition tasks. CNNs comprise of convolutional and pooling layers and a classification layer for the output of the neural network.The convolutional layer includes maps and has several parameters. The activation over the non-overlapping rectangular regions determines the output of the pooling layer. The classification layer consists of one neuron per target class as the last layer of the network.

The TRNAVA LeNet model was first trained to obtain a predictive accuracy of 50.96%. The final model TRNAVA LeNet 10, obtained a predictive accuracy of 81.41%. The image data is processed by the Inner Product layer, which transforms the simple input vector into a single vector output. The image data is then processed by the convolutional layer, which matches the input image with a set of learnable filters, each producing one feature map in the output image. The pooling layer reduces the image size to reduce computation and control overfitting. Transformations like feature-scaling and mean-subtraction were used for data flow through the model. Unlike other approaches, this model takes a high dimensional set of features. Areas of improvement include reduction in error caused due to blurring and low lighting.

Image recognition using a Quadratic Convergence Learning Algorithm of Synergetic Neural Network[11] presents a learning algorithm of synergetic neural network which can be realised using a 2-layer architecture. The main feature of this algorithm is its high accuracy and quadratic convergence rate. This enables the synergetic neural network to recognise images more efficiently. The initial steps of the algorithm involve weight update methods and weight adjustment as the learning process until the error is minimum. They show that the convergence is faster when the learning rate is within a specified range. The new algorithm proposed in this paper needs only several epochs to get high accuracy even with a large initial system error.

The Currency Recognition Module (CuRM) of a Novel Approach to a camera based assistive aid[5] which helps the user to identify denomination of indian currency uses input in the form of an image of the paper note(which is sent to the Currency Recognition Module), to which user is pointing, and then clicked by the on-device camera. Bag Of Words algorithm is used to implement the module; SIFT descriptor is used to retain the key point features [12]. Bag of Words algorithm helps in plotting a histogram of occurrence of key points. These histograms are then send to 1-v/s-all SVM classifier. The vocabulary trained is 1000 feature-long. The classifier is trained on 1000 images and tested on 700 images.

A Vision Based Method for Object Recognition[13] This paper focuses on splitting an image into N*N image blocks and extracting robust SIFT (Scale-Variant Feature Transform) from each block. The recognition is a two-stage process, training and testing stage. In the first stage, sample images are collected and SIFT features are extracted from them before sending to train classifiers. In the testing stage, new samples are collected, features are extracted and using the trained classifiers, class labels of the testing samples are predicted.

The Methods for feature Extraction mentioned in this paper are:

Pixel magnitude and Orientation Histogram

Find magnitude and orientation of each pixel and produce a histogram to represent the image gray values with orientation on the x-axis and magnitude of the y-axis. The orientations are divided into several bins, and for each orientation the magnitude is accumulated into the corresponding bin of the orientation histogram.

Image Block Orientation Histogram

Each image block produces an orientation histogram. These histograms can be concatenated to obtain the SIFT features (vector). If the number of histograms are L, the dimension of SIFT vector is N*N*L. The features are then sent to trained classifiers for recognizing objects.

The training classifiers used for recognition in the abovementioned process are:

KNN classifier

Class label is determined by those of its k nearest neighbours in the training samples. If the number of nearest neighbours from one class is largest, the testing image is assigned to the class label.

Naive Bayes Classifier

Based on the Bayesian decision rule and the following two assumptions: the features of every class are independent; the probability distribution of every class is a Gaussian function.

Decision Tree

The training of the decision tree classifier is to construct a tree on the basis of the training samples. At the root node, all the training samples are input. Then, according to one feature, the sample travels forward along different branches. The samples from the same class should walk as together as possible along the same branch. When it comes to the leaf node, all the samples are from the same class.

SVM (Support Vector Machine) Classifier

This is a linear classifier that attempts to find a hyperplane to split the samples of two classes as well as possible. The hyperplane has the maximum margin to the samples. If the samples cannot be split linearly in the low-dimensional space, it can project them into a high-dimensional space with the kernel functions so that they can be split by a hyperplane.

