An Automatic Approach for Translating Simple Images into Text Descriptions and Speech for Visually Impaired People

Mrunmayee Patil
Department of Computer Engineering
Dr.D.Y.Patil SOET
Lohegaon,Pune.

Ramesh Kagalkar
Assistant Professor Department of Computer Engineering
Dr.D.Y.Patil SOET
Lohegaon,Pune .

ABSTRACT
Image processing is a rapidly growing field of research. Images are of different file formats and of different things, places, humans, scientific, astrological and many such. An image is a collection of several pixels arranged in rows and columns. These images are captured, processed and stored for various uses. For common people it is very easy to identify and analyze general images but for the blind and physically disabled people it is difficult. Unfortunately, there is no prior medium or interface for such needy people to communicate with the world. Blind or visually impaired people are usually those people who are neglected by the society, so there is always a need to help such people. Hence, we propose a new technique of converting images into text as well as speech using techniques provided by image processing like pre-processing, image segmentation, edge detection, object detection and speech synthesis. In this paper we first introduce image to text conversion need for blind people and system overview of image to text and speech conversion system. Edge detection plays an important role in this system where Canny edge detection algorithm is used to detect objects from images. Object recognition is done on the basis of color, size, texture and shape of the object.

Keywords
Image Processing; Image Segmentation; Speech Synthesis; Text to Speech Conversion; Edge Detection.

1. INTRODUCTION
The field of image processing is widely used in many areas of research. Common people can easily sense what all is going around them and can view all that is present in this universe. Blind people or visually impaired people find it difficult to interact with the world and they cannot exactly sense the things so they need some or the other human intervention. There are around 15 million blind people in India. In order to make provision for such people there is need of converting images into text and speech However, there is a need to develop an interface for such visually disabled people to communicate with the world. The proposed system makes a better provision for converting captured images as well as stored images to be converted into text and speech. In this conversion process there are various techniques used such as image pre-processing, image segmentation, edge detection and text to speech synthesis. Step by step execution of these techniques helps to achieve the final output. Input to the system is an image and final output is speech output.

2. RELATED WORK
Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu [1] 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 his 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 form 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[2][3][4][5][6]. Paper [7] introduced by Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang 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.

S. Shahnawaz Ahmed, Shah Muhammed Abid Hussain and Md. Sayeed Salam [8] introduced a model of image to text conversion for electricity meter reading of units in kilo-watts 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 nonuniform 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.

In [9] Fan-Chieh Cheng, Shih-Chia Huang, and Shanq-Jang Ruan gave the technique of eliminating background model form video sequence to detect foreground and objects from any applications such as traffic security, human machine interaction, object recognition and so on. Accordingly, motion detection approaches can be broadly classified in three categories: temporal flow, optical flow and background subtraction.

CBIR(content based image retrieval) system enables digital images to be processed and used to extract the features vector on the basis of low level properties of the image. Color, texture and shape are to be considered. A solutions to this is given in [10].

Another approach in feature extraction is given in analyzing edges of image. Jain and vailaya [11] used edge direction method to build edge direction histogram firstly we have to find edges of image and then quantize them. It had limited performance.

Then Shandehzadeh [12] improved this method by considering the correlations between edges by using a weighted function.

3. SYSTEM OVERVIEW
In the proposed system our main focus is on identifying input image and converting it into relevant text and speech. So firstly we create a database of images. The input to the system can be images form database or captured image. These images when taken as input from the system, system checks for similar kind of images and in database in order to identify the objects from the image. Once objects are detected then it identifies the image and system gives text output. This generated text is then undergoes

into text to speech synthesis and we get the speech output. This kind of approach may help the blind people to know the scenario around them. Following figure: 1 gives idea about system overview.

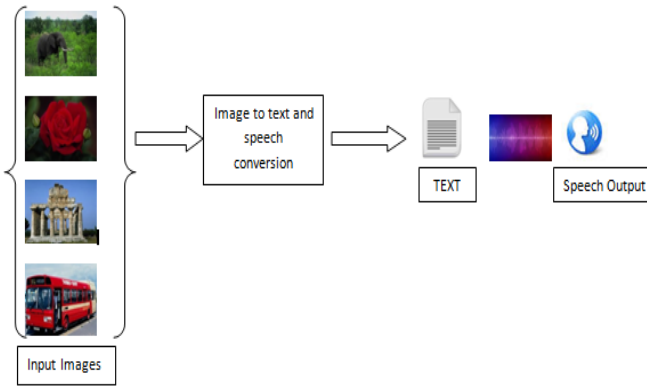

Fig 1: System Overview

4. PROPOSED SYSTEM ARCHITECTURE AND WORKING

In the proposed system our main goal is to identify objects in the image. When objects in the image are recognized it is easy to identify that image and convert it into respective test and speech. Techniques mainly used in the conversion process are:

- Pre-processing
- Gray scaling
- Edge detection
- Segmentation
- Feature extraction
- Speech synthesis

Based on these techniques of image processing we convert image into text as well as speech.

1) Pre-processing: Pre-processing of images involves mainly of removal of low-frequency background noise, removing reflections and masking portions of images. Pre-processing technique enhances data images in order to do further computational processing. Some common methods of pre-processing are:

- Smoothing: Spatial smoothing of images.
- Background Subtraction: Removal of unwanted background.
- Close (Dilate+Erode): Perform dilation followed by erosion on a binary image.
- Dilate: Perform dilation on a binary image.
- Erode: Perform erosion on a binary image.
- Max: Maximum value over neighboring pixels.
- Median: Median value over neighboring pixels.
- Mean: Mean value over neighboring pixels.
- Min: Min value over neighboring pixels.
- Open (Erode+Dilate): Perform erosion followed by dilation on a binary image.

2) Gray scaling: In gray scaling each pixel value of an image is represented using shades of gray. These kind of images are also known as black and white images. Each pixel intensity is expressed within the range of minimum and maximum where range is 0(black) and 1(white), any fractional value is in between.

Fig 2: Example of grayscaling. (A) Original image. (B) Gray scale image.

3) Edge Detection: Edge detection is the technique used to identify the fine edges in digital images. It identifies the points in the image at which the brightness of image changes very sharply. Point at which the image brightness changes are organized into set of curved line segments known as edges. Edge detection is mainly an important tool in the field of image processing for detection of features and feature extraction.

Approaches of edge detection-

Two main methods of edge detection are search based and zero crossing based. Search based method detects edges by computing the edge strength. The zero crossing based method searches for the zero crossing second-order derivative expression computed form images in order to find edges.

4) Segmentation: The aim of segmentation is converting an image into more meaningful and easy to analyze portions. Segmentation does the job of partitioning an image into multiple segments which help to locate the objects and boundaries (curves, arcs, lines, etc.) in an image. With the help of image segmentation we can assign a label to each pixel which then same labels share the certain characteristics. We can characterize the pixels in a region with respect to the characteristics such as color, intensity or texture.

5) Feature extraction: In the fields of machine learning, pattern recognition and image processing, feature extraction plays an important role of building derived values which are known to be the features. These features are intended to be informative, non-redundant, facilitating the subsequent and generalized steps. Extracted features should contain relevant data from input data. This technique plays an important task in our proposed system.

Feature Extraction is the key concept in CBIR. A certain number of features for each image are extracted, describing its high level content information. Then, according to the similarity of these vectors, we can compare two specific images to each other. This class uses different techniques to extract features related to a single or the group of images. There are two methods in this class which extracts the features:
- extractSingleImage : which extracts the features for a single image. It is used to extract the features of the input image.

- extractAll : which is used to extract the features for all of database images

6) Speech Synthesis: It is the process of artificially producing the human speech. Systems used for such purpose are called speech synthesizer which can be a software or hardware product. Concatenations of several pieces of recorded speech are stored in database and then synthesized speech is created.

A text -to-speech conversion system is used in our work to convert the generated text of images into speech output.

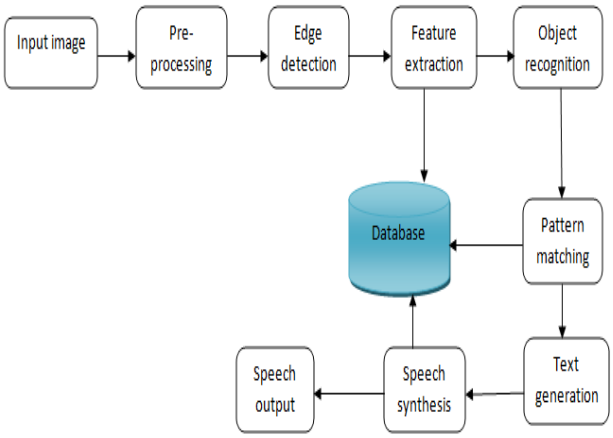

Fig 3: Architecture of proposed system

5. FLOW OF PROPOSED SYSTEM

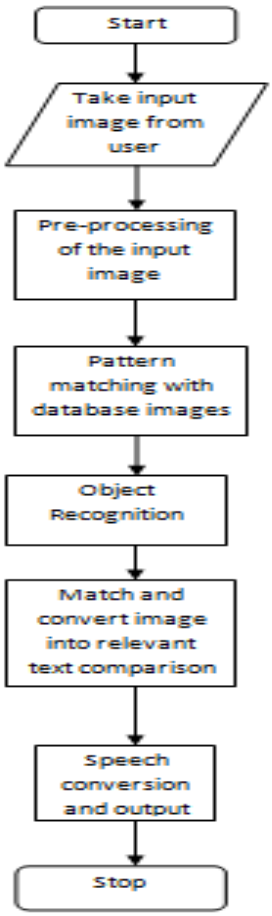

Fig 4: Flow of proposed system

In the above figure 4, the flow of our proposed system of image to text and speech conversion is given. Here, the input to the system is an image file. Then pre-processing and pattern matching phases are carried out. In these phases the gray scaling of images is done. After segmentation and edge detection the next phase is object detection. In this phase, the important objects of the image are analyzed. Then the matching and comparing of images with database are carried out. The relative text for the identified objects is generated and then that text is converted into speech which is easy to be heard by the blind people.

6. ALGORITHMIC STEPS

Canny edge detection algorithm is used in our proposed work.

It uses multi-stage algorithm to detect several ranges of edges in an image. It helps to extract structural information from objects and thus reduces the amount of data to be processed. Among the edge detection methods developed till now, canny edge detection algorithm is one of the most strictly defined and provides good and reliable detection.

Steps include:

1. Firstly apply the Gaussian filter to smooth the image for removal of noise.
2. Secondly, we find the intensity gradients of that image.
3. Then we apply the non-maximum suppression for removal of spurious response to edge detection.
4. Next we apply double threshold to determine potential edges.
5. Finally, only strong edges are considered and suppress all other edges that are weak and not connected to strong edges.

   Mathematical Model:

   Let S={Ii, Si, Ti, DBi, Qi, TRi} where

   Ii=Input images {I1,I2……In},

   Si=Speech of particular image {S1,S2……Sn},

   Ti=Text of particular image {T1,T2…..Tn},

   DBi=Database of Images, text and speech,

   Qi=Query image testing,

   TRi= Training on images.

Working of image to text and speech conversion system

The input to the system is an image which is given as input to the system by browsing it. Image with proper file extension and valid images are expected. System takes the input image and checks for relevant images in database. Once identified, the objects in the image are analyzed and then the text output is given in grid view. The below screenshot gives us the overview.

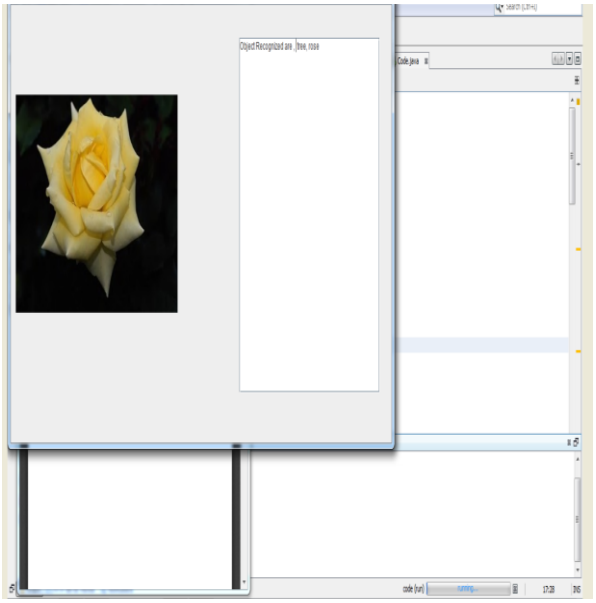

Fig 5: Image to text and speech conversion system

7. RELEVANT MATHEMATICS

1) Smoothing of images-

Smoothing of gray scale images is done by applying convolution with the Gaussian function $f(x) = 1/(sqrt(2*pi)*window)*exp(-x^2/(2*windo$

Parameters are:

Window : an odd positive integer controlling the width of the Gaussian.

2) Gaussian Filter-

Smoothing of images is done by Gaussian Filter.

The equation for Gaussian Filter kernel with size of 2k+1*2k+1 is given below:

$$H_{ij} = \frac{1}{2\pi\sigma^2} * \exp(-\frac{(i-k-1)^2+(j-k-}{2\sigma^2}$$

3) A simple edge model-

The edges obtained from natural images are not at all ideal edges. Gaussian smoothed step edge are the simplest extension of ideal step edge model. A one-dimensional image f has exactly one edge placed at x=0. It is modelled as:

$$f(x) = \frac{I_r - I_l}{2}\left(\operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right)+1\right) + I_l$$

At the left side of the edge, the intensity is $I_l = \lim_{x\to-\infty} f(x)$, and right of the edge it is $I_r = \lim_{x\to\infty} f(x)$. The scale parameter σ is called blur scale of the edge.

4) Segmentation-

A Gaussian model used for marginal distribution is given as:

$$\frac{1}{\sigma(l_i)\sqrt{2\pi}} e^{-\frac{(f_i-\mu(l_i))^2}{2\sigma(l_i)^2}}$$

8. ANALYSIS

During the analysis & testing phase we tried comparing the images with database. The percentage of accuracy and object recognition with database is given in below table.

Sr. No.	Object	Percentage Accuracy
1.	Horse	100%
2.	Dinosaurs	100%
3.	Flower	90%
4.	Mountain	60%
5.	Food	75%
6.	Human	80%
7.	Nature Scene	70%

Table 1: Result analysis

Graphical Representation:-

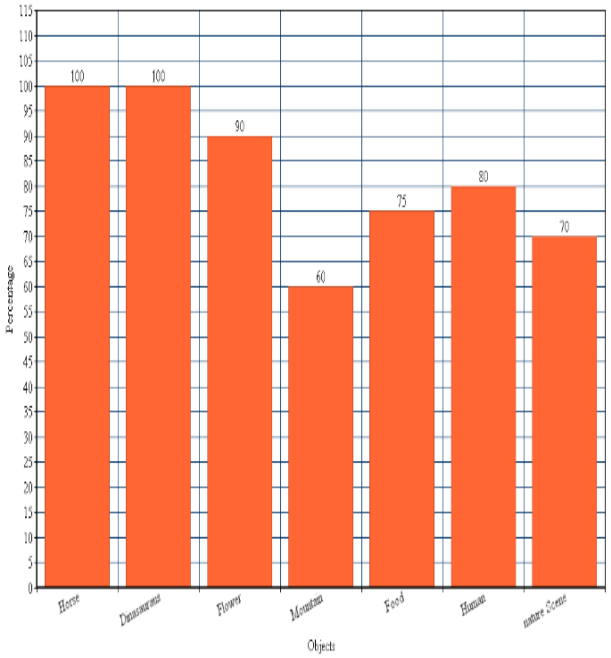

Fig 6:- Graph for object matching with database

Fig 7: Result given by system for the keyword "historical monument".

For 10 images, 9 are correct.

Thus from the analysis of normal person using the system it shows that the objects matching with the database is accurate and clear. From the analysis it is clear that system is accurate and except some flaws which are acceptable. Thus proposed system can be efficient but it still depends on the ability of user which is using it.

9. CONCLUSION

The proposed image to text as well as speech conversion system provides the solution to the problems faced by blind people. In proposed system we have applied a simple and fast method which works suitably to recognize image and convert it into text as well as speech. It is low time-consumption approach, so that the real time recognition ratio is achieved easily. In the proposed system Canny edge detection algorithm is used which will recognize the input image by detecting the edges of objects in the image. It is capable of handling the different input images and translates them into text and speech. The proposed system is designed to translate The dataset contains the number of hand images that are taken from multiple user of different size which helps to recognize the correct output to any user using the system. The proposed system is trained on predefined dataset.

In future work we are looking for the dynamic system which will identify the dynamic actions or images taken dynamically. The dynamic image recognition system contains not only shapes but also the many other objects in image. So the system will continuously recognize the dynamic movements in videos/ images. Also in future work we are trying to have the advanced technology such as video conferencing, and try to make android application.

10. ACKNOWLEDGMENT

I would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. Sincere thanks to my guide Prof. Ramesh Kagalkar sir his support, co-operation, and motivation provided to me during the work for constant inspiration, presence and blessings. I take this opportunity to express my profound gratitude and deep regard Prof. Arti Mohanpurkar, HOD and Asst. Professor Department of Computer Engineering for her exemplary guidance, monitoring and constant encouragement throughout the course of this work. I'm also thankful to reviewer for their valuable suggestions and also thank the college authorities for providing the required infrastructure and support.

11. REFERENCES

[1] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg and Tamara L. Berg, "Baby Talk: Understanding and Generating Simple Descriptions," IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 35, NO. 12, DECEMBER 2013.

[2] Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu, "I2T: Image Parsing to Text Description" ,IEEE transactions on image processing, 2008.

[3] Iasonas Kokkinos, Member, IEEE, and Petros Maragos, Fellow, IEEE "Synergy between Object Recognition and Image Segmentation Using the Expectation-Maximization Algorithm", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 8, AUGUST 2009.

[4] Fan-Chieh Cheng, Shih-Chia Huang, and Shanq-Jang Ruan, Member, IEEE "Illumination-Sensitive Background Modeling Approach for Accurate Moving Object Detection", IEEE TRANSACTIONS ON BROADCASTING, VOL. 57, NO. 4, DECEMBER 2011.

[5] DHIRAJ JOSHI, JAMES Z. WANG and JIA LI, The Pennsylvania State University, "The Story Picturing Engine—A System for Automatic Text Illustration", ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 2, No. 1, February 2006.

[6] *Munawar Hayat, Mohammed Bennamoun and Senjian An* "Deep Reconstruction Models for Image Set Classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[7] *Mina Makar, Member, IEEE, Vijay Chandrasekhar, Member, IEEE, Sam S. Tsai, Member, IEEE, David Chen, Member, IEEE, and Bernd Girod, Fellow, IEEE,* "Interframe Coding of Feature Descriptors for Mobile Augmented Reality", *IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 8, AUGUST 2014.*

[8] *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 Trans. PAMI, vol. 22, no. 12, 2000.

[9] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based 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.

[10] A. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2d-3d hybrid approach to automatic face recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 11, pp. 1927–1943, 2007.

[11] S. Feng, D. Xu, X. Yang, *Attention-driven salient edge(s) and region(s) extraction with application to CBIR,* Signal Processing 90, pp. 1–15, 2010.

[12] A. Vailaya, A. Jain, H.J Zhang, *On Image Classification: City Images vs. Landscape,* Proceeding of the IEEE workshop on Content-Based Access of Image and Video Libraries, pp. 3-8, 1998.

[13] J. Shanbehzadeh, F. Mahmoudi, A. Sarafzadeh, A.M. Eftekhari-Moghaddam, *Image Retrieval Based on the Directional Edge Similarity*, Proceeding of the SPIE: Multimedia Storage and Archiving Systems, Vol. IV, Boston, Massachusetts, USA, pp. 267-271, 1999.

12. AUTHOR's PROFILE

Mrunmayee G. Patil Research Scholar Dr. D. Y. Patil School of Engineering & Technology, Pune, University of Pune, Maharashtra, India. She has received her Bachelor's Degree in Information Technology from Padmashree. Dr. D. Y. Patil Institute of Engineering and technology, Pimpri, Pune, Maharashtra with first class distinction in 2013 from University of Pune. Currently she is pursuing her Master's Degree in Computer Engineering from Dr. D. Y. Patil School of Engineering & Technology, Pune, University of Pune.

Prof. Ramesh. M. Kagalkar was born on Jun 1st, 1979 in Karnataka, India and presently working as an Assistant Professor, Department of Computer Engineering, Dr. D. Y. Patil School Of Engineering and Technology, Charoli, B.K.Via – Lohegaon, Pune, Maharashtra, India. He is a Research Scholar in Visveswaraiah Technological University, Belgaum. He had obtained M.Tech (CSE) Degree in 2005 from VTU Belgaum and he received BE (CSE) Degree in 2001 from Gulbarga University, Gulbarga. He is the author of text book Advanced Computer Architecture which cover the syllabus of Visveswaraiah Technological University, Belgaum. He has published many research paper in international conference.