iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VI **Month of publication:** June 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Translating Images into Text Descriptions and Speech Synthesis for Learning Purpose

Yogesh N. Shinde[1], Mrunmayee Patil[2]
Department of Computer Engineering, Dr.D.Y.Patil SOET, Lohegaon, Pune.

Abstract— Image to text and speech conversion system can be useful for improving accessibility of images for visually impaired as well as physically challenging people understand the scenario from the images and also train the system as that of human brain. The techniques of image segmentation and edge detection play an important role in implementing proposed system. The system generates text descriptions for an input image given by the user. Object wise generation of sentences, preposition and conjunction mapping is a challenging task. The framework formulates the interaction between image segmentation and object recognition in the framework of Canny algorithm. The system goes through various phases such as pre-processing, feature extraction, object recognition, edge detection, image segmentation and Text To Speech (TTS) conversion. The proposed system database consists of huge set of sample images, which help to perform training of database. The accuracy of proposed system is achieved due to the proper recognition of objects and sentences are formed making use of the recognized objects. These sample images consists of several categories of images. The system mainly consists of two main modules such as image to text and text to speech. An image to text module generates text descriptions in natural language based on understanding of image. A text to speech module generates speech synthesis in English from description of natural language.

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

I. INTRODUCTION

Image processing is one of the most growing field in research and technology in to- days world. Image processing uses hardware and software as computing resources to pro- vide an efficient interface to process an image. Image processing uses various techniques such as image filtering, image pre-processing, image segmentation, image compression, image editing and manipulation, feature extraction, object recognition. An image can be defined in a function of two real variables f(r, w) where f as the amplitude (e.g. brightness) and of the image at the real co-ordinate position (r, w). The image can be of any file for- mats. These file format helps us to discriminate different types of images. In today's world there are around 285 million people who are visually impaired; out of which 39 million are visually impaired and 246 have low vision. Such people have very low scope to understand what exactly is going on in their current environment. There is no such interface which is easily available for such disabled people to interact with the world. Providing efficient interface for such people is of great need.

II. RELATED WORK

The author 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 [1]. 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.

The author 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 [8]:

A. Read the image and convert it into a three dimensional array of pixels.
B. Convert the image from color to black and white.
C. Removal of shades caused due to non-uniform light,
D. Turning black pixels into white ones and vice versa,
E. Threshold the image to eliminate pixels which are neither black nor white,

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

F. Removal of small components,

G. 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.

III. SYSTEM OVERVIEW

In proposed work of image to text and speech conversion system we develop a cost efficient and user friendly interface for visually impaired people. The primary motivation is to provide a blind person with a friendly speech interface with computer and to allow such people which are physically and visually challenged to use the system for understanding any type of scenario. One important approach to develop this system is to make any visually impaired person to analyze what is going on around him/her. visually impaired people usually rely on their partners or sense the scenario by their senses. In order to make a visually impaired people more and more independent we developed this system. Many challenges are faced by a visually impaired person in his/her day-to-days life while interacting with the world every day. Proposed framework goes through various phases.

Pre-Processing
Feature Extraction.
Image Segmentation
Text Conversion
Text to Speech synthesis

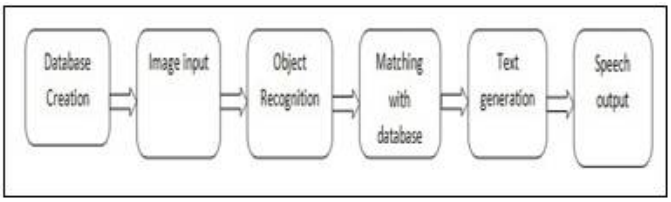


Fig 1: System Overview

A. Edge Detection

A set of connected pixels that forms a boundary between two disjoint regions is known as an edge. The task of segmenting an image into regions of discontinuity is done using edge detection. Edges usually occur on the boundary of two different boundaries in an image. Edge detection helps to clearly identify the changes in region of an image where gray scale and texture change in the regions of an image.