The paper Object Recognition using Neural network with Optimal feature Extraction[14] shows how the process of object recognition can be structurally divided into two independent subtasks: feature extraction and object recognition. Feature extraction, involves obtaining the most useful features from the training set to reduce dimensionality and improve the recognition process. The main method used for feature extraction is Principle Component Analysis(PCA) which has optimal properties among other feature extraction methods. Next these dominant features are used to train a feed-forward neural network (multilayer perceptron network or MLP) for object recognition.

So, from the above research, the methodology best suited for object recognition is implementing optimal feature extraction to train a multilayer neural network. The advantage of using feature extraction is to overcome curse of dimensionality by extracting only the useful

features. It also reduces the time required to train the neural network. We aim at gaining high accuracy by using a neural network to minimise the error in recognition.

IV. COMMUNICATING WITH USER

Communication with the visually impaired can be of two basic varieties; Auditory or Vibrations. For certain projects the use of vibration sensors fit better than auditory warnings (Object Detection and Navigation systems), whereas certain systems require an auditory communication module (Object/ Image Identification).

For Example the Smart Stick for the Blind[2] use a dual feedback mechanism i.e. an Auditory as well as a vibratory circuit, This enhances the overall feedback received by the blind user who receives the outputs generated in different formats of vibration i.e high, low, medium and strong vibrations.

Whereas Darshan: Electronics Guidance For The Navigation Of Visually Impaired[4] system detects the obstacle using ultrasonic sensors and sends feedback in form of beep sound via earphone.

The "Teletact"[15][16] is a hand held laser telemeter. To communicate the distance information to the user, they used tactile or audio interface. For tactile there are two vibrating devices located on two different fingers. The first finger codes the distances between 3 and 6 meters by a discrete vibration and the distances between 1.5 and 3 meters by a strong vibration. The second finger is only concerned with the distances under 1.5m, i.e. the alerts. For the audio interface, the distance is coded up to 15 meters. 28 different musical notes correspond to 28 unequal distance intervals (intervals are shorter for short distances), the higher the tone, the shorter the distance. Combining perceptions of the musical notes and from proprioception gives the user information about the shape of the obstacle and the position of the user's body relative to it.

The TBE module of a Camera based assistive aid for the visually Impaired[5] uses a Text to Speech converting module to convert analysed Text Blocks into a final speech output.

Novel indoor navigation system for Visually Impaired and blind people[1] makes use of a ceiling camera which tracks the individual by placing a marker on his head to determine the person's location/co-ordinates in the current environment. The individual can communicate with the system via a microphone (speech to text) and commands from the system are transmitted via speaker (text to speech).

Image Parsing to Text Description[17] proposed an image parsing to text description that generates text for images and video content. Image parsing and text description are the two major tasks of this framework. It computes a graph of most probable interpretations of an input image. This parse graph includes a tree structured decomposition contents of scene, pictures or parts that cover all pixels of image.

Over past decade many researchers from computer vision and Content Based Image Retrieval (CBIR) domain have been actively investigating possible ways of retrieving images and videos based on features such as color, shape and objects[18][19][20][21][22].

Heterogeneous Domain Adaptation and Classification by Exploiting the Correlation Subspace[23] presents a novel domain adaptation approach for solving cross domain pattern recognition problem where data and features to be processed and recognized are collected for different domains

A Novel Substitute for the Meter Readers in a Resource Constrained Electricity Utility[24] introduced a model of image to text conversion for electricity meter reading of units in kilowatts by capturing its image and sending that image in the form of Multimedia Message Service (MMS) to the server.

The server will process the received image using sequential steps:

1. Read the image and convert it into a three dimensional array of pixels.
2. Convert the image from color to black and white.
3. Removal of shades caused due to non uniform light.
4. Turning black pixels into white ones and vice versa.
5. Threshold the image to eliminate pixels which are neither black nor white.
6. Removal of small components.
7. Conversion to text.

