Image to voice converter is software computer science

Image to Voice converter is software or a device to recognize an image and convert it into human voice. The purpose of the conversion is to provide communication aid for blind people to sense what the object in their hand or in front of them. This converter is also suitable for children at the age of three until six years old for early education part.

In this project converter, it consists of image processing and sound generation. For an image processing, it is a series of calculation techniques for analyzing, reconstructing, compressing, and enhancing images. When an object is inputting, an image will captured through scanning or webcam; analyze and manipulate of the image, accomplished using various specialized software applications such as MATLAB and output like a printer or a monitor.

Image processing has several techniques, including template matching, KNN (K-Nearest Neighbour), thresholding and etc. For the template matching, it is a technique for finding small parts of an image to match with the template image; it is also used to identify printed characters, numbers, and other small, simple objects. KNN (K-Nearest Neighbour) is an algorithm that can work very well in practice and easy to understand. It is also a lazy algorithm that does not use the training data points to do any generalization. Besides, thresholding technique is one of the most important approaches to image segmentation. It is a non-linear operation that can converts a gray-scale image into a binary image.

The purpose of image processing in this project is to analysis of a picture using techniques that can identify shades, colours and relationships that cannot be observed by the human eye. Besides that, an image processing is used to solve identification problems, i. e. in forensic medicine or in establishing weather maps from satellite photos. It assigns with images in bitmapped graphics form that have been scanned in or taken with digital cameras. For sound generation is to generate a sound through window sound library or play a wav file from computer.

Problem Statement

Nowadays, many visually impaired people still using blind man's stick to sense the road of the direction and object in front of them in this society. With just only a plain stick and a pair of covered eye, it is difficult for a human to get sense of their direction. Probably, they would not know what the objects around the people which had been blinded eye. As we can see the economy nowadays is getting worse, most of the people or family members were getting busy on their busy work life; they have no extra time to spend on the handicap people to give them a good care. In this case, for all the handicap people especially blind people, they have to get use to it on their living style. In order than that, this product is also available to help the small kid's to improve the ability on distinguishing or differentiate the daily use objects. This is the reason why the product mentioned above was developed.

Project Aim and Objective:

The aim of this project is to develop an Image to Voice converter which able to recognize an image from the webcam and then convert it into sound by window sound library or wav file with good performance. To achieve the main objective of this project, there are sub-objectives need to be carry through as follows:

To develop a unique image recognition algorithms for shapes and colours for real time application using MATLAB.

To analyze the performance of the image recognition algorithm in term of accuracy and time processing.

To develop an algorithm to convert recognized image to voice using MATLAB.

To analyze the performance of image to voice conversion algorithm.

Test the performance of the closed loop interface for the image and sound processing converter system.

To develop Graphical User Interface (GUI) of the image to voice converter for case of user finding.

Project Scope/Limitation

The scope of this project is to construct a unique image to voice converter within a period of time at cost not to exceed RM200. Referring to this project, it consists of hardware which is webcam and software which is MATLAB. The system of this project is to capture an image using webcam, then recognize an image and generate a sound using MATLAB with several techniques. This product specially created for visually impaired people or to improve small kid's learning capability. There was few limitation of this project which specified as follows:

Shape limitation

Colour limitation

Resolution limitation

Distance limitation

Literature Review

Image processing is a technique to convert an image into digital specification and go through some actions on it, so as to get an enhanced image or to collect some advanced information from it. It is a kind of signal exemption in which input is image, like video frame or photograph and output may be image or features related with that image. Frequently, image processing institution consist of treating images as two dimensional signals while applying already set signal processing techniques to them[1]. For the image recognition process can be divided into several algorithms which are image acquisition, image pre-processing, image segmentation, image representation and image classification. For the image acquisition, it is a digital image that captured by one or a few image sensors, such as various types of light-sensitive cameras, range sensors, tomography devices, radar, ultra-sonic cameras and etc. According to the type of sensor, the outcome of an image data is an generally two dimensional image, a three dimensional capacity, or an image order. The pixel values usually correspond to strength of light in one or a few spectral bands, but can also be involved many physical measures, such as depth, absorption or reflectance of sonic or electromagnetic waves, or nuclear magnetic resonance.

Image pre-processing is one of the algorithms that can increase the dependability of an optical inspection. This algorithm can be categorized into two categories which are image enhancement. Image enhancement requires intensifying the different features of images either for display or analysis targets. The enhancements techniques are edge enhancements, noise filtering, magnifying and sharpening an image. Several filter operations which increase or reduce certain image features allow an easier or faster evaluation. For examples, mean filter, median filter, wiener filter, and etc. With continuous use, an image will becomes degraded and has many errors. Image restoration is the process used to restore the degraded image. This process is also used to correct images read from different sensors that show up murky or out of focus[2].

Next, image segmentation is performed to assemble pixels into salient image areas, for example, areas corresponding to specific surfaces, objects, or inherent sections of objects. Segmentation could be used for object recognition, occlusion boundary estimation within motion or stereo systems, image density, image editing, or image database. The traditional image segmentation method can be divided into several techniques including gray threshold segmentation method, edge extraction method, regional growth method and split consolidation method and etc. Threshold technique was applied in this project. It is a technique that deals with gray-scale images. For the moment of the influence of noise or illumination, it can be assumed that the majority of pixels belonging to the objects will have a relatively low gray-level, whereas the background pixels will have a relatively high gray-level. For example, Black is represented by a gray-level of 0, and White by a gray-level of 255. Based on this observation, we can divide the pixels in the image into two dominant groups, according to their gray-level. These gray-levels may serve as “ detectors’ to distinguish between background and objects in the image. On the other hand, if the image is one of smooth-edged objects, then it will not be a pure black and white image; hence this would not be able to find two distinct gray-levels characterizing the background and the objects. This problem intensifies with the existence of noise[3]. In order to overcome the ill influence of noise and shading, there are two methods that can solve this problem which are Otsu known as “ Global Threshold” and Neighbourhood known as “ Adaptive Threshold”.

For the image representation, all information is commonly represented in binary. This is real of images as well as numbers and text. However, an important differentiation needs to be made between how image data is shown and how it is stored. Displaying includes bitmap representation while storing as a file includes many image formats, such as jpeg and png[4]. There are few techniques for image representation which are Roundness ratio known as Circularity, Fourier Descriptors and etc.

The intent of the image classification procedure is to sort all pixels in a digital image into one of several land cover categories, or “ themes”. This categorized data may then be used to deliver thematic maps of the land cover present in an image. Ordinarily, multispectral data are used to carry out the classification and truly the spectral pattern present within the data for each pixel is used as the numerical basis for categorization. The purpose of image classification is to determine and describe, as a distinct gray level or colour, the characteristics occurring in an image in terms of the object or kind of land cover these characteristics practically express on the ground[5]. The technique for this algorithm is using template matching and KNN (K-Nearest Neighbour).

Table : Comparison of image sensors for image acquisition[6, 7]

Types of Image Sensor

Strength

Weakness

1

Webcam

– allow face to face interaction

– low cost

– easy to use

– low resolution

– not portable

– no optical zoom lenses

– no auto-focus

2

Digital Camera

– high resolution

– portable with batteries

– has optical zoom lenses

– has auto-focus

– high operating speed

– less durability

– battery consumption faster

– high cost

– many complex function

From the Table 1, it can be seen that both image sensors have its own strengths and weaknesses. This research will more focus on webcam due to this image sensor is using for this project. Webcam can be used to connect with computer to capture an image for image recognition. On the other hand, it is easy to use and cheaper compare with digital camera which is more complex and high cost. However, the megapixel of digital camera is higher than webcam.

..

Table : Comparison of several types of filter for image pre-processing[2, 8]

Types of filter

Strength

Weakness

1

Median filter

– more robust

– more smoothing

– provide good results

– memory consuming

– complex computation

2

Mean filter

– intuitive

– simple to use

– smoothing

– not good in sharpen images

– susceptible to negative outliers

3

Wiener filter

– short computation time

– controls output error

– straightforward to design

– results often too blurred

– spatially invariant

From the Table 2, it can be seen that all filters have its own strengths and weaknesses. This research will focus on two types of filter which are median filter and mean filter. Median filter have been chosen for this project is because median filter is more robust on average than mean filter and so a not representative pixel in a neighbourhood will not influence the median value significantly. Since the median value needs to be the value of one of the pixels in the neighbourhood, the median filter does not establish new unrealistic pixel values when the filter straddles an edge. This is because of the median filter is better at preserving sharp edges than the mean filter. Also, median filter removes the noise level more than mean filter.

Table : Comparison of threshold techniques for image segmentation [9, 10]

Threshold Techniques

Strength

Weakness

1

Otsu

– fast

– ease of coding

– easy to use

– less sensitivity

– assumption of uniform illumination

– does not use any object structure or spatial coherence

– complex computation

2

Neighbourhood

– produce a good result

– less computation

– memory consumption

– time consumption

– sensitive

From the Table 3, it can be seen that both techniques have its own strengths and weaknesses. Otsu’s method, named after its inventor Nobuyuki Otsu, is a global threhold that consists of many binarization algorithms[11]. This method involves iterating through all probable threshold values and computing a measure of propagates for the pixel levels each side of the threshold, i. e. the pixels that can be falls in background or foreground. The purpose is to find the threshold value where the total of foreground and background propagate is at its minimum. Neighbourhood which known as adaptive threshold is used to separate desirable foreground image objects from the background based on the difference in pixel intensities of each region. The differences between both methods were Otsu uses a histogram to threshold the image and the Neighbourhood method uses a histogram to threshold the pixels in a small region/neighbourhood around the pixel. In addition, Otsu methods suffer less errors occur that are caused by the sensitivity of the local algorithms to image noise compare with the Neighbourhood methods.

Table : Comparison of the two techniques for image representation[12]

Techniques of Image Representation

Strength

Weakness

1

Roundness Ratio

very fast algorithm

scale, position and rotation invariant

high accuracy if image shape can be preserved properly after segmentation

susceptible to errors if object shape is changed due to improper segmentation

2

Fourier Descriptor

– medium speed

– produce a good result

– low computation cost

– overcome the weak discrimination ability

scale, position and rotation invariant

– difficult to obtain high order invariant moments

– cannot deal with disjoint shapes

From the Table 4, it can be seen that both techniques have its own strengths and weaknesses. Roundness is defined in term of a surface of revolution like cylinder, cone or sphere where all marks of the surface alternated by any plane vertical to a common axis in case of cylinder and cone are equal in distance from axis. As the axis and centre do not exist, measurements have to be made with consultation to surfaces of the figures of revolution only. The circularity of the outline is to measuring roundness[12]. Fourier Descriptors are used to describe the feature of contour of shape. It was founded in the early sixties last century by Cosgriff and Fritzsche. According to the Fourier analysis theory, Fourier coefficients can be often generated by Fourier transformation. Lower frequency coefficients have the general shape of the signature, and higher frequency coefficients have the more information about the shape. As the harmonic amplitude and the phase angle can represent the Fourier Descriptor, and Fourier coefficients are usually normalized by dividing the first Fourier coefficient separately. Because there are some fast algorithms in computing the coefficient of Fourier series, many recognition systems in machine vision using these coefficients as shape features.

Table : Comparison of several techniques for image classification [13-15]

Techniques for Image Classification

Strength

Weakness

1

Template Matching

– easy to implement

– high degree of flexibility

– high accuracy of detection

– shape limitation

– computation speed

– susceptible to scaling and rotation

2

K-Nearest Neighbour

– easy to implement

– very effective

– improve accuracy

– improve run-time performance

– poor run-time performance if the training set is large

– very sensitive

– outperformed by more exotic techniques

3

Neural Network

– minimize energy function

– high accuracy

– easy to use

– unstable

– curse of dimensionality

– space consumption

From the Table 1. 5, it can be seen that all techniques have its own strengths and weaknesses. This research will focus on two techniques which are Template Matching and K-Nearest Neighbour. The standard template matching technique is known as simple mechanism, high accuracy of detection, and is used as a general model assessment and error estimation. Hence, it plays a very important role in image processing, and is commonly used in object detection and recognition. But the contradiction between rapidity and accuracy is exceptional. The main factors affecting rapidity are searching calculation, and operations of template matching. Appropriately decreasing positions and similarity computing precision can increase the speed of template matching obviously. That is becoming a focus in this field. Many studies focus on improving the searching algorithm, decreasing the matching times by decreasing the matching points on the template of images, which need to be detected so that rapidity is realized. The typical algorithms are pyramid algorithm, genetic algorithm and so on. Each matching operation is based on the template matching, thus it is necessary to pay attention to improving the computation speed of template matching fundamentally[14]. The intuition underlying Nearest Neighbour Classification is quite straightforward, examples are classified based on the class of their nearest neighbours, it is often useful to take more than one neighbour into account so the technique is more commonly referred to as K-Nearest Neighbour (KNN) Classification where k-nearest neighbours are used in determining the class. Since the training examples are needed at run-time, i. e. they need to be in memory at run-time; it is sometimes also called Memory-Based Classification. Because induction is delayed to run time, it is considered a Lazy Learning technique[13].

. Analysis on Similar Products and Paper Literatures

Oral Image to Voice Converter by Takaaki HASEGAWA and Keiichi OHTANI[16]:

In this paper, the authors propose a new speech communication system to convert oral image into voice, " Image input Microphone". This system synthesizes the voice from only the oral image. This system provides high security and is not affected by acoustic noise, because actual utterance is not always necessary to input. Moreover, since the voice is synthesized without recognition, this system is independent of languages.

Simulations to convert oral image to voice about Japanese five vowels are carried out as basic investigation. A vocal tract area function is estimated from the oral image, and PARCOR synthesis filter is obtained from the vocal tract area function. The PARCOR synthesis filter is driven by a pulse train. The performance of this system is evaluated by hearing tests of the synthesized voice. As a result, audible voice has been synthesized and the mean recognition rate of Japanese five vowels has been 91%.

This paper describes a system to convert oral image into voice with considering human’s lip-reading ability. In the proposed system, the voice is directly synthesized only from the oral image without recognition, and actual utterance is not always necessary to input. They use both the feature of a tongue and the feature of lips obtained from the oral image. Therefore this system is not affected by the acoustic noise, and simultaneously, it provides high security because of no utterance input capability.

The system structure of this product is using a vocal tract area function which is equivalent to the transfer function of the vocal tract as a parameter. “ Indirect” means synthesis via the vocal tract area function. The vocal tract area function is obtained from the PARCOR analysis of speech signals, and speech signals are synthesized by inverse processing of PARCOR analysis. Therefore if the vocal tract area function is estimated from oral image signals, they can convert the oral image to the corresponding voice. Human utters various voice by changing the vocal tract, and each articulator moves not independently but cooperatively in utterance, It is generally known that the information of articulation is obtained from lip-reading.

Software Comparison

Table below shows that the two comparison of the software between MATLAB and C++.

Table : Comparison of software between MATLAB and C++[17]

Types of Software

Strength

Weakness

1

MATLAB

– easy to learn

– fast numerical algorithms

– inexpensive software

– fast development

– slow processing

– complex computation

2

C++

– mature standard

– large community

– fast

– complex computation

– difficult to debug

– low level programming

From the Table 6, it can be seen that both types of software have its own strengths and weaknesses. MATLAB is software that has been widely used in image processing and computer vision community. Multiple image analysis function has been build into this software; it is very useful image analysis tools for end user. C++ is a standard template library (STL), computer graphics, and image processing. Based on C++ template mechanism, the library accepts all C++ build-in types as the image data, although certain functions are only valid to subset of build-in types. MATLAB has been selected due to the project analysis characteristic. MATLAB version R2010b will be used to analyze the image quality and performance in this project.

Project Methodology

This project has been divided into hardware and software. For the hardware section is the webcam as the input and speaker as the output. For the software section is using MATLAB to recognize image to sound with several image processing techniques.

Block Diagram

Webcam

Image Segmentation

(Thresholding)

Image Acquisition

(Acquire image)

Image Preprocessing

(Median filtering)

MATLAB

Image Representation

(Roundness Ratio)

Sound Generation

(WAV file)

Image Classification (Template Matching using KNN)

Speaker

Figure 1: Block diagram of Image to Voice converter.

The block diagram shown in Figure 1 is the basic concept on the system interface that needed to be carried out. Base on the block diagram, first prepared a webcam. Then, capture the image in front of the webcam. After that, perform a median filtering in image pre-processing using MATLAB. It will filtered unwanted signal or noise inside the image. Next is image segmentation, referring to the literature review, the most suitable method is using Otsu's method in thresholding techniques to convert grayscale image into binary image to do segmentation. Secondly, find the largest object and do the image representation using roundness ratio to calculate the ratio of the largest object to determine which one is the nearest to the template ratio. Next stage is image classification, using template matching with KNN techniques to find the small part of the image to match with the template image.