B. Canny Algorithm

This algorithm focuses mainly on three main aims of low error rate, minimize distance between real edge and detected edge and minimum response i.e. one detector response per edge to detect the edges in an image. The canny edge detector is an edge detection operator that uses a multi stage algorithm to detect a wide range of edges in images.

C. Image Segmentation

Image segmentation is another important aspect necessarily required to divide an image into regions or categories which then helps to identify correctly the object in an image. Segmentation functions on the properties shown by the pixels in an image, every pixel which belongs to same category has similar gray scale value whereas pixels of different categories have dissimilar values. Segmentation is often one of the critical steps in ana- lyzing the images because additional overhead of moving to each new pixel of an image while working with object in an image. Once image segmentation is done successfully, the other stages in image analysis

are much easier. While considering a fully automatic conversion algorithm, the success of image segmentation is partial and sometimes requires manual intervention. Segmentation mainly has two main objectives:

1) Divide or decompose the image into parts for further processing,
2) Perform change in organizing the pixels of image into higher-level units so that the objects become more meaningful.

IV. FLOW OF PROPOSED SYSTEM

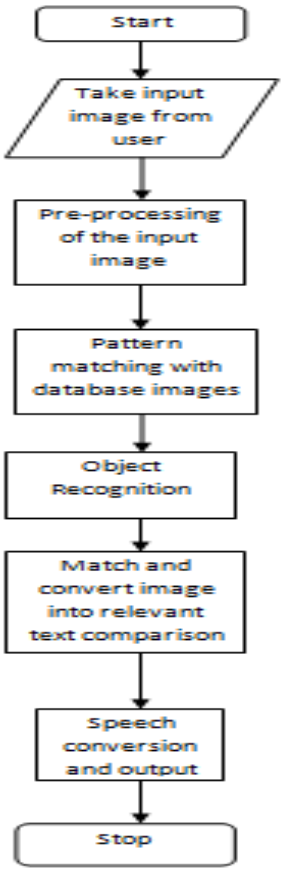


Fig 2: Flow of proposed system

The field of image recognition and computer vision is major fields of research in todays world. Language, whether written, spoken or typed, makes much of human communication. This language describes the visual world either around us or in the form of images and video. Combining visual imagery and visually descriptive language is a challenge for computer vision. The proposed work mainly focuses on generating text descriptions for particular input from user and speech output for generated text. The output of pro- posed system which is an automatically generated description of images has many related applications such as improving accessibility of images for visually impaired. The implementation of the system for automatic translation of image to text as well as speech, works on conversion system of images for getting text and speech output by generating proper image descriptions for appropriate images. This approach will help the betterment of understanding of people who are illiterate or learning new things. Also visually im- paired people can judge the scenario around them after the image to text descriptions are done into speech. We make prior use of computer vision and image processing techniques for this conversion process. The input image given to the system generates proper text description and speech by analyzing the objects in the image. For this purpose, the input image goes through pre-processing, feature extraction, edge detection and object recognition phases. Canny edge detection algorithm is used to detect the edges. After the edge detection process, the arcs, curves and lines help to identify the objects. Once objects are known their features are extracted like color and texture. Then the object keywords are fetched from the database and using a proper preposition and sentence forming words the appropriate sentence is build and text descriptions are generated by the system. After

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

the text generation process speech output is generated for the text description of the image.

V. ALGORITHMIC STEPS

Algorithm for Training the Images in Database

Algorithm 1 Algorithm for training the images in database

Input:

Set of database images.

Steps:

1. Color extraction using RGB values
   → RGB-HSV extraction
2. Edge detection
   → Detect edges
   → Obtained RGB image into binary values
3. Texture extraction
4. Save these features into a file
5. Save objects with corresponding features
6. Set mapping

Output:

1. Features of all images.
2. Objects and prepositions or conjunctions mapping of all images.

All the steps are carried out one after the other for proper working and correct output generation. Initial phase is give each image input to the system for training the database. Then color extraction, edge detection and feature extraction techniques are carried out. After this process the eighteen features of each image are generated and saved into a file. After this process, mapping each object with its relevant features is done in the set mapping phase. In this way the training on all the images in the database is performed.