The paper on HMM based speech synthesis systems[25] discusses one of the most popular techniques of speech synthesis, a statistical parametric speech synthesis based on Hidden Markov Models(HMM). In this system, context-dependent HMMs are trained from databases of natural speech and even the speech waveforms are generated from the HMMs itself. It consists of two parts: training and synthesis. The training involves extraction of the spectral and excitation parameters from the speech database, which is then modeled by the HMM. The synthesis part does the inverse of training. The text to be synthesized is first converted to a context dependent label sequence. Then an utterance HMM is constructed by concatenating the context-dependent HMMs according to label sequence.

Several papers talk about using Hidden Markov Models to convert text to various forms of speech, for example Text-To-Audio-Visual Speech Synthesis Based On Parameter Generation From HMM[26] describes a technique for synthesizing auditory speech and lip motion from arbitrary text. Some papers have tried to improve on this model, such as the paper on A Bayesian Approach To HMM-Based Speech Synthesis[27]. The Bayesian method is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters. In the proposed framework, all processes for constructing the system can be derived from one single predictive distribution, which represents the basic problem of speech synthesis directly. Using HMM as the likelihood function and assuming some approximations, it can be regarded as an application of the variational Bayesian method to the HMM-based speech synthesis. Experimental results show that the proposed method outperforms the conventional one in a subjective test.

As our project aims to convey the type of object in front of the module, not just the presence of it we will need a final output of speech, after the identification process is complete. Thus we will need an Image to text, and a Text to speech convertor connected in series to get our required output. The Image to text module will be achieved by the second phase (Object identification), which will then be pipelined with a text to speech convertor that will use the principles of Markov chains (HMM) to generate speech waveforms from the text.

V. CONCLUSIONS

We have finalised the methodology to achieve the three phases of our project: Object Detection, Object Identification and Communication with the User with the least possible errors.

For our first phase we will use a still camera for edge detection and depth perception. To improve the accuracy of the depth perception, we may use ultrasonic sensors. Captured still images will be refined using affine transformation, and analysed using concepts of edge or region based segmentation.

For object identification we will use the principles of neural networks and implement optimal feature extraction, to assess patterns and provide the most accurate identification. This method not only minimises error, it also requires the least training time for the algorithm to recognise objects.

Our method of Communication with the user will be auditory, and we will implement image to text to speech conversion to relay the identity of the object in front of the module to the user. The image to text output will be generated in the object identification phase of our project, and the principles of Markov chains will be used to generate speech waveforms.

VI. ACKNOWLEDGMENT

This Project is aided by B.M.S. College of Engineering, and is guided by Prof. Rekha G. S., Associate Professor in the Computer Science Engineering Department. The authors would like to thank their guide and their college for giving them the opportunity to work on this project.

REFERENCES

[1] Kabalan Chaccour and Georges Badr, "*Computer vision guidance system for indoor navigation of visually impaired people"* , IEEE 8th International Conference on Intelligent Systems, 2016

[2] Shruthi Dhambare and Prof. A. Sakhare, "*Smart stick for Blind: Obstacle Detection, Artificial vision and Real-time assistance via GPS"*, 2nd National Conference on Information and Communication Technology (NCICT) 2011, Proceedings published in International Journal of Computer Applications® (IJCA)

[3] Vigneshwari C, Vimala V, Sumithra G, "*Sensor Based Assistance System for Visually Impaired,"* International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 10 - Oct 2013

[4] Marut Tripathi, Manish Kumar, Vivek Kumar and Warsha Kandlikar, "*Darshan: Electronics Guidance For The Navigation Of Visually Impaired Person,"* international journal for research in applied science And engineering technology (ijras et)

[5] Ishaani Mittal, Apoorva Mittal, S Indu, *A novel approach to a camera based assisstive aid for Visually Impaired,* 3rd International Conference on Recent Advances in Information Technology 2016

[6] René Farcy, Roger Leroux, Alain Jucha, Roland Damaschini, Colette Grégoire, Aziz Zogaghi, *electronic travel aids and electronic orientation aids for blind people: technical, rehabilitation and everyday life points of view,* Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments Technology for Inclusion – 2006

[7] Katsumi Moriwaki, Yuki Katayama, Katsuyuki Tanaka, Takahiro Shinkyo, *Recognition of moving objects by image processing and its applications,* Proceedings of IEEE conference 23-25 Nov 2009

[8] K. Moriwaki, et al., "Recogniton of Moving Objects by Image Processing and its Applications to Machine Control", *Proceedings of 2010 International Symposium on Flexible Automation,* paper no. JPL2544, pp.1–7 2010.