After matching done, it will automatically generate a sound from the computer with WAV file.

Flow Chart

Start

Acquire image from webcam

Perform median filtering

Colour Space Conversion

Thresholding using Otsu

Image labelling

Find the largest object

Image Representation

-roundness ratio

Template matching using KNN

Is the image matched?

No

Yes

Generate Sound

Figure 2: Flow chart of Image to Voice converter.

Based on Figure 2, before the beginning of image recognition, first, acquire an image in front of the webcam, and then the acquired image will go through image enhancement process to perform median filtering to filter some unwanted noise and sharpening the image. After that, the image will perform a colour space conversion which is convert the image colour space to another colour space, i. e. RGB, HSV, YCbCr and etc. The purpose of converting the colour space is to ensure that the converted image to be as same as the possible to the original image. Next, perform a threshold technique using Otsu's method to calculating a measure of spread for the pixel levels each side of the threshold. The reason of doing this is to separate the objects from the background. Once the thresholding technique is done, perform a image labelling by taking the outside lines in the image and label them as occluding the background. After that, find the largest object and do the image representation using roundness ratio to calculate which object is similar to the template ratio. Then, perform a template matching techniques to find a match between the template and a portion of the image. The template that most closely matches the object is then found using the KNN method to do a matching system with the database image. If the data is matched, it will generate a sound automatically by using MATLAB to load the wav file from the computer or laptop. After that, it will repeat the procedure starting from the first step. If the data is unmatched, it won't generate a sound and it will go back to the first step and repeat the procedure again until the data is matched.

Project’s Method

Median Filter

Median filters are nonlinear rank-order filters based on replacing each element of the source vector with the median value, taken over the fixed neighbourhood of the processed element. These filters are widely used in image and signal processing applications. The purpose of median filtering is to removes impulsive noise, while keeping the signal blurring to the minimum[18].

Otsu’ Method

Otsu’s method is a widely used method of segmentation, also known as the maximum infra-class variance method or the minimum inter-class variance method. This method involves iterating through all the possible threshold values and calculating a measure of spread for the pixel levels each side of the threshold, i. e. the pixels that either falls in foreground or background. The aim is to find the threshold value where the sum of foreground and background spreads is at its minimum[11].

Roundness Ratio/Circularity

Roundness is defined as a condition of a surface of revolution like cylinder, cone or sphere where all points of the surface intersected by any plane perpendicular to a common axis in case of cylinder and cone. Since the axis and centre do not exist physically, measurements have to make with reference to surfaces of the figures of revolution only. For measuring roundness, it is only the circularity of the contour which is determined[12].

Template Matching

The classical template matching method is charactered as simple mechanism, high accuracy of detection, and is used as a general model evaluation and error estimation. Therefore, it plays a very important role in image processing, and is widely used in object detection and recognition. It is a technique for finding small parts of an image to match with a database image[14].

K-Nearest Neighbour (KNN)

K-Nearest Neighbour (KNN) is a branch of simple classification and regression algorithms. It can be defined as a lazy method. It does not use the training data points to do any generalization. Although classification remains the primary application of KNN, it can use to do density estimation also. Since KNN is non parametric, it can do calculation for arbitrary assignation[19].

Project Specification

This project is divided into 3 main sections which are hardware, software and project estimate cost.

Hardware

The hardware was using for this project is Logitech HD Webcam C310, below is the basic requirement of the webcam:

logitech-hd-webcam-c310. png

Figure 3: Logitech HD Webcam C310[20]

Windows Vista, Windows 7 (32-bit or 64-bit) or Windows 8

1 GHz

512 MB RAM or more

200MB hard drive space

Internet connection

USB 1. 1 port (2. 0 recommended)

Software

The software for this project is using MATLAB for image recognition and sound generation.

Project Estimate Cost

The estimate cost for this project is RM89 which was the Logitech HD Webcam C310, because this project was basically software based project and the software to be used is MATLAB from college engineering lab.

Gantt Chart