Algorithm for Testing the Image

Algorithm 2 Algorithm for testing the image

Input:

Single image.

Steps:

1. Color extraction using RGB values
   → RGB-HSV extraction
2. Edge detection
   → Detect edges
   → Convert obtained RGB image into binary values
3. Texture extraction
   → Color
   → Shape
   → Texture
4. Calculate euclidean distance to compare features with database images
5. Get all the matching features
6. Get objects relevant to features
7. Map objects with relevant prepositions and conjunctions
8. Sentence generation
9. Convert generated text to speech by using the voice engine of system.

Output:

Text and speech output for an input image.

VI. EXPERIMENTAL RESULTS

For evaluation of this system, we use dataset taken from UCI repository which contains of total 14 categories of images. The image dataset contains of about 1000 images. From this set we make use of around 400 images for the purpose of training the proposed system. In the existing system, two forms of quantitative evaluation were performed, automatic evaluation using standard methods

Volume 4 Issue VI, June 2016
ISSN: 2321-9653

for evaluating generated sentences and human forced evaluations to directly compare the results between their method and several previous methods. In each case they also quantitatively evaluate and compare to two previous approaches for image description generation used on the same dataset. The first comparison method is the bottom up HMM approach from Yang et al. , which detects objects and scenes, and then hallucinates plausible verbs for generation (using text statistics). The second comparison method is a retrieval-based approach from Farhadi et al. This method detects objects, scenes and actions and then retrieves descriptive sentences from similar images through the use of a meaning space.

TABLE I

COMPARISON OF TIME TAKEN BY PROPOSED SYSTEM FOR TEXT AND SPEECH GENERATION.

Id	Category	Text generation (In milliseconds)	Speech generation (In milliseconds)
1.	Sea Shore	11246	4382
2.	Historical monument	6137	3245
3.	Bus	5167	3211
4.	Elephant	5917	4174

Table 1 shows the time taken by the proposed system for generating text descriptions and speech translations. The categories of images from dataset considered for text and speech translations are sea shore, historical monument, bus and elephant images. The text generation time is comparatively more than the time taken for speech generation. The text generated by the proposed system consults the database tables of object and preposition for proper matching object names and prepositions. When the text is displayed on the text field of the screen, the speech is generated for the same text using the voice engine of the laptop.

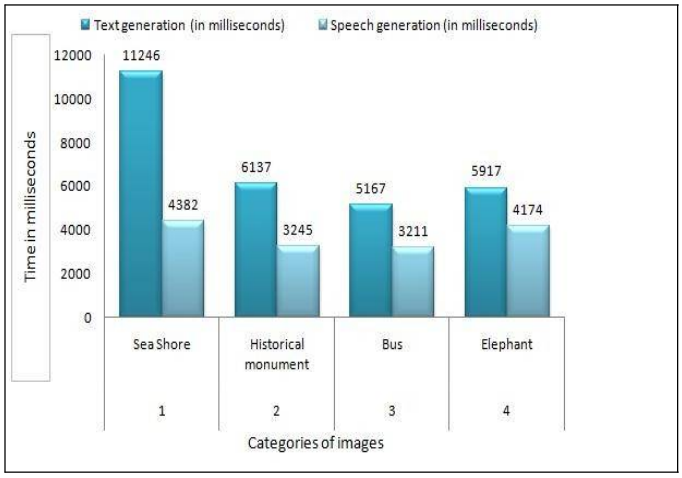


Fig 3: Comparison graph for text and speech generation.

VII. CONCLUSIONS

In proposed system we have applied a simple and fast method which works suitably for recognize image and convert it into text as well as speech. It is less time consumption approach, so that the real time recognition ratio is achieved easily. In the proposed system Cannys 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 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 images taken in the form of video. 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 which is a challenging work. Also in future work we are trying to have the advanced technology such as video conferencing and also make android application.