[9] K. Moriwaki, "Recognition of Moving Objects and Safe Zone by Image Processing", *Preprints of the 5th IFAC Symposium on Mechatronic Systems,* pp. 695–700, 2010.

[10] Pavol Bezak, *Building Recognition System Based on Deep Learning,* Slovak University of Technology in Bratislava Advanced Technologies Research Institute, MTF

[11] R. Q. Sheng, Hong Qiao, Bing Chen, *Image Recognition Using a Quadratic Convergent Learning Algorithm of Synergetic Neural Network,* Proceedings of the International Conference on Robotics, Intelligent Systems and Signal Processing China - October 2003

[12] D. G Lowe, *Object recognition from local scale-invariant features In Computer vision*, 1999. The proceedings of the seventh IEEE international conference on, pages 1150–1157, 1999.

[13] Yiting Zhang and Jianning Liang, *A Vision based Method for Object Recognition,* 3rd International Conference on Information Science and Control Engineering - 2016

[14] Jiann-Der Lee, *Object Recognition Using a Neural Network with Optimal Feature Extraction,* Mathl. Comput. Modelling Vol. 25, No. 12, pp. 105-117, 1997

[15] Farcy R., Damaschini R. (1997), *"Triangulating laser profilometer as three-dimensional space perception system for the blind"*, Applied Optics, vol. 36, pp. 8227-8232.

[16] Farcy R., Damaschini R. (2000), *"Guidance – Assist systems for the blind"*, EBIOS 2000, Amsterdam,3-5 Juillet

[17] Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu, "*I2T: Image Parsing to Text Description*" IEEE Conference on Image Processing, 2008 .

[18] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "*Content-based image retrieval at the end of the early years*," IEEE Transactions PAMI, vol 22, no. 12, 2000.

[19] Y. Rui, T. S. Huang, and S. F. Chang, "*Image retrieval: Current techniques, promising directions, and open issues,"* Journal of Visual Communication and Image Representation, vol. 10,1999.

[20] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "*Contentbased multimedia information retrieval: State of the art and challenges,"* ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 2, no. 1, pp. 1-19, Feb. 2006.

[21] C. Snoek and M. Worring, *"Multimedia video indexing: A review of the state-of-the-art,"* Multimedia Tools Appl, vol. 25, no. 1, 2005.

[22] R. Datta, D. Joshi, J. Li, and J. Z. Wang, *"Image retrieval: Ideas, influences, and trends of the new age,"* ACM Computing Surveys, vol. 40, no. 2, pp. 1-60, Apr. 2008.

[23] Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang, *"Heterogeneous Domain Adaptation and Classification by Exploiting the Correlation Subspace,"* IEEE Transactions on Image Processing, vol. 23, no. 5, May 2014.

[24] S. Shahnawaz Ahmed, Shah Muhammed Abid Hussain and Md. Sayeed Salam, *"A Novel Substitute for the Meter Readersin a Resource Constrained Electricity Utility"* IEEE Trans. On Smart Grid, vol. 4, no. 3, Sept. 2013.

[25] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, Keiichi Tokuda, *The HMM-basedSpeech SynthesisSystem (HTS) Version 2.0*, 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007.

[26] Masatsune Tamura, Shigekazu Kondo, Takashi Masuko, and Takao Kobayashi, *Text-To-Audio-Visual Speech Synthesis Based On Parameter Generation From HMM,* Auditory-Visual Speech Processing Australia 1998

[27] Kei Hashimoto, Heiga Zen, Yoshihiko Nankaku, Takashi Masuko, Keiichi Tokuda, *A Bayesian Approach To HMM-Based Speech Synthesis,* Accoustics, Speech and Signal Processing 2009