Form segmentation with tesseract

Tanvir Hasan

Reg No: 12101005

Anup Kumar Kar

Reg No: 12101033

Sayem Hossain

Reg No: 12101046

submitted in partial fulfilment of the degree of

Bachelor of Science, with Honours

at the University of Asia Pacific, Dhaka,

Bangladesh.

March 25, 2016

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Md. Shiplu Hawlader, Lecturer, Department of Computer Science and Engineering, University of Asia Pacific. We also declare that no part of this thesis and thereof has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

Md. Shiplu Hawlader
Supervisor

Signature

Tanvir Hasan
Candidate

Anup Kumar Kar
Candidate

Sayem Hossain
Candidate

Approval

The Thesis Report " Form segmentation with tesseract" submitted by Tanvir Hasan, Reg.NO.12101005, Anup Kumar Kar, Reg.NO.12101033, Sayem Hossain, Reg.NO.:12101046 students of Spring-2012, to the Department of Computer Science & Engineering, University of Asia Pacific, has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science & Engineering and approved as to its style and contents.

Approved as to the style & contents by	
Dr. Bilkis Jamal Ferdosi Assistant Professor Department of Computer Science & Engineering, University of Asia Pacific	
Dr. Khan Md. Anwarus Salam Assistant Professor Department of Computer Science & Engineering, University of Asia Pacific	
Md. Akhtaruzzaman Adnan Assistant Professor Department of Computer Science & Engineering, University of Asia Pacific	
Md. Habibur Rahman Lecturer Department of Computer Science & Engineering, University of Asia Pacific	

Acknowledgements

First of all, thanks to Almighty Allah for giving us the potency and energy to complete this thesis successfully.

We want to express out gratefulness towards our thesis supervisor Md. Shiplu Hawlader for his valuable advices and important suggestions. His regular and active supervision and erudite directions from the beginning to the end were the driving forces for the successful completion of the research work.

We would like to convey our thankfulness to all of our teachers at the Department of Computer Science and Engineering, University of Asia Pacific. Discussions with many of them have helped us tremendously in improving the quality of our work. We also thank the department for providing us with resources which were necessary for the preparation of the thesis.

And last but not the least, we would like to express thanks to our parents and family members for their tremendous support and inspiration.

Abstract

Image segmentation is an important component in many image analysis and computer vision tasks. Particularly, the problem of efficient interactive foreground object segmentation in still images is of great practical importance in image editing and has been the interest of research for a long time. Classical image segmentation tools use either texture, color or edge (contrast) information for the purpose of segmentation. Deformable models, Graph-cut, GrabCut etc. are some prominent methods used for the segmentation of a foreground object. Object segmentation methods have helped in many computer vision areas, such as scene representation & interpretation, content based image retrieval, object tracking in videos, medical applications etc.

Most object segmentation techniques in computer vision are based on the principle of boundary detection. These segmentation techniques assume a significant and constant gray level change between the object(s) of interest and the background. However, this is not true in the case of textured images. In textured images, there exist many local edges of the texture micro units (texels), due to the basic nature of a texture image. In case of textured images, the object bound- ary is defined as the place where texture property changes. So to perform the correct segmentation in case of textured images, there is a need to incorporate the textural information in the segmentation process.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Segmentation of Images is the widely investigated ?eld of image processing, image analysis and important module of early vision problem. It is the process to cluster a form image into some isolated image regions corresponding to individual surface, objects or some natural part of the object. It is the process of separating an image into some disjoint or distinct regions whose characteristic such as intensity; colon texture etc are similar. No two such regions are similar with respect to these characteristics. in digital image processing, digital image analysis usually involves a ?low-level? and a ?high-level? processing. In low-level analysis, the representation of an image is transformed from a numerical array of pixel intensities to a symbolic set of image primitives: edges and regions. In high-level analysis, object labels (or interpretations) are assigned to these primitives, thereby providing a semantic description of the image.

An image is segmented for different kind of implementations like object recognition to extract data from it, occlusion boundary estimation within motion or stereo systems, image compression, image editing or image database lookup. The output of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in a region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristic . Image segmentation is a fundamental part of the ?low level' aspects of computer visiOn and has many practical applications such as in medical imaging, industrial automation and satellite imagery.Traditional methods for image segmentation have approached the problem either from localization in class space using region information or from localization in position, using edge or boundary information. for monochrome images generally are based on one of two basic properties of gray- level values: discontinuity

and similarity. In the ?rst category, the approach is to partition an image based on abrupt changes in gray level. The principal areas of interest within this category are detection of isolated points and detection of lines and edges in an image. The principal approaches in the ?rst category are based on edge detection, and boundary detection. Basically, the idea underlying most edge-detection techniques is the computation of a local derivative operator. The ?rst derivative of the gray-level pro?le is positive at the leading edge of a transition, negative at the trailing edge, and zero in areas of constant gray level. Hence the magnitude of the ?rst derivative can be used to detect the presence of an edge in an image. In this paper we are proposing a form segmentation, an image segmentation methodology using threshold techniques that will be applied to a form image and data matching algorithms to detect different objects and clusters.

The goal of segmenting a form image is to extract text data isolating texts of different part of that image by segmenting form into several sections. It becomes easy to extract digital data by processing individual objects into texts if we can first isolate that object scenario and use it for processing.

1.1 Objective

Not yet done

1.2 Motivation

In recent years, with the advancement of digital era, we are facing a problem of converting handwritten data into digital data for the purpose of saving them in database or repositories in order to analyse or perform operation on the data or keeping log for the future. We still need to process handwritten documents and forms manually that takes lot of times and resources and increases the possibilities to make mistake when processing that data. That causes serious Inefficiency and pain worth not to do it manually.

Several offices including government and non-government organizations gets a lot of applications and different kind of forms every day that needs to be processed immediately with efficient measure and more accuracy. In this paper we have proposed an efficient way to process those analogue data into digital texts through form segmentation, a technique of processing form image into digital text. Experimental results shows it?s accuracy and efficiency to process a form that is lot less time consuming

and effective.

1.3 Overview of this book

In this chapter we have introduced the problem of predicting amino acid interaction network in protein. Rest of the chapters are organized as follows.

In Chapter 2, we will acquire some background knowledge by discussing about protein structure including amino acid, primary, secondary and tertiary structure of protein.

In Chapter **??**, we are going to discuss about some existing researches on protein structure prediction and amino acid interaction network prediction.

In Chapter **??**, we will discuss about amino acid interaction network and verify network properties of amino acid in protein with PDB data.

In Chapter **??**, we are going to formulate the amino acid interaction network prediction problem theoretically and mathematically.

In Chapter **??**, we will present a new algorithm based on multi-objective optimization and ant colony optimization to predict the formulated problem of amino acid interaction network prediction.

In Chapter **??**, we will analyze the algorithm with some PDB data and show the result.

Finally, Chapter **??** concludes the document.

Chapter 2

Background Study

Base of a good research is the understanding of the background terms and definition. To understand the Image processing and character reorganisation, one have to clearly understand about Image filtering, Image Transformations and tesseract OCR Engine operation. In this chapter as background knowledge discovery we will discuss about some Image processing term and tesseract.

2.1 Gray Scale

In photography and computing, a grayscale or greyscale digital image is an image in which the value of each pixel is a single sample, that is, it carries only intensity information. Images of this sort, also known as black-and-white, are composed exclusively of shades of gray, varying from black at the weakest intensity to white at the strongest

2.2 Blur

Blurring is a very powerful operation used in image processing and procedural texture generation. Blurs involve calculating weighted averages of areas of pixels in a source image for each pixel of the final blurred image. Blurring images so they become fuzzy may not seem like a useful operation, but actually is very useful for generating background effects and shadows.

2.2.1 Gaussian smoothing

The Gaussian smoothing operator is a 2-D convolution operator that is used to 'blur' images and remove detail and noise. The idea of Gaussian smoothing is to use this 2-D

distribution as a 'point-spread' function, and this is achieved by convolution. Since the image is stored as a collection of discrete pixels we need to produce a discrete approximation to the Gaussian function before we can perform the convolution. I Gaussian smoothing technique is widely used effect in graphics software, typically to reduce image noise and reduce detail.

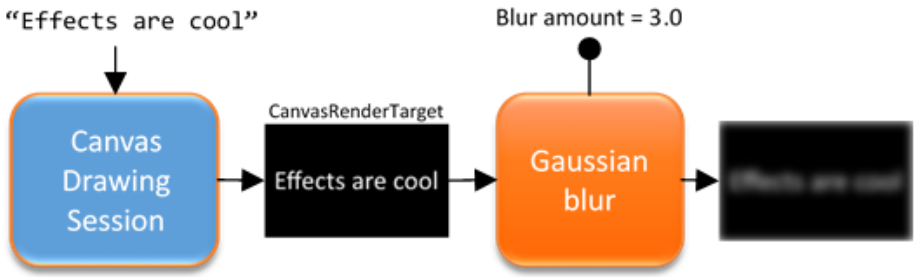


Figure 2.1: Gaussian smoothing

2.3 Thresholding

Thresholding is one of the most basic techniques for what is called Image Segmentation. When the threshold technique is applied on an image, we get segments inside the image and each of segments are represent something.

- Thresholding is a simple way of segmentation.
- We separate out various regions of an image regarding to objects which we want to analyze. The object pixels and the background pixels are the backbone of this separation.
- To differentiate the pixels we are interested in from the rest, we perform a comparison of each pixel intensity value with respect to a threshold.
- Once we have separated properly the important pixels, we can set them with a determined value to identify.

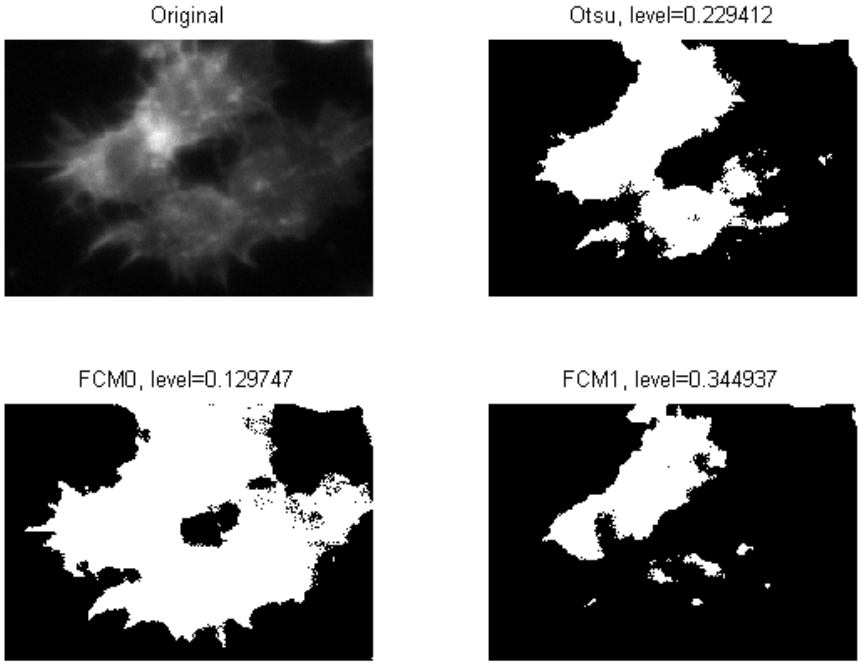
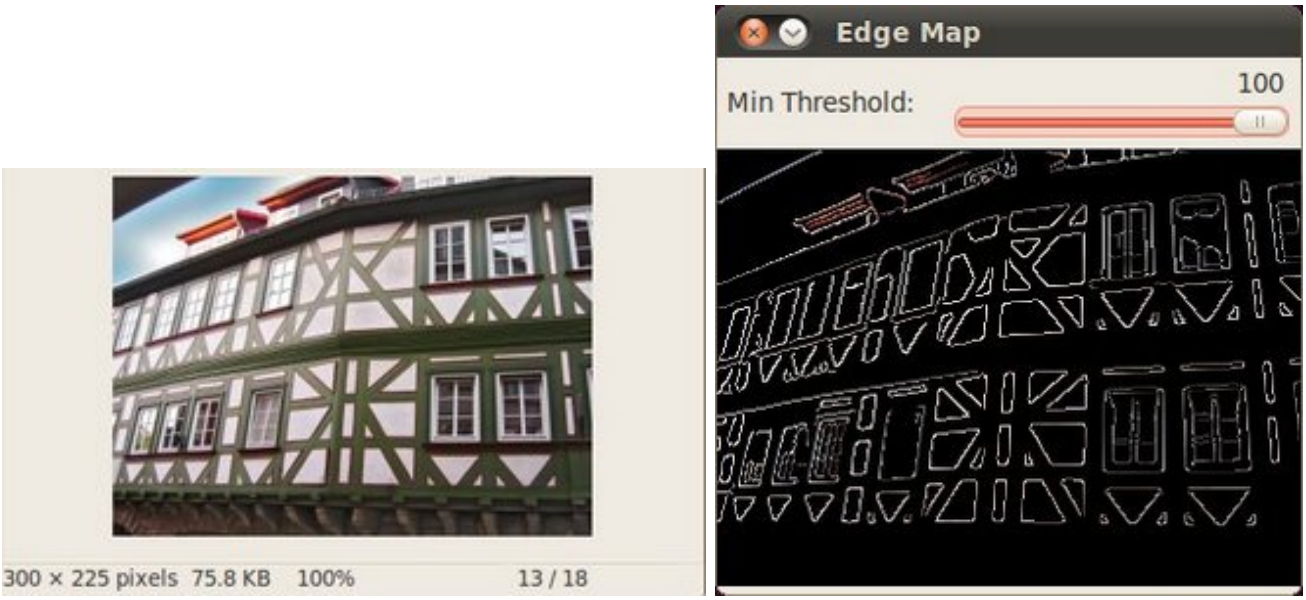


Figure 2.2: Thresholding

2.4 Edge Detection

In digital image processing, edge detection is an important subject matter. Edge detection is a crucial step in object recognition. It is a process of finding sharp discontinuities in an image. The discontinuities are abrupt changes in pixel intensity which characterize boundaries of objects in a scene. In short, the goal of edge detection is to produce a line drawing of the input image. The Canny operator is also known as the optimal detector, developed by John F. Canny in 1986. There are multiple steps to implement the Canny operator. First, a Gaussian filters is used to smooth the image to remove noise in an image. Second, compute the gradient magnitude. Third, apply the algorithm to remove the pixels that are not part of an edge. Last step is involved with the use of hysteresis thresholding along edges. Hysteresis is based on two thresholds which are upper and lower. If a pixel gradient is higher than the upper threshold, then the pixel will be marked as an edge and if a pixel gradient is below the lower threshold, then the pixel will be marked as a non-edge.

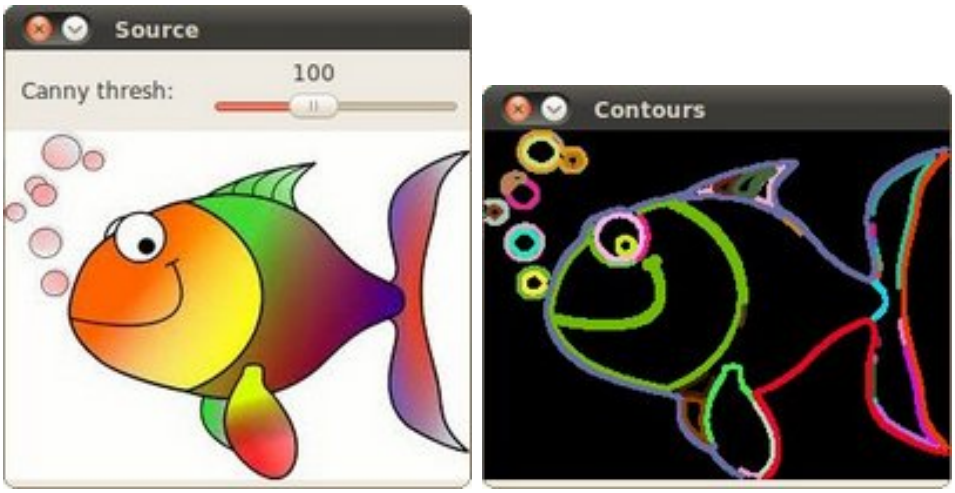


(a) Original Image (b) Edge Detected Image

Figure 2.3: Edge Detection

2.5 Contour

Although algorithms like the Canny edge detector can be used to find the edge pixels that separate different segments. The next step is to be able to assemble those edge pixels into contours. A contour is a list of points that represent, in one way or another, a curve in an image. This representation can be different depending on the circumstance at hand. There are many ways to represent a curve. Contours are represented in OpenCV by sequences in which every entry in the sequence encodes information about the location of the next point on the curve.



(a) Original Image (b) Contour Detect

Figure 2.4: Find Contour

2.6 Tesseract OCR engine

Tesseract is an open source optical character recognition (OCR) engine. HP originally was originally started it as a project. Later it was modified, improved and taken over by Google and later released as open source in year 2005.

Tesseract is considered as one of the most accurate free software OCR engines currently available. A large variety of other OCR software now uses it as a base. It is an excellent quality OCR program, with a large amount of flexibility, a solid codebase, and a large, engaged community of interested people around it.

This thesis investigates the principles of optical character recognition used in the Tesseract OCR engine and techniques to improve its efficiency and runtime.

Optical character recognition (OCR) method has been used in converting printed text into editable text in various applications over a variety of devices such as Scanners, computers, tablets etc. But now Mobile is taking over the computer in all the domains but OCR still remains one not so conquered field. This paper focuses on improving the Tesseract OCR efficiency for Bangla.

This thesis presents a preprocessing technique being applied to the Tesseract Engine to improve the recognition of the characters keeping the runtime low.

2.7 Why Tesseract

The first relevant criterion in Tesseract is the fact that is free and open source (FOSS), which is an advantage and a key point in the research development.

Usually, whenever Tesseract is compared to another free OCR tool, it is the best whether in terms of recognition rate or speed. Even, when it is compared with the Finereader commercial tool, Tesseract arrives to rub it and managed to overtake for handwritten writing.

2.8 Optical character recognition systems

An OCR system is a system that takes a text image as input and applies certain treatments through modules making up the system in order to output editable file with the same text.

The architecture of an OCR system varies from one system to another as needed.Optical Character Recognition systems are usually composed of the following phases:

- Preprocessing phase: prepares the sensor data to the next phase. It is a set of treatments allowing image quality increasing.
- Segmentation phase: delimits document elements (line, word, character). By applying good segmentation techniques, we can increase the performance of OCR systems.
- Feature extraction phase: defines features characterizing the delimited elements of a document. Feature extraction is one of the most important steps in developing a classification system. This step describes the various features characterizing the delimited elements of a document.
- Classification phase: recognizes and identifies each element. It is performed based on the extracted features.
- Post-processing phase: it is an optional phase. It may be automatic or manual.

Several approaches and techniques have been developed for each module.