ISSN: 2321-7782 (Online)
Volume 1, Issue 6, November 2013
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

Image to Sound Conversion

Jaiprakash Verma[1]
Assistant Professor
Nirma University,
Institute of Technology
Ahmedabad – India

Khushali Desai[2]
Nirma University,
Institute of Technology
Ahmedabad – India

Barkha Gupta[3]
Nirma University,
Institute of Technology
Ahmedabad – India

Abstract: ***In this paper we are proposing a system to convert pictures into sound. This proposed system will identify the object from its picture and person will be able to listen to the name of the objects, in the picture. Here we will first get the image from a digital camera then by removing noise by Grey scale and after this Thresholding will be applied. In image processing, thresholding is used to split an image into smaller segments, or junks, using at least one color or grayscale value to define their boundary. After that Object recognizing can be done with Memory-Based Object Recognition Algorithm and object will be identified as Textual name and this name will be stored in the database. Then Optical Character Recognition will be applied to convert text into machine Text. This Text will be converted finally to sound.***

Keywords: Removing Noise, Introduction to Greyscale, Thresholding, Optical Character Recognition, Memory-Based Object Recognition Algorithm, Text to sound conversion.

I. INTRODUCTION

In this world or say the era of technology, where technology is transforming our life in a way you had never imagined before. Technology is not just changing our life but also helps make our life easier to live. For instance reminders, mobile phones, hearing head are one of those amenities. So we have proposed this Image to sound conversion System.

Here Image will be taken through any means of Picture taking say it camera or your mobile phone.

First task after taking picture we will remove noise. Noise reduction is the process of removing noise from a signal. All recording devices, both analogue and digital, have traits which make them susceptible to noise. Noise can be random or white noise with no coherence, or coherent noise introduced by the device's mechanism or processing algorithms. In the case of photographic film and magnetic tape, noise (both visible and audible) is introduced due to the grain structure of the medium. In photographic film, the size of the grains in the film determines the film's sensitivity, more sensitive film having larger sized grains. In magnetic tape, the larger the grains of the magnetic particles (usually a ferric oxide or magnetite), the more prone the medium is to noise. [1]

We all are familiar with the black and white TVs and Pictures of 19th century despite of gradual innovations to color photography; monochromatic (B & C) photography remains popular.

The digital revolution has actually increased the popularity of monochromatic photography because any digital camera is capable of taking black-and-white photographs. [2]

The use of many shades of gray to represent an image. Continuous-tone images, such as black-and-white photographs, use an almost unlimited number of shades of gray.

Conventional computer hardware and software, however, can only represent a limited number of shades of gray (typically 16 or 256). Gray-scaling is the process of converting a continuous-tone image to an image that a computer can manipulate [3].

After this Thresholding will be done, Thresholding is one of the most important approaches to image segmentation, in this method; pixels that are alike in grayscale (or some other feature) are grouped together. [4]

Now after all the alimentary work that we were here coming the main portion which is Recognition of the object.

Object recognition - in computer vision is the task of finding and identifying objects in an image or video sequence. Humans recognize a multitude of objects in images with little effort, despite the fact that the image of the objects may vary somewhat in different viewpoints, in many different sizes / scale or even when they are translated or rotated. Objects can even be recognized when they are partially obstructed from view. This task is still a challenge for computer vision systems. Many approaches to the task have been implemented over multiple decades. [5]

Here we have used Memory-Based Object Recognition Algorithm to do full fill our needs. This Object recognition is useful in applications such as video stabilization, automated vehicle parking systems, and cell counting in bio imaging.

This system can also be helpful for blind people as they can take the pic and can listen what is happening around them. Following are some statistics data of blind people.

- 285 million people are estimated to be visually impaired worldwide: 39 million are blind and 246 have low vision.
- About 90% of the worlds visually impaired live in developing countries.
- 82% of people living with blindness are aged 50 and above.
- Globally, uncorrected refractive errors are the main cause of visual impairment; cataracts remain the leading cause of blindness in middle- and low-income countries. [6]

II. How Greyscale actually works

All greyscale algorithms utilize the same basic three-step process:

- Get the red, green, and blue values of a pixel
- Use a fancy math to turn those numbers into a single gray value
- Replace the original red, green, and blue values with the new gray value. [2]

Greyscale Algorithm

When describing grayscale algorithms, let's focus on step 2 – using math to turn color values into a grayscale value. So, when you see a formula like this:

Gray = (Red + Green + Blue) / 3Recognize that the actual code to implement such an algorithm looks like:

For Each Pixel in Image

```
{
Red = Pixel.Red
Green = Pixel.Green
Blue = Pixel.Blue
Gray = (Red + Green + Blue) / 3
Pixel.Red = Gray
Pixel.Green = Gray
Pixel.Blue = Gray
}
```

This formula generates a reasonably nice grayscale equivalent, and its simplicity makes it easy to implement and optimize. However, this formula is not without shortcomings – while fast and simple, it does a poor job of representing shades of gray

relative to the way humans perceive luminosity (brightness). But that is not our purpose so we will take this algorithm and move ahead. [2]

III. THRESHOLDING

In image processing, thresholding is used to split an image into smaller segments, or junks, using at least one color or grayscale value to define their boundary. A possible threshold might be 40% gray in a grayscale image: all pixels being darker than 40% gray belong to one segment, and all others to the second segment. It's often the initial step in a sequence of image-processing operations.

- BinaryThreshold(T,M))≡ foreach Pixel in SourceImag

```
if (Pixel > T)
    DestImage[Pixel.Position]=M
else
   DestImage[Pixel.Position] = 0
```

- InverseBinaryThreshold(T,M) ≡ foreach Pixel in SourceImage

```
if (Pixel > T)
        DestImage[Pixel.Position] = 0
else
        DestImage[Pixel.Position] = M
```

- TruncateThreshold(T,M) ≡ foreach Pixel in SourceImage

```
if (Pixel > T)
        DestImage[Pixel.Position] = M
else
        DestImage[Pixel.Position] = Pixel
```

- ToZeroThreshold(T) ≡ foreach Pixel in SourceImage

```
if (Pixel > T)
        DestImage[Pixel.Position] = Pixel
else
        DestImage[Pixel.Position] = 0
```

- InverseToZeroThreshold(T,M) ≡ foreach Pixel in SourceImage

```
if (Pixel > T)
        DestImage[Pixel.Position] = 0
else
        DestImage[Pixel.Position] = Pixel
```

Above are the most popular functions which you can use for this purpose. [7]

IV. MEMORY BASED OBJECT RECOGNITION ALGORITHM

To use the Memory based Object Recognition we have referred idea of Randal C. Nelson's Research work at University of Rochester for Object Recognitions.

As this is a memory based Algorithm we will need something to store our memory and then when we want to do the reorganization we will be in the need of comparing.

So, we must first prepare a database against which the matching will take place. To do this, we first take ample amount of images of each object, covering the region on the viewing sphere over which the object may be encountered.

How: The exact number of images per object may vary depending on the features used and any symmetries present, but of the patch features we use, obtaining training images about every 20 degrees is sufficient, so to cover the entire sphere at this sampling requires about 100 images.

For every image obtained, the boundary extraction procedure is run, and the best 20 or so boundaries are selected as keys.

The basic recognition procedure consists of four steps, as follows:

First, potential key features are extracted from the image using low and intermediate level visual routines.

In the second step, these keys are used to access the database memory (via hashing on key feature characteristics and verification via local context), and retrieve information about what objects could have produced them, and in what relative configuration.

The third step uses this information, in conjunction with geometric parameters factored out about the key features regarding the position, orientation, and scale, to produce hypotheses about the identity and configuration of potential objects. These "pose" hypotheses serve as the loose global contexts into which information is integrated.

This integration is the fourth step, and it is performed by using the pose hypotheses themselves as keys into a second associative memory, where evidence for the various hypotheses is accumulated. Specifically, all global hypotheses in the secondary memory that are consistent (in our loose sense) with a new hypothesis have the associated evidence updated. After all features have been so processed, the global hypothesis with the highest evidence score is selected. Secondary hypotheses can also be reported. [8]

V. Storage Part of Objects and Results

First of all we will need the bulk of data already available with us to compare the newly came picture. The original database will be having the pictures unique id with its various key features as discussed above like position, orientation, scale, viewing dimensions from various angles etc.

Now, whenever the new picture's processed data will also be stored with its various key features to make hypothesis. After that 2 hypotheses comes from above algorithm they will be stored in the final result set which now will be extracted as text to convert it into sound.

We need 2 or more hypotheses in the final result set to take proper decision to determine the exact thing or to give various options to the person.

For example, if there is a picture like bottle then it may have 2 hypotheses. Either it may be a bottle or a jar. So we can take an exact decision about that picture by observing surrounding objects of it. If surrounding pictures are wooden cabinets, and shelves (see Figure 1) then it may be a jar, or else if surrounding pictures are door of fridge, other food related pictures than it may be a bottle.

Fig1. Examples of Bottles and jars in the picture. With color and gray scale effect.

VI. Optical Character Recognition

Optical character recognition, usually abbreviated to OCR, does the mechanical or electronic conversion of scanned images of hand or type written or printed text into machine-encoded text.

It is rampantly used as a form of data entry from some sort of original paper data source, whether documents, mail, or any other printed records.

As has been a common method of digitizing printed texts, they can be electronically searched, stored more trimly, displayed on-line, and used in machine processes such as machine translation, text-to-speech and text mining.

Early versions of OCR needed to be programmed with images of each character, and worked only for one font at a time. "Intelligent" systems with a high degree of recognition accuracy for most fonts are now common. Some systems are capable of reproducing formatted output that closely approximates the original scanned page including images, columns and other non-textual components.

We need OCR technique if there exist any written text or something. This will directly store to final result set of output as text without matching its key features with the database. [9]

VII. FINALLY TO SOUND

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech so it is said TTS; other systems render symbolic linguistic representations like phonetic transcriptions into speech. [10]

Synthesized speech can be created by joining pieces of recorded speech that are stored in a database. As a Systems differ in the size of the stored speech units; a system which needs to stores phones provides the largest output range, but may lack clarity. [10]

For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can include a model of the vocal area and also other human voice characteristics to create a completely "synthetic" voice output. [11]

VIII. DATABASE FOR SYNTHESIS

Now as we have synthesis the text by making unit selections we have to store complete recorded speech. To accomplish this large database is needed. The database would be created for each utterance of the recorded speech. The whole speech will be divided into segments. Some examples of such segments are individual phones, diaphones, half phones, words, phrases, sentences etc. These segments can be divided using a specially modified speech recognizer set which needs some manual corrections afterwards. Manual corrections will be done by visual representations such as waveforms and spectrograms. After storing each segment of recorded speech the major task would be index creation. The index would be created on the bases of segmentations as well some acoustic parameters like frequency, duration, neighboring phones etc.

To make these speeches run we need the desired target utterance. This can be achieved by determining the best chain of candidate units from the database. This process is achieved by generating a special weighted decision tree. Unit selection provides greatest naturalness, because it applies only a small amount of “Digital Signal Processing”. At the time of conclusion some system makes less use of signal processing to smooth the waveform. Maximum naturalness requires best unit selection and it can be achieved by having huge database to store the speech unit from all its aspects.

IX. CONCLUSION

So by doing the image acquire to removing noise, Object Identification, from there to OCR to TTS we can systematically develop something which will definitely help the people. Google is undertaking one project where you can click the image upload and Google image will tell you at which place this photo was taken. So this system can also come in handy for blind people to let them know their surroundings and even reading.

References

1. Rubin, P.; Baer, T.; Mermelstein, P. 1981 "An articulatory synthesizer for perceptual research". Journal of the Acoustical Society of America
2. http://en.wikipedia.org/wiki/Noise_reduction 2013
3. http://www.tannerhelland.co/3643/grayscale- image-algorithm-vb6/ 2011
4. http://www.webopedia.com/TERM/G/gray_scaling.html
5. http://itee.uq.edu.au/~elec600/elec4600_lectures/1perpage/lectanal4.pdf 2003
6. http://en.wikipedia.org/wiki/Outline_of_object_recognition 2013
7. http://www.who.int/mediacentre/factsheets/fs282/en/ 2013
8. http://weblog.benjaminsomme.com/blog/2012/05/19/introduction-to-image-thresholding/ 2012
9. http://www.cs.rochester.edu/~nelson/research/recognition/algorithm.html
10. http://en.wikipedia.org/wiki/Optical_character_recognition 2013
11. http://en.wikipedia.org/wiki/Speech_synthesis 2013