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.

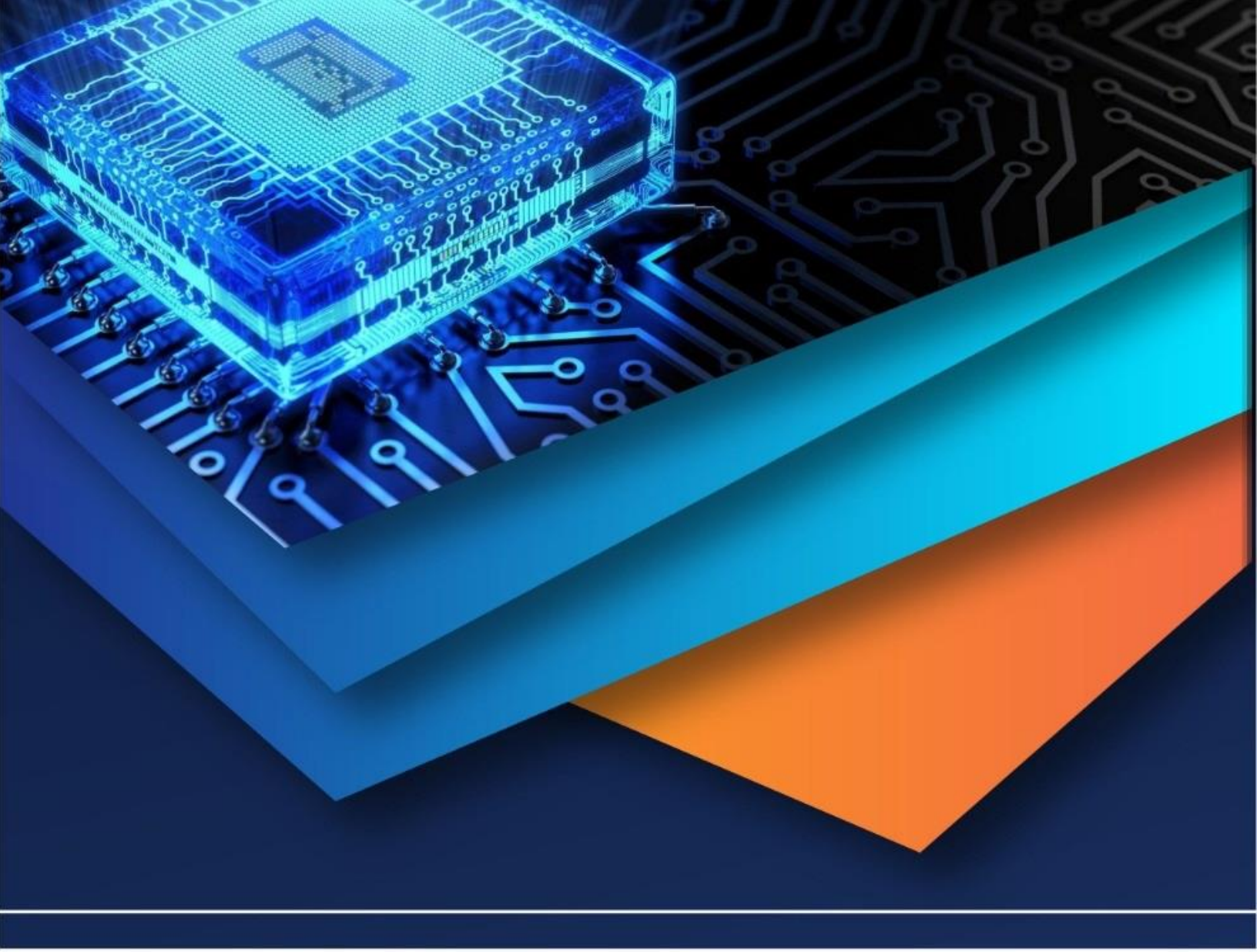
[9] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 2, no. 1, pp. 1–19, Feb. 2006.

[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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY