UI2CODE: Computer Vision Based Reverse Engineering of User Interface Design

Mulong Xie

A thesis submitted for the Course
COMP4540 Software Engineering Research Project
The Australian National University

October 2019

Except where otherwise indicated, this thesis is my own original work.

Mulong Xie
11 October 2019

Acknowledgments

Who do you want to thank?

Abstract

UI designs consist of heterogeneous UI elements, such as natural images, drawings, graphic symbols, and computer-rendered objects (e.g., texts, buttons, boxes). Existing image segmentation techniques start with a fine partition into small regions, and gradually merge them into larger and larger ones. When applied to UI designs, this bottom-up strategy results in many primitive low-level shapes (i.e., oversegmentation). We propose a novel, top-down, divide-and-conquer approach that segments a UI design into a tree of UI design elements, corresponding intuitively to the perceptual boundary and organization of UI design elements. Experiments on large numbers of UI designs from Google play and Dribbble confirm the effectiveness of our approach for segmenting UI designs with complex visual composition of heterogeneous UI design elements. Our approach enables a new way of data-driven design applications on Google play and Dribbble design data, which is impossible before due to the lack of element-level metadata.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis Statement

I believe A is better than B.

1.2 Introduction

Put your introduction here. You could use \fix{ABCDEFG.} to leave your comments, see the box at the left side.

You have to rewrite your thesis!!!

1.3 Thesis Outline

How many chapters you have? You may have Chapter 2, Chapter **??**, Chapter **??**, Chapter 6, and Chapter 7.

Chapter 2

Background and Related Work

At the begging of each chapter, please introduce the motivation and high-level picture of the chapter. You also have to introduce sections in the chapter.

Section 2.1 xxxx.

Section 2.2 yyyy.

2.1 Motivation

2.2 Related work

You may reference other papers. For example: Generational garbage collection is perhaps the single most important advance in garbage collection since the first collectors were developed in the early 1960s. (doi: "doi" should just be the doi part, not the full URL, and it will be made to link to dx.doi.org and resolve. shortname: gives an optional short name for a conference like PLDI '08.)

2.3 Summary

Summary what you discussed in this chapter, and mention the story in next chapter. Readers should roughly understand what your thesis takes about by only reading words at the beginning and the end (Summary) of each chapter. [Tian et al., 2016]

Chapter 3

Why Not Deep Learning?

Speaking of object detection, the first idea coming to mind for most contemporary researchers in related fields is the set of deep learning based techniques. Recent thriving and development of neural network endow the significant capability to the deep learning based object detection models [Gandhi, 2018]. Popular techniques, represented by the RCNN family (RCNN, Fast RCNN, Faster RCNN etc.) [Girshick et al., 2014a; Girshick, 2015; Ren et al., 2015] and the one step end-to-end methods (YOLO, SSD etc.) [Redmon et al., 2016; Liu et al., 2016] as well as the semantic segmentation methods (Mask RCNN etc.) [He et al., 2017], have performed remarkable ability to capture objects under a variety of complex environments. However, on the ground of my experiments and analyses, most of those approaches are designed to detect targets in the natural scenes, but they can hardly fit directly into the detection task of the artificial graphic human-computer interface, such as the user interface (UI) components detection. Multiple factors contribute to such unsatisfied performance in this mission, including the specified requirements and practices when an user interface is designed, and the essential mechanisms of deep learning that are inappropriate to be used in this case.

This chapter summarizes these particular characters in the human-computer interface and briefly introduces the principles and problems of some state-of-the-art deep learning based object detection methods. Then the conclusion about the necessity of the domain-specific approach is drawn at the end on the basis of analyses and comparisons.

3.1 Characters of the Human-computer Interface

Based on the statistics, we observe that the artificial interface have some common visual attributes that differ from the natural scenes. These distinctions make the conventional deep learning approaches hard to perform their abilities as effectively as they do in natural images.

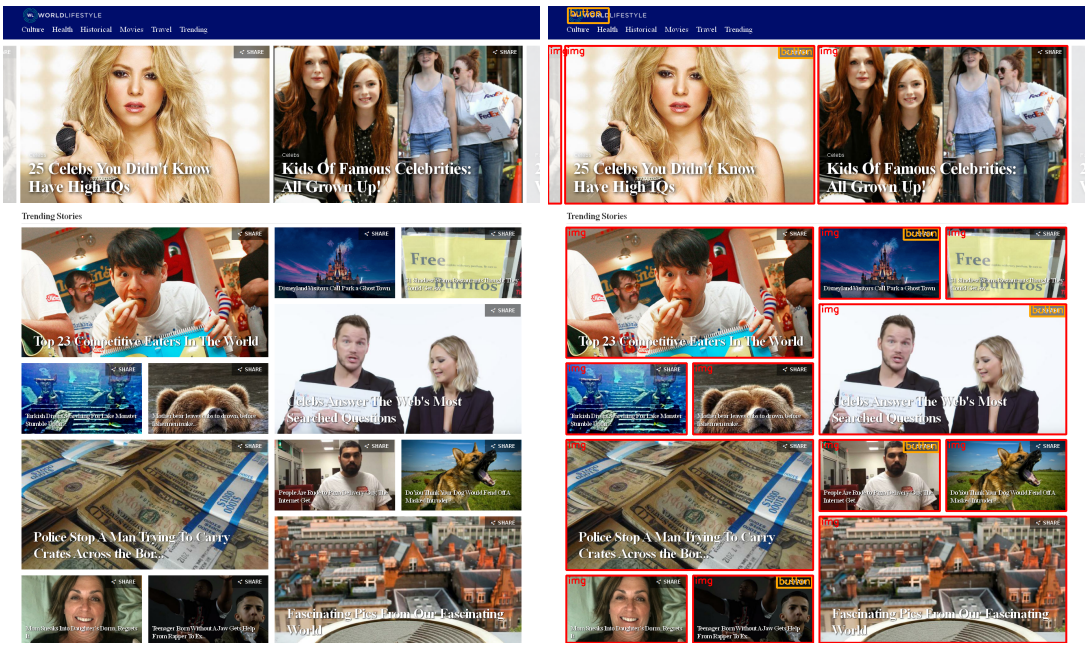
Property 1: The contents in the graphic user interface are heterogeneous. To be more specific, an interface design could contain various components, including the functional elements, such as the button and input box, and the static resource that is responsible for displaying information, such as the image and text. The detailed

categories are defined in the table 5.2 of the Chapter 5. The real challenge lies in the significant diversity of individuals in the same group, especially for image components as shown in Figure 3.1(a). That is, the contents of an image elements can be literally everything. For example, a selfie can be put on the interface as an image component, while a group picture that contains plenty of people can also be regarded as a single image, as well as a natural scenery photo or even a cartoon illustration can become an image element on the interface. It is also possible that an individual element is a clipping section of an image. In other words, the variation in the same class can be significant, which differs from the conventional object detection tasks that identify the more or less similar targets as the same class.

Property 2: The components on an user interface are picked. A variety of components might be allocated in a compact layout, and they would also overlap with or superimpose on others. It is common to put some buttons and text on the background image where the colourful and various contents is displayed. Besides, some independent elements sometimes should be treated as a whole. For instance, a special object named image button is widely used in the mobile application interface, in which a piece of text is placed in the centre of an background image in shape of rectangle or oval. Detection in this case requires accurate component segmentation that does not separate the integrated objects while identifies it as a whole. In addition, some mobile applications that work on a small screen are designed in a tight style where the elements are close to each other, which rises the difficulty for precise localization of UI components.

Property 3: The user interface component's shape is arbitrary. This property is especially for the width, height and aspect ratio. Although there are some rules for user interface design in terms of the size and shape, as shown in Figure 5.10, the elements' character in same class can still vary to a large degree. As the first attribute, this variation poses a great negative influence on the accuracy of localization, because most deep learning based object detection methods achieve localization by bounding box regression [Girshick et al., 2014b; Lee et al., 2019]. Regression essentially is a statistical approach modeling the relationship between a dependent variable and one or more explanatory variables [Freedman, 2012]. For example, in linear regression, the relationships are modeled by the linear predictor functions estimated from the known data. However, if the data is too dispersed, the estimated functions are hard to be accurate and effective [SEAL, 1967]. Thus, localizing methods that rely on the bounding box regression is not robust in human-computer interface components detection.

Property 4: The position and boundary of components in an user interface demand of absolute precision. In web and mobile application development, developers implement the components in the way that sets the accurate size (width, height or aspect ratio) and places them in an exact position (pixel distance from the boundary). This character asks for as accurate the detection of components on the interface as possible. But as mentioned in the third property, the bounding box regression that deep learning based object detection methods use is a statistic estimated function, which is inadequate to predict the one hundred percent precise result.



(a) Heterogeneous contents of images on web page (b) Desirable detection result

Figure 3.1: A real web interface design. The 3.1(a) represents a common case in human-computer interface design. Various colourful and heterogeneous images are used to display information, in which the contents can be everything and may confuse the neural networks. Figure 3.1(b) is the result of my approach using inventive image processing algorithms. It also shows the desirable result of UI components detection that each image should be regarded as an individual element (labeled with red bounding box), in stead of oversegmenting its detailed contents. Despite the complexity, those components should be precisely identified and localized for further code generation.

Therefore, the component detection in this case requires some domain-specific approaches that adapt to the particular properties of the user interface. Since the analysis through the data and the popular deep learning methods draws the conclusion that the end-to-end neural networks do not well fit into this mission, I propose a technique that combines with some conventional computer vision and image processing algorithms to accommodate the specified characters in the artificial interface design. Before introducing my own approach, comprehending the mechanisms of deep learning based object detection techniques is critical to bypass their defects.

3.2 Deep Neural Network's Mechanism

Although the deep learning approaches are dominated in contemporary computer vision field, some natural deficiencies of those techniques still exist and the flaws

indeed cause some issues. Some of the defects are exposed in the mission of human-computer interface component detection, as stated previously. I dig into several popular object detection methods to try to summarize their mechanisms and attempt to reveal the inadequacies.

Generally, two directions of object detection techniques are mostly studied, the region proposals (RCNN, Fast RCNN, Faster RCNN) and single shot methods (SSD and YOLO) [Gandhi, 2018]. Those approaches are proven effective in natural scenes, and they are so inspiring that a wide variety of derivative techniques refer to them. This section briefly summarizes the their principles and analyzes the flaws that prevent them from performing well in human-computer interface components detection.

3.2.1 Region-based Methods

The reason why we cannot directly build a classic CNN followed by a fully connected layer to proceed object detection is that the length of the detection's output is not fixed [Gu et al., 2018]. That is, the number of occurrences of targets is variable. To address this issue, an idea is that selects various regions of interest from the input image and classifies them respectively with a CNN to inspect the presence of objects in the regions [Gandhi, 2018]. The region proposal approaches derive from this idea and adopt some strategies to alleviate the computational problem that the huge number of regions in different spatial locations and various aspect ratios.

Region-based Convulutional Neural Network: To solve the problem of selecting a huge number of regions, Ross Girshick et al. proposed a technique that utilizes the selective search [Uijlings et al., 2013] to extract smaller amount (always 2000) of regions from the input image. They are named region proposals. Thus, the number of regions that need to be classified reduce to 2000. Then, those regions are resized and thrown into a convolutional neural network followed by a dense layer of 4096 in size at the end. The network hence output a 4096-dimensional feature vector for each region. Those vectors are fed into a pre-trained Support Vector Machine (SVM) to classify the presence of objects. The RCNN uses the bounding box regression [Felzenszwalb et al., 2010] to increase the accuracy of object localization.

Several problems existing in the RCNN. First, although it reduces the number of regions, but it still has to classify 2000 region proposals. Second, as the long time (47s average) it takes to process all regions, it can hardly be used real time. Third, it locates objects by using bounding box regression, but as mentioned in the last section, this method cannot predict the actual outline and precise location of objects but just perform statistical regression.

Fast RCNN: Ross Girshick et al. then proposed the Fast RCNN to solve some of the drawbacks of the previous version RCNN. The improvement of this approach is that it builds a CNN to generate a convolutional feature map by feeding the input image, in stead of processing the thousands of region proposals every time on the original image. The Fast RCNN uses a region of interest (RoI) pooling layer to

recognize the region proposals from the feature map and resizes them into squares to be the input of the final fully connected layer. In the end, those RoIs are turned into feature vectors, which then be fed into a softmax layer to predict the classes of regions, as well as inputted into a bounding box regressor to localize the accurate positions of objects.

Although the Fast RCNN makes some improvements especially in terms of the processing time, the essential mechanisms are similar to its predecessor. They all try to propose regions by some means and identify the presences and classes of objects in these regions. Meanwhile, they all utilize a bounding box regressor to predict the precise locations of targets.

Faster RCNN: Both the two previous techniques apply the selective search to pick region proposals. One drawback of the selective search is that it is a time-consuming process, as well as a fixed algorithm which can not learn for itself like the neural network. In order to bypass this shortcoming, the authors of the Faster RCNN proposed a region proposal network (RPN) [Ren et al., 2015] that is capable of learning the region proposals. Similar to the Fast RCNN, this method uses the CNN encoder to generate feature maps from the input image. Then the RPN is applied separately to predict the region proposals, which followed by a non-maximum suppression (NMS) [Canny, 1987] to refine the resulting predictions.

In this process, one novelty is that the Faster RCNN adopts a variety of anchor boxes in different sizes and shapes at each sliding-window location. The RPN predicts k proposals simultaneously at each location. The prediction consists of the *regression* layer which has totally $4k$ outputs presenting the coordinates of k boxes and the $classification$ layer where the outputs are $2k$ scores to estimate probability of object/non-object of each proposal. Those k proposals are encoded into k anchor boxes that are centered at the sliding window.

Therefore, we can observe some common grounds from the series of region based approaches. One salient similarity is that they all propose multiple regions and conduct prediction and regression on those regions independently. On the other hand, another branch of deep learning based object detection approaches try to process the input image as a whole other than focusing on local section of the image, and they complete the detection in a single convolutional neural network. To some degree, they convert the detection problem into regression problem, and hence also are called regression-based methods [Zhang et al., 2019].

3.2.2 Single Shot Methods

This section analyses two representatives of the regression-based approaches [Zhang et al., 2019], the You Look Only Once (YOLO) proposed by Redmon et al. and the Single Shot Multibox Detector (SSD). They all perform object detection by using a neural network to an entire image and pose the detection problem as regression problem [SACHAN, 2017].

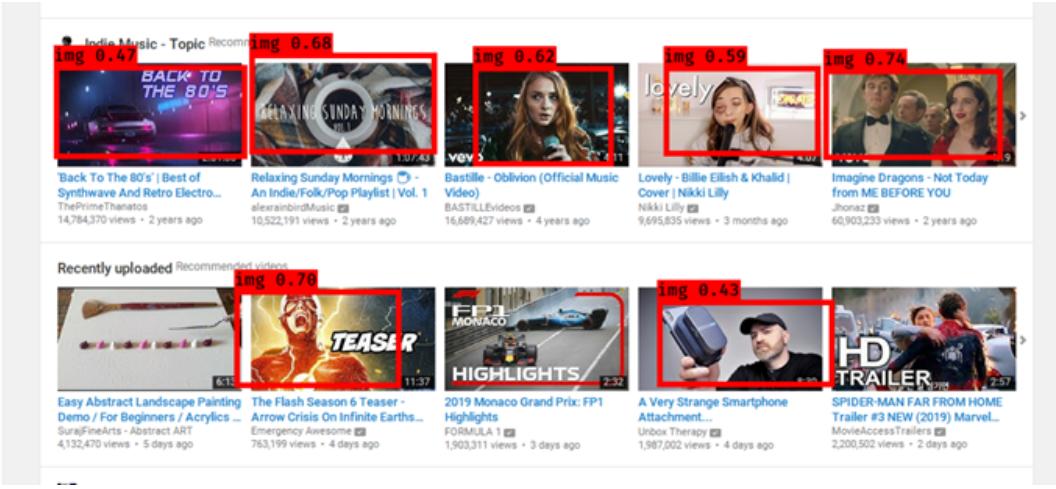


Figure 3.2: The detection result of YOLO. The predicting images are labeled in red bounding boxes, from which we can observe that the accuracy of the localization cannot fit the strict requirement in real User interface design, as stated in *Property 4*. The main reason for the deficiency is that the localization is achieved by the regression of the offset values of bounding boxes.

YOLO: Essentially, the YOLO achieves detection by learning the class probabilities and the coordinates for each bounding box. This technique splits the input image into S x S grids, in which each grid cell has B bounding boxes. The outputs of this model are class probabilities and position coordinates of the bounding boxes. Then those whose probabilities are above a threshold are chosen and used to localize the objects within them. In the training stage, if the center of a ground truth object falls into a grid cell, this cell is responsible for detecting it in the way that finds the bounding box with the maximum intersection over union (IOU) over the true object.

SSD: The single shot multibox detector, as its name suggests, is also a single-shot detector for multiple categories which is faster than its previous state-of-the-art technique. The speed advantage mainly comes from the design that does not resample the bounding box or generate proposals as the RCNN series do. The improvement in accuracy attributes to combination of predictions of multi-scales feature maps with various resolutions, which handles the problem of variation in object's size. This method still use a fixed set of default bounding boxes, and it apply small convolutional filters applied to feature maps.

3.3 Conclusion

We analyze the four particular characters of the data we aim to process and the mechanisms of the state-of-the-art deep learning based object detection techniques in the

preceding parts of the text. This section combines the observations and draws the conclusion that the new domain-specific approach is needed to satisfy the requirements of detection task in human-computer interface.

In summary, the object detection task consists of two major parts, the localization of the targets and the classification. As mentioned in the beginning of the section 3.2.1, the key difficulty lies in how to find the objects accurately and efficiently, while the classification is already well handled by CNN. To this end, a variety ideas have been proposed, and two directions are most widely studied among them, the region-based and the one shot methods.

The primitive region-based approaches, such as the RCNN and the Fast RCNN all adapt the selective search algorithm to propose potential regions where object might be. However, some of their drawbacks prevent them from working well in the human-computer interface. First, the selective search depends on the similarity of regions, while the components in an interface can be heterogeneous as the *Property 1* analysed in section 3.1. Thus the effectiveness of the selective search is undermined. Second, because of the *Property 1* and *Property 2*, it is very easily that the components are oversegmented, especially for the image components that contain colourful and various contents. Oversegmentation is a common issue for region segmentation methods, which means integrated objects are segmented into multiple sub-components by mistake [EGGLESTON, 2015]. Consequently, the single image component would be split into many sections that will be counted as false positive.

The Faster RCNN avoids using the fixed selective search algorithm and turns to a learnable RPN to propose potential regions. In detail, the anchor boxes are responsible for presenting the regions and fed into regression layer and classification layer. However, as the analysis in the *Property 3* and *Property 4* of human-computer interface, the huge variances of shape in the UI components hamper the effectiveness of logistic regression, while the highly precise localization is required.

Same problems exist in the one-shot methods. Because they are fundamentally regression based approaches, they still suffer the issues brought by the *Property 3* and *Property 4* when applied in the artificial interface. In addition, the oversegmentation is also a challenge for all methods using default bounding box hypothesis. The heterogeneous contents of the image element always affect the object judgement of the bounding box and produce misrecognition. For example, improperly identifying a part of image as an individual interface component.

Therefore, simply relying on machine learning is inadequate to satisfy the demands in the human-computer interface detection. Interestingly, the seemingly primitive image processing algorithms fit well into this task. Several reasons related to the unique characters of the UI interface cause this result, including the pixel-level processing and the insensitivity to changes after being conducted some approaches. I will elaborate them in the later sections with technical details.

3.4 Comparison

Due to a lack of time, I did not reproduce all the aforementioned deep learning based approaches, but just compared my method with the YOLO to illustrate the result. This YOLO model is on the basis of VGG16, I retrained the last 10 layers of this network through totally 13230 images with five different UI compoennt classes refer to table 5.2. Several results are visualized below:

Chapter 4

Data Collection

First of all, the initial step of this project is to inspect related data and to build the datasets for further analyses and experiments. Particularly, to meed the specific goal of this project, the datasets should consist of a variety of images of graphic user interface (GUI) designs, which should include both website data and mobile application data. Thus, this chapter introduces the methodologies and issues in the data collection process, as well as the utilization of some published datasets.

4.1 Web Page Dataset

The objective of this project is to detect the components in the human-computer interface or user interface (UI). The general input data should be an interface design image , such as a screenshot of a real web page. Therefore, I collected such screenshots by crawling over a variety of websites to build the web page dataset, the element's annotation are also gathered in this process if possible. In detail, the annotation is directly fetched from the source code of website, which consist of the tag name (class), size (width and height) and location (coordinates) of the element. However, various issues occurred and made this process far more difficult than it looks at the first glance. This section presents those problems and gives solutions to settle or bypass them.

4.1.1 Dataset Construction

I designed a system that automatically mines the screenshots of websites by using a Python based web application testing framework, Selenium. Generally, the web crawler is chosen to collect data from numerous websites, but the drawback of conventional crawling agent hampered me from doing so.

Web crawler stands for a program script that systematically browse the Internet. It usually takes a list of initial links (URLs) as the so-called *seeds*, [Kobayashi and Takeda, 2000] then visits them one by one and downloads the web page for further analysis. The downloaded web page is also known as the document object model (DOM) with a tree structure [Whitmer, 2009]. All kinds of web component information, including the hierarchy, size and location of the element can be retrieved by parsing the DOM tree.

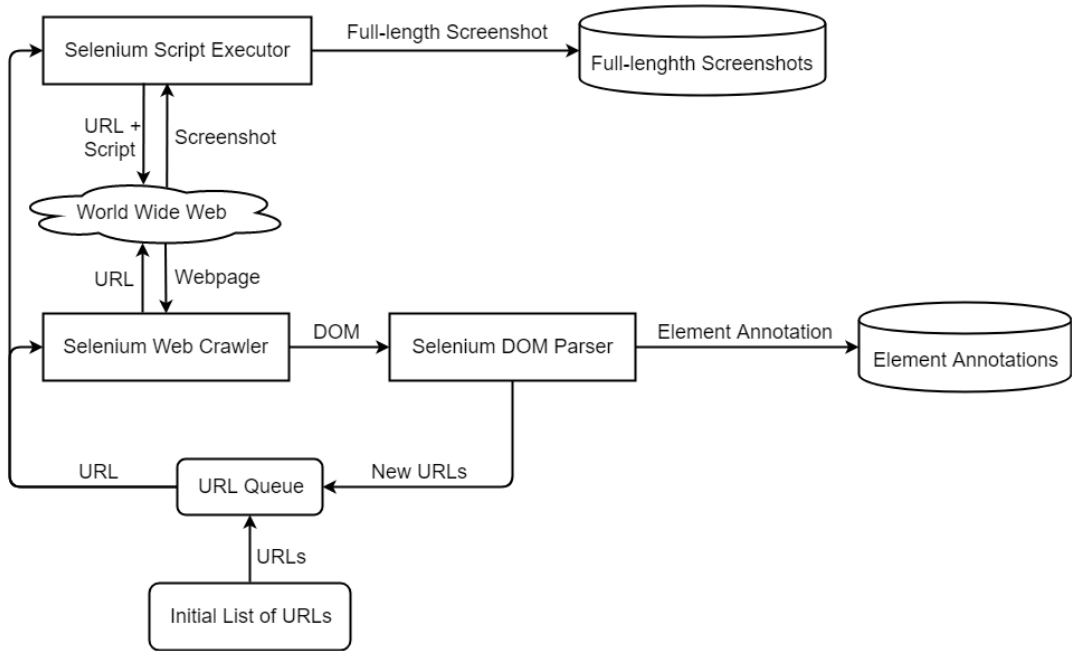


Figure 4.1: The architecture of the data collection system. The initial list of URLs are pushed into the URL queue first, and then the Selenium Web Crawler and the Selenium Script are fed with a new URL pop up from the queue. The Script Executor runs the script to scroll over the web page on the World Wide Web and generates a full-length image. Meanwhile, the Crawler downloads the DOM file of the web page and conveys it to the Parser to retrieve the elements' annotations. In the end, the dataset consisting of screenshots and corresponding annotations is generated.

A common limitation for web crawlers, such as the *BeautifulSoup* [Richardson, 2015], is that they can only conduct parsing on the basis of the static documentation, while they do not support interaction with the website. In other words, they work by means of downloading the DOM and then analyzing the DOM file without any communication with the web. But the data we desire is the full-length screenshot of pages, which cannot be directly obtained by normal crawling agents. Therefore, I turned to the test automation [Moizuddin, 2019].

4.1.1.1 Selenium

Originally, test automation in software testing is defined as the using of software separate from the tested one to execute a series of commands and compare the actual outcome and the predicted outcome [Huizinga and Kolawa, 2007]. In our case, I leverage this feature to execute the script that scrolls down the page to the bottom to obtain the full-length snapshot.

Selenium is a software developed for web test automation. It is composed of several components for aiding the development of web application test automation.

Its own independent integrated development environment (IDE) is implemented as a Chrome Extension, which supports recording, editing and debugging of web application tests. Selenium also has some APIs for various programming languages such as Python and Java. In order to convey messages between the browser and the client, Selenium developed the *WebDriver*, a module accepting the commands via the API and passing them to the browser. The WebDriver is specific to browsers in the way that it sends specified instructions to them to run and retrieves the results. [Moizuddin, 2019].

Selenium can well perform the full page screenshot capture by executing the scripts, as well as the DOM retrieval by web crawling. Besides, it has been implemented as a package in Python which is convenient to be incorporated into the whole UI2CODE system. Hence I built the data collection module by it.

4.1.1.2 Breadth-first Search

Generally, there are two strategies for web crawling, the breadth-first search and the depth-first search [Kobayashi and Takeda, 2000]. The depth-first search in this context is the common practice for crawlers. In this way, the crawling agent visits the first initial URL and retrieves all the new links in this page. The new links are pushed into the top of a stack. After completing the process of the current page, the next URL is pop out and parsed, and the new links on that site are pushed into the stack again. This process repeats until the stack goes empty or reaching an end condition. On the other hand, the breadth-first means visiting through all the initial list of URLs first and adding the new links fetched from the crawled websites to the end of a queue. The new links will not be accessed until the given ones are all visited.

The breadth-first search strategy is adopted. We want to collect interface designs as various as possible, so that the components detection techniques can be thoroughly tested and developed. Thus, the crawler does not go deep of a single website, but visits various websites as many as possible.

In summary, the data collection system is composed of three modules: the web crawler responsible for DOM file downloading from the given link; the parser takes the DOM file as input and parse it to retrieve element annotations, such as tag, size and position; the script executor run a piece of JavaScript code to scroll through the web page to take the full-length screenshot. The entire process is demonstrated in Figure 4.1.

4.1.2 Problems in Web Crawling

4.2 Mobile Application Dataset

Chapter 5

User Interface Components Detection

UI2CODE is a system automatically recognizing the semantic contents of the input image of human-computer interface and generate the front-end code that can implements the same visual effect and expected functionalities of the given image design. To this end, I propose a precise and purpose-built user interface components detection pipeline that works particularly better than popular objection detection methods [Ren et al., 2015; Redmon et al., 2016; Redmon and Farhadi, 2018] in graphic human-computer interface.

This technique utilizes computer vision and image processing algorithms to detect and locate objects. After selecting the candidate regions where are likely to be UI components, a classifier is built to recognize those areas and categorize them into different classes, such as *button, input bar, images and text*. Meanwhile, the text recognition is achieved by an effective Optical Character Recognition (OCR) technique, the Connectionist Text Proposal Network (CTPN) [Tian et al., 2016]. In the end, the results from those modules are merged and refined to produce the final result.

Performance of the image processing technique compared with machine learning methods, especially in our task, is significantly better in terms of accuracy of components position detection and recall. Detailed reasons for those strengths are stated below, one of the most interesting interpretations is that the deep neural network observes the texture rather than the shapes of objects [Geirhos et al., 2019; Cepelewicz, 2019; Geirhos, 2018], which arouses some deep-going thoughts about the nature of deep learning methods, and I elaborate those thinkings in Chapter 3.

5.1 Architecture

I assume the input to this tool to be an image of an interface, which can be either a screenshot of a real application (web or mobile) or a conceptual design drawing. We focus on segmenting and classifying the possible human-computer interface components on the image.

Based on the common practice of developer and the properties of the front-end programming languages, three categories including totally five types of graphic user

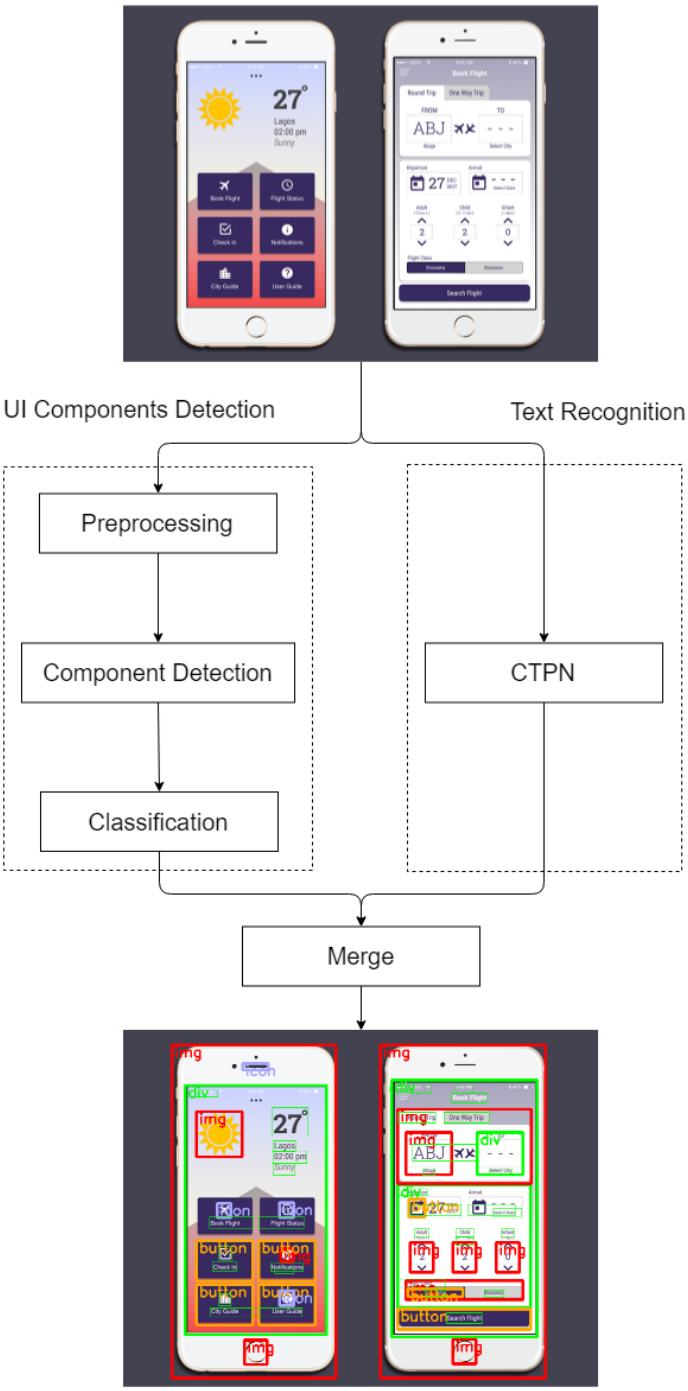


Figure 5.1: The architecture of the UI components detection pipeline. The input image here is a conceptual design drawing of a mobile application, and it is processed by two independent branches to segment the UI components and detect the text regions respectively. Then the results are merged and refined to get the final result.

interface components are defined in this process: interactive elements *(button, input box)*, static resource *(image, icon)* and layout structure *(block)*, see Table 5.2.

To perform precise segmentation, I implement this tool as a three-phase pipeline consisting of pre-processing, components detection and classification, combined with text detection achieved by CTPN. The results from both branches are then integrated at the end to produce the final prediction. The illustration of the overall architecture is shown in Figure 5.1.

5.2 Pre-processing

The first step for of the UI components detection pipeline is pre-processing. It transforms the input image into specific form that is convenient for further processing. Three substeps are conducted here: gray-scale image conversion, gradient calculation and binary image conversion.

In our task, we treat all contents on the image as parts of individual components without consideration of their own detailed information. For example, an image on an interface design should be regarded as a single element instead of a combination of the real contents in it, as shown in Figure 5.2.

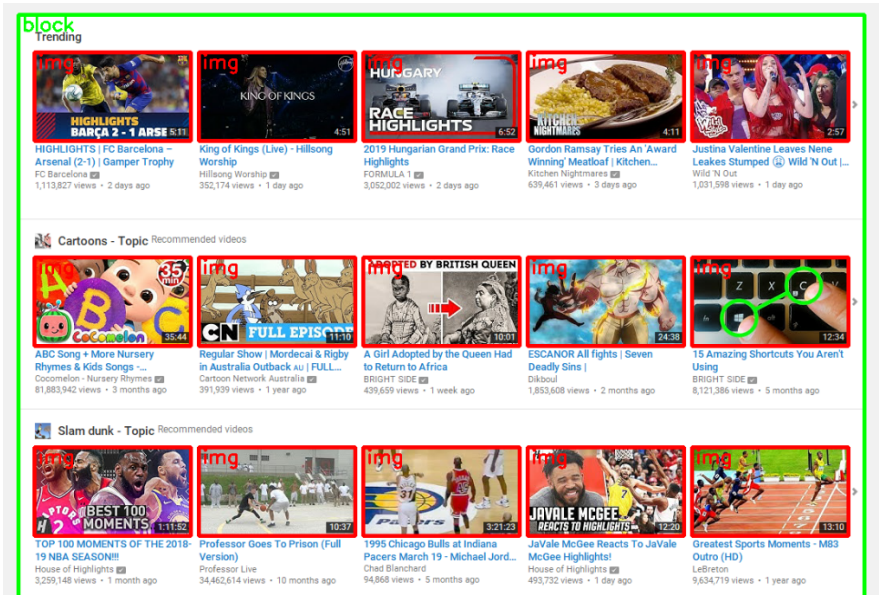


Figure 5.2: A section of screenshot of YouTube website. Plenty of image components (labeled in *img* with red bounding boxes) with colourful contents appear on this user interface, but we want to leave out the information of the real contents and treat them as parts of individual UI components.

To this end, I try to find a means to convert the colorful and complicated image into a simple form that does not contain redundant information this task does not need and is convenient to segment components. The popular related algorithms, such as Canny edge detection [Canny, 1986] and findContour method in OpenCV

based on techniques proposed by Suzuki et al. [Suzuki and be, 1985; Team, 2012], do not work well in this case, because those processing always leave the texture details and disconnect the contents in an image, as shown in Figure 5.3. So, I propose a new method to satisfy this purpose.

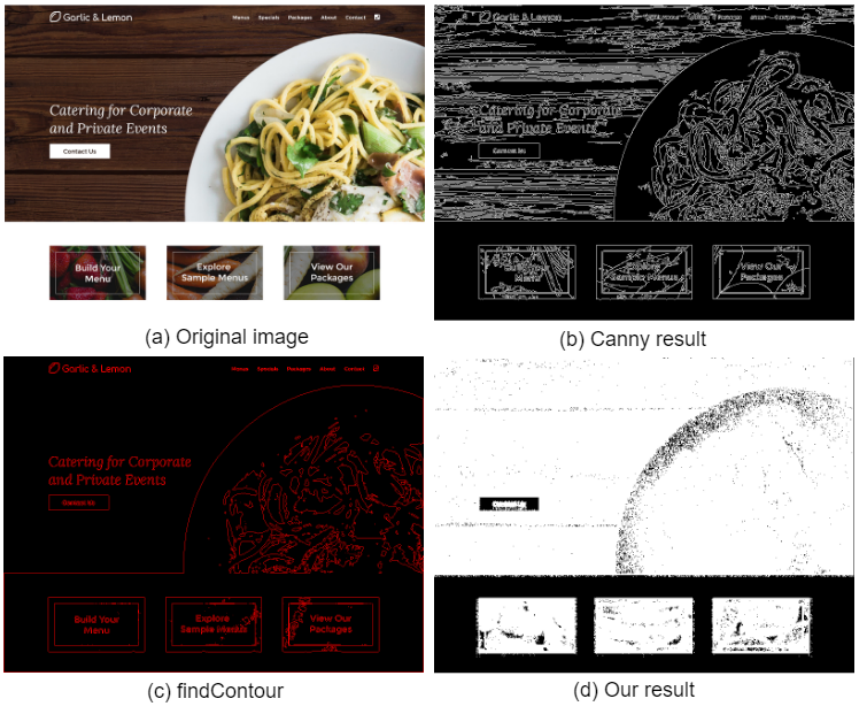


Figure 5.3: The picture (a) is the original image; the image (b) is the result of Canny algorithm, which extracts too many details of texture; the picture (c) is the result of findContour function in OpenCV library, and it focuses on calculation of the boundary of objects; (d) is the binary image processed by our method, which convert the components to a simple binary image consisting of few integrated objects without too many redundant texture information.

5.2.1 Gradient Calculation

The gradients of a digital image measure how each pixel changes in terms of significance and direction [Jacobs, 2005]. Popular techniques of image gradient calculation are Roberts cross operator, Prewitt operator, and Sobel operator [Roberts, 1963; Prewit, 1970; Sobel, 1968]. We can acquire two pieces of information from the gradient of each pixel, the direction of the change and the magnitude of this change.

However, unlike other common computer vision tasks that deal with the natural scene, we do not care for the changing direction as much as about the magnitude, because we focus on determining whether a pixel is a part of the potential components rather than the detailed information of how it changes. Therefore, we calculate

the magnitude of gradient by formulae below:

$$gx = \frac{\partial f(x,y)}{\partial x} = f(x+1,y) - f(x,y) \tag{5.1}$$

$$gy = \frac{\partial f(x,y)}{\partial y} = f(x,y+1) - f(x,y) \tag{5.2}$$

$$G(x,y) = |gx| + |gy| \tag{5.3}$$

where: f(x,y) is the pixel vale for point (x,y) in the image; gx and gy are the gradients in the x direction and y direction respectively; G(x,y) is the magnitude of gradient value at pixel point (x,y).

The result of this step is a grey-scale map (a two-dimensional matrix in which the value of each pixel is on the scale of 0-255) reflecting the significance of gradient of the original image, as shown in Figure 5.4(b).

5.2.2 Binarization

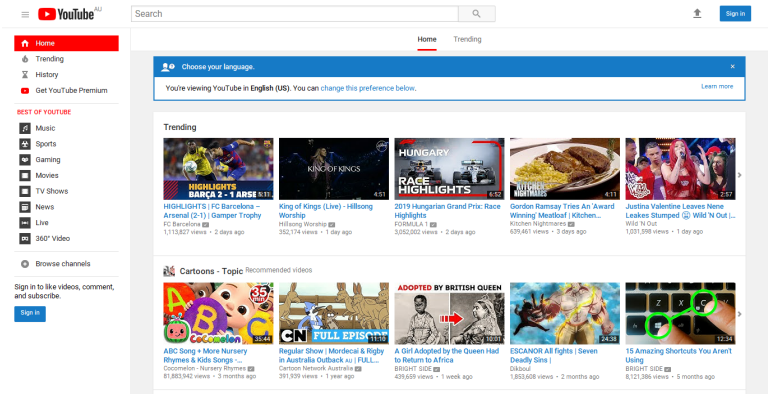
The second substep of pre-processing is to binarize the gray-scale map, the purpose of this process is to intensify the component regions from the background.

One particular observation on graphic user interface that differs from natural scene is that the regions where there is little or no gradient change are more likely to be background. On the other hand, pixels with large gradient should be parts of the foreground objects (interface components). With this regard, I set a small gradient threshold to label each pixel as either foreground or background.

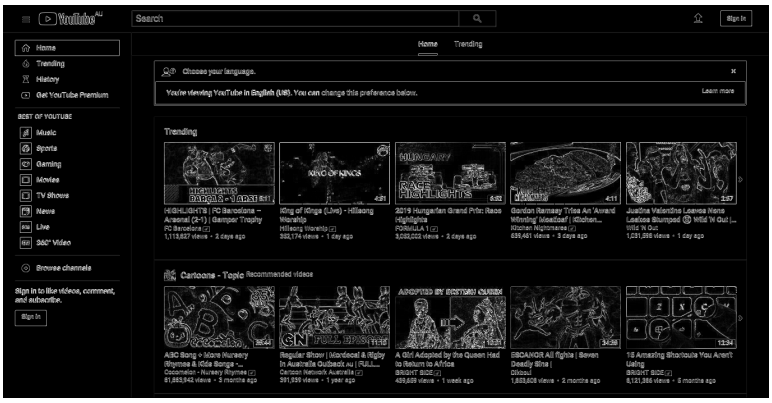
The goal of binarization is to assign a binary value to every pixel in an image [Shapiro, 2002]. T.Sezgin and Sankur summarized the thresholding methods into six categories: the histogram shape-based , clustering-based, entropy-based, object attribute-based, spatial methods, local methods [Mehmet Sezgin, 2004]. My technique is similar to the Entropy-based methods, which use the entropy of the foreground and background regions. Specifically, as mentioned above, the gradients of the pixels in the foreground are usually larger than zero, while the background of graphic user interface is always unicolor and the gradients are zero. In our case, this step assigns either 255 (white) or 0 (black) to each pixel; points whose value is 255 means could be parts of an interface component; value 0, on the contrary, means this point is part of the background, as demonstrated in Figure 5.4(c).

But for different datasets, the gradient property would be slightly different, which requires adjustment of the threshold. For instance, the images in Rico dataset are more compact than the screenshots of real web pages, so the threshold should be slightly higher to better segment regions; and the images of Google Play Poster and Dribbble datasets can be lower resolution than are web pages, so the background is more blurred and its gradient can be higher than zero.

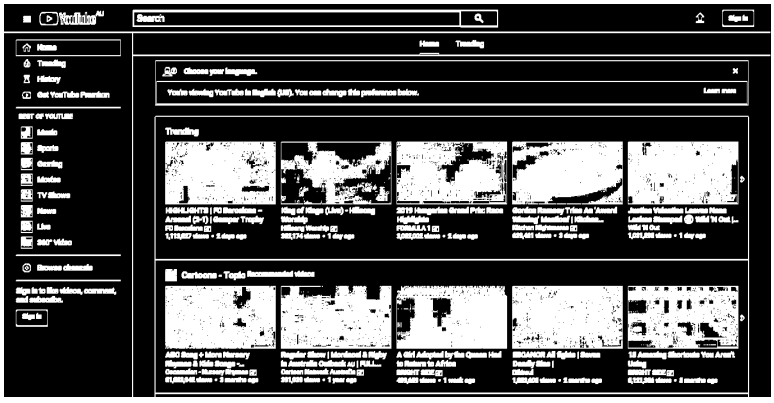
In the end, a binary map that contains clear foreground objects is produced by the pre-processing for further operations.



(a) Original input image



(b) Gradient map



(c) Binary map

Figure 5.4: The visualized demonstration of the pre-processing. The original image Figure 5.4(a) is given as the input; and the process calculates the magnitude of the gradient for each pixel to produce a gray-scale map Figure 5.4(b); then according to the observation of foreground and background in the human-computer interface, a binary map Figure 5.4(c) is generated

5.3 Component Detection

Based on the acquired binary map, this stage tries to extract component candidates and heuristically categorize the connected regions that could be potential user interface elements. To this end, this process contains several substeps: Connected Components Labelling, Component Boundary Detection, Rectangle Recognition, Block Recognition, Irregular Components Selection and Nested Components Detection.

Three sets of objects yield from this process, *Block, Image and Interface Components*. Detailed categories and classes of UI components are defined in section 5.4, at this stage, the connected components are only grouped to three aforementioned general classes according to some heuristic rules based on the size and aspect ratio in order to save processing time in the next step.

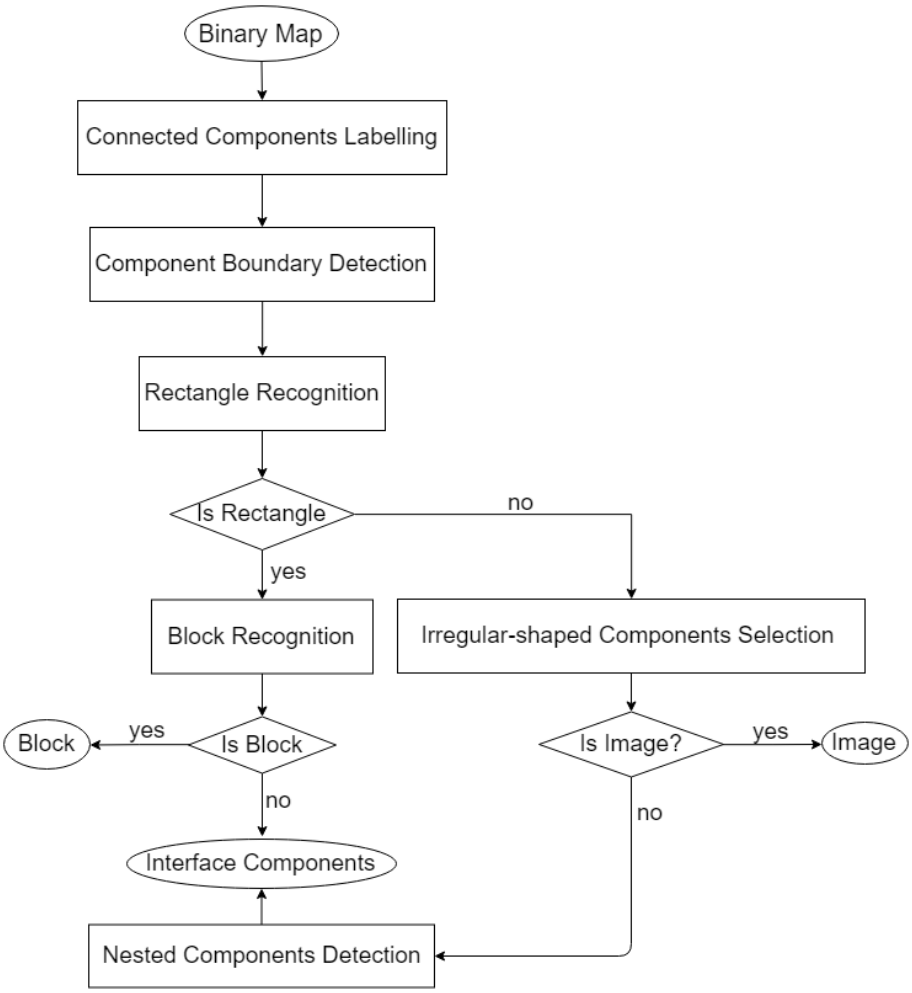


Figure 5.5: Flow chart of the Component Detection pipeline. This pipeline takes binary map as input, and the result of this step consists of three categories of UI elements: *Block, Image and Interface Components*

5.3.1 Connected Components Labelling

This process refers to the connected-component labeling (CCL) algorithm, which is demonstrated in Figure 5.6, the purpose is to assign each pixel a label identifying the connected component to which this pixel belongs [Samet and Tamminen, 1988; Dillencourt et al., 1992]. In other words, this substep aims to segment connected components form the binary image.

I refer to the Seed Filling algorithm in computer graphic [Vincent and Soille, 1991] to implement my own method. The pseudocode of this technique is shown in Algorithm 5.3.1.

Algorithm 1 Connected-component labeling

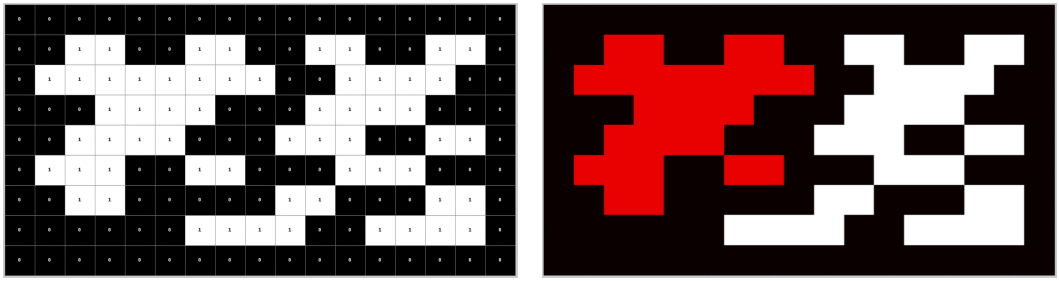
Input: Binary map
Output: An array of components, each component contains a group of points that constitute it

```
1: Components ← [ ]
2: MarkingMap ← Zeros(BinaryMap.shape)
3: for Point in BinaryMap do
4:     if Point is 255 and MarkingMap[Point] is 0 then
5:         MarkingMap[Point] ← 1
6:         Neighbors ← new Queue.push(Point)
7:         Component ← new Stack.push(Point)
8:         while Neighbors.Length > 0 do
9:             NextPoint ← Neighbors.pop()
10:            for Neighbor in NeighborsofPoint(NextPoint) do
11:                if Neighbor is 255 and MarkingMap[Neighbor]is 0 then
12:                    MarkingMap[Neighbor] ← 1
13:                    Neighbors.push(Neighbor)
14:                    Component.push(Neighbor)
15:                end if
16:            end for
17:        end while
18:        Components.push(Component)
19:    end if
20: end for
21: return Components
```

Alternatively, this algorithm can be written as:

(1) Start from the top-left point, initialize a MarkMap with the same shape as the input image and fill it with zeros to indicates if points are already labelled, go to (2).

(2) If this pixel is foreground (its value is 255), and it hasn't been labelled (MarkMap is zero at this position), then add it to a store queue and count it as a point of a new component, and go to (3); otherwise repeat (2) for the next pixel in the binary map.



(a) Binary map (b) Connected-component labelling result

Figure 5.6: Connected-component labelling demonstration. The 5.6(a) shows foreground (white points) and the background (black points); the 5.6(b) is the CCL result, where two connected components are labeled in red and white colour.

(3) Pop out an element from the store queue, inspect all of its neighbors. If the neighbor is foreground and hasn't been labelled, then add this point into the store queue and count it as a point of current component. Repeat (3) until the store queue is empty then go to (4).

(4) Scan the next pixel in the binary map and go back to (2) until all pixels in the input image are inspected.

As described in above algorithm, the output of this function is an array of components. The way I store a component is presenting every connected component as a list of points (coordinates), where the points are all connected foreground in each list.

5.3.2 Component Boundary Detection

For each connected component, I calculate its outer boundary by a four-border method. The resulting boundary consists of four borders: border-top, border-bottom, border-left and border-right.

Because the sole information of a component acquired from the last step is the points that constitute this component, so for the sake of efficiency, I implement this method by means of storing the borders in form of $\{position : boundaryvalue\}$ through dictionary data structure, *position* is key of the dictionary and *boundaryvalue* is value. For instance, the format of points in $BorderTop$ and $BorderBottom$ is $\{column : maxrow\}$ and $\{column : minrow\}$, which indicates the vertical boundary values of this column in this component. In the same way, points in $BorderLeft$ and $BorderRight$ are stored as $\{row : mincolumn\}$ and $\{row : maxcolumn\}$ to show the horizontal boundary value of each row.

In Algorithm 5.3.2, the Borders are initialized as four empty dictionaries that will eventually be filled with points in the aforementioned form. $BorderTop[Column]$ means retrieving the vertical boundary value (minimum row index) at this column of the component, and if this value is larger than the row index of the current point

Algorithm 2 Four-border Boundary Detection

Input: Component consisting of points that belong to it
Output: Boundary containing four borders: top, bottom, left, right; Each border is a list of points

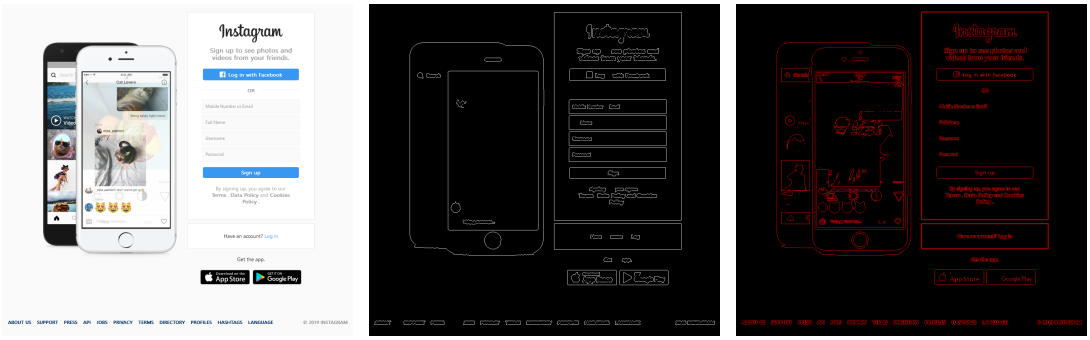
```
BorderTop, BorderBottm, BorderLeft, BorderRight ← { }, { }, { }, { }

for Point in Component do
    Row, Column ← Point[0], Point[1]
    if Column not in BorderTop or BorderTop[Column] > Row then
        BorderTop[Column] ← Row                ▷ Choose the smaller row as top
    end if
    if Column not in BorderBottom or BorderBottom[Column] < Row then
        BorderBottom[Column] ← Row             ▷ Choose the larger row as bottom
    end if
    if Row not in BorderLeft or BorderLeft[Row] > Column then
        BorderLeft[Row] ← Column               ▷ Choose the smaller column as left
    end if
    if Row not in BorderRight or BorderRight[Row] < Column then
        BorderRight[Row] ← Column              ▷ Choose the larger column as right
    end if
end for
Boundary ← [list(BorderTop), list(BorderBottm), list(BorderLeft), list(BorderRight)]
return Boundary
```

in the loop, then this point should be considered as the new top value at this column in the component, done by $BorderTop[Column] \leftarrow Row$.

For future convenience, the borders are transformed from dictionaries to list in the way that converts point information stored in the dictionary $\{position : boundaryvalue\}$ to a list $(position, boundaryvalue)$, so that all the points in the borders become the format $(position, boundaryvalue)$, which are better to retrieve and change in further process. This conversion is done by $list(BorderTop)$. Finally, this function returns a list of the four borders in order of top, bottom, left and right.

Unlike other popular contour detection algorithms, especially the findContour function implemented in OpenCV [Team, 2012], this task does not care much about the precise outer borders and hole borders for each object, because the purpose at this stage is to select the potential graphic interface components and filter out those unlikely to be interface elements, instead of acquiring the detailed texture information of each object. Besides, the findContour function is highly sensitive to parameters, which means this method is not robust enough and requires adjustment of parameters for various input images.

On the contrary, the four-border boundary detection algorithm is sufficient and more efficient in this case because it only calculates the outer boundary, and is not overly subject to parameters.



(a) Original image (b) Result of Four-border (c) Result of findCountour

Figure 5.7: Demonstration of the Four-border boundary detection. The 5.7(a) is the original input image from a web interface design; the result of this algorithm is shown in figure 5.7(b), which does not contain the fine grained details of the components but only the outer boundaries; the 5.7(c) shows the result of findContour function in OpenCV library, is detects more precise border of objects but the performance is unstable and sensitive of the parameters.

5.3.3 Rectangle Recognition

Another observation on human-computer interface is that most of the elements have regular shape. For example, pictures on a website are always be displayed in rectangle regions, and buttons and input boxes are usually round or rectangular. Thus, a rectangle detection algorithm is introduced as a heuristic process for interface components detection.

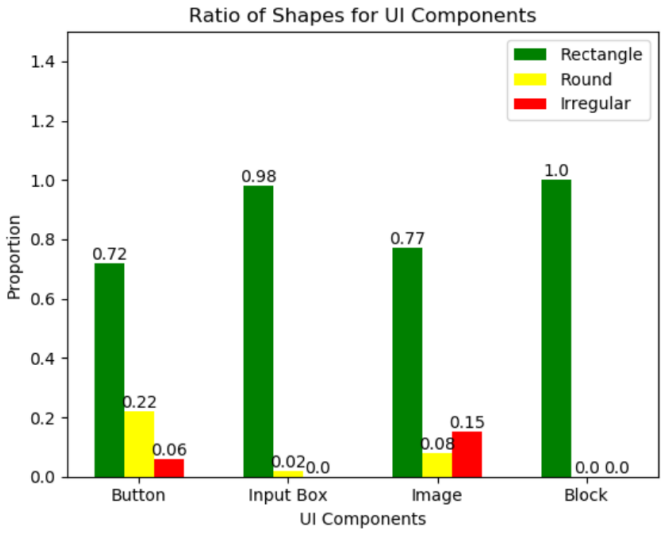


Figure 5.8: The proportion of UI components with different shapes.

The statistics is collected from three different datasets, in which the numbers of *Buttons*, *InputBoxes*, *Images*, *Blocks* are 10566, 3460, 39998, 1568 respectively. The

statistics is consistent with the observation mentioned before, a large proportion of human-computer interface components are in rectangular shape, which proves the necessity of the rectangle detection process.

Existing techniques, such as approxPolyDP in OpenCV [Team] library and Hough transform [Duda and Hart, 1972], are too complicated and rather unnecessary in our task. We only estimate whether the component is a rectangle or not, but the approxPolyDP method based on Douglas-Peucker algorithm [Douglas and Peucker, 1973] involves too much computation to calculate the precise polygonal curves. Hough transform is also too computationally expensive because it examines four parameters to detect a rectangle, which projects the information into a four-dimension computation space.

Therefore, I propose a simple and efficient method to detect rectangle by calculating the smoothness of boundary. In addition, this method measures the dentation to filter out concave objects. The pseudocode is show in Algorithm 5.3.3.

Algorithm 3 Rectangle Detection

Input: A component's boundary consisting of four borders: top, bottom, left, right
Output: Boolean value to indicate if this component is rectangular shape

```
1: smoothBorderNo ← 0
2: smootness ← 0
3:
4: for Border in Boundary do
5:     for i in range(Border.length) do
6:         difference ← Border[i] − Border[i + 1]
7:         if different = 0 then
8:             smootness ← smootness + 1
9:         end if
10:    end for
11:    if smoothness/Border.length > 0.8 then
12:        smoothBorderNo ← smoothBorderNo + 1
13:    end if
14: end for
15: if smoothBorderNo = 4 then
16:    return True
17: else
18:    return False
19: end if
```

In this implementation, *Boundary* is an array of size four, which contains four borders for top, bottom, left and right directions. For each border, $Border[i]$ means the ith element in it, and $Border[i+1] - Border[i]$ calculates the variance or gradients between two adjacent points in the same border, which indicates the smoothness of this border.

5.3.4 Block Recognition

We define a bordered region enclosing various multiple elements as block, Figure 5.9 presents two examples. As explained in the section 5.4, block is a layout structure, which could be regarded as a frame or a box. Block is usually rectangular and hollow, but it is hard to be recognized by the machine learning methods directly because it is often too variant and is easily to be misidentified as image element, especially when the block containing images as shown in Figure 5.9(c).

However, the simple and general shape attributes (rectangular and hollow) can be well captured by the image processing methods. Therefore, a block recognition algorithm based on pure image processing is proposed here to identify the block region.

One detailed observation of block is that it can be likened to a wireframe, as demonstrated in Figure 5.9(b) and Figure 5.9(d). In other words, there should a gap between the border lines and the contents in it. Therefore, the algorithm identifying block is designed by detecting the gaps, and it only check the rectangular objects because of the shape attribute. The Python style pseudocode is shown in Algorithm 5.3.4. It just illustrates the basic idea of this algorithm, the real implementation would be more sophisticated because of the complicity of input images.

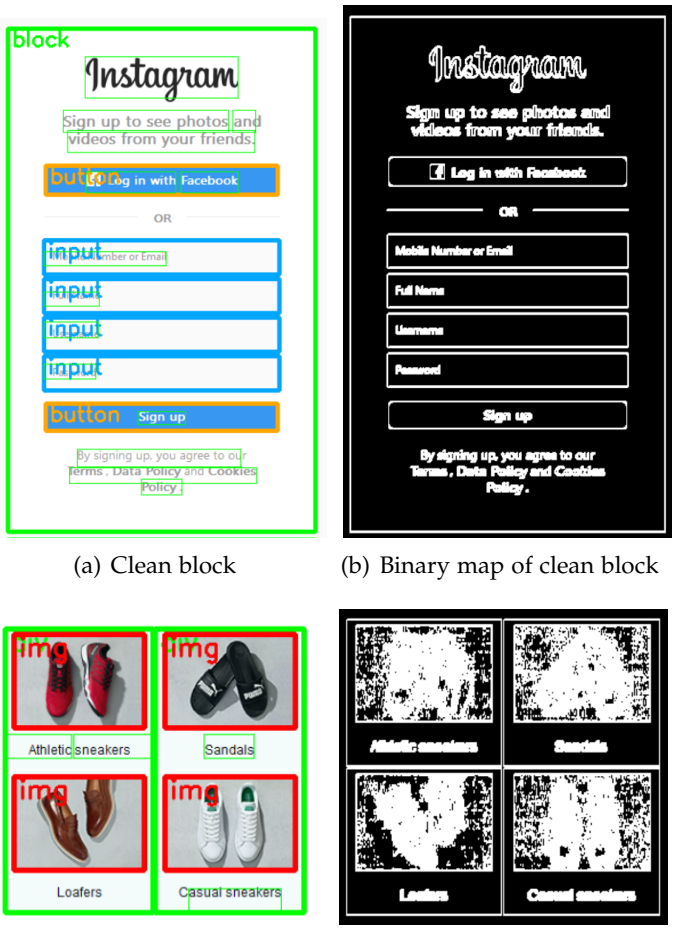
Algorithm 4 Block Recognition

Input: Boundaries of rectangular components; Binary map; Border thickness
Output: Boolean value to indicate if this component is block

$GapTop, GapBottom, GapLeft, GapRight \leftarrow True, True, True, True$
for i **in** $range(1..MaxBorderThickness)$ **do**
    if $\sum BinaryMap[BorderTop + i, BorderLeft + i : BorderRight - i] = 0$ **then**
        $GapTop \leftarrow False$
    end if
    if $\sum BinaryMap[BorderBottom - i, BorderLeft + i : BorderRight - i] = 0$ **then**
        $GapBottom \leftarrow False$
    end if
    if $\sum BinaryMap[BorderTop + i : BorderBottom - i, BorderLeft + i] = 0$ **then**
        $GapLeft \leftarrow False$
    end if
    if $\sum BinaryMap[BorderTop + i : BorderBottom - i, BorderRight - i] = 0$ **then**
        $GapRight \leftarrow False$
    end if
end for
if $GapTop = True$ **and** $GapBottom = True$ **and** $GapLeft = True$ **and** $GapRight = True$ **then**
    return $True$
else
    return $False$
end if

Where the $BorderTop$, $BorderBottom$, $BorderLeft$, $BorderRight$ are the boundary val-

ues of this component, $BorderTop, BorderBottom$ are the minimum row index and the maximum row index, and $BorderLeft, BorderRight$ are the minimum column index and maximum column index of it. The $\sum BinaryMap[BorderTop + i, BorderLeft + i : BorderRight - i]$ stands for summing up all pixels from column $BorderLeft + i$ to column $BorderRight - i$ in row $BorderTop + i$. If the amount is zero, column $BorderRight - i$ is hollow and is considered as gap.



(a) Clean block (b) Binary map of clean block

(c) Block containing image (d) Binary map of block containing image

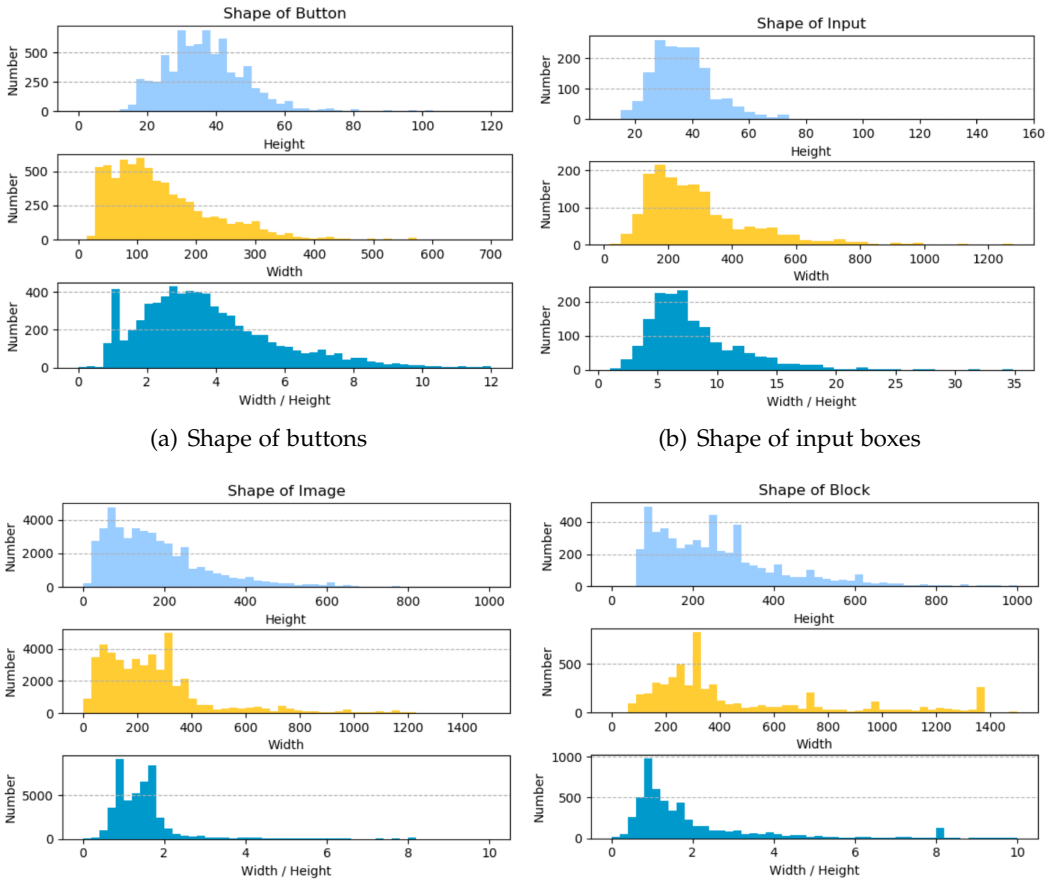
Figure 5.9: The demonstrations of a block. Blocks are drawn with green bounding box, and they usually are bordered regions where contain multiple components.

5.3.5 Irregular Shaped Components Selection

On the ground of the statistics 5.8, I observe the rule that human-computer interface components always have regular shapes (rectangle or round or oval), and the irregular objects are more likely to be image elements. However, exception still exists in functional UI components, although it is relatively rare. Thus, I add this step to

check all irregular objects and estimate whether they should be selected as potential UI components or be filtered out based on heuristics of size and aspect ratio of elements.

According to the datasets collected from web set and mobile applications whose statistics is show in Figure 5.10, some rules of the interface design in terms of the scale and length-width ratio are exposed and can be used as heuristic knowledge.



(a) Shape of buttons

(b) Shape of input boxes

(c) Shape of images

(d) Shape of blocks

Figure 5.10: Statistics of components' shape. For each type of UI components, three kinds of information are collected: height, width and aspect ratio (width / height). The amount of *Buttons*, *InputBoxes*, *Images*, *Blocks* are 10566, 3460, 39998, 1568 respectively from totally three different web and mobile application datasets.

From the above statistics, we can see the distribution of the shapes for different components. All of those distributions are long-tail and the majority of data concentrate in specific ranges. For example, most of buttons' height is in the range of 20 pixels to 60 pixels, and most of their width are between 40 pixels and 300 pixels, the major aspect ratio of buttons is on a scale of one to eight.

Therefore, adhoc filters is built according to the heuristics 5.1. Four rules are defined here and the filters are scattered over the whole process when implemented.

The *SmallComponent* rule leave out those small noises; the *AbnormalAspectRatio* rule is checked for all the components to filter out those impossible to be UI elements; for all irregular objects, the *IrregularImage* estimation is conducted to decide if we can directly affirm the irregular component is image; and the *Block* judgement should be pass for all blocks.

Number	Name	Heuristics
1	Small Component	$Area < 175 \bigwedge Perimeter < 70$
2	Abnormal Aspect Ratio	$width/height < 0.4 \bigvee width/height > 20$
3	Irregular Image	$isIrregular \bigwedge height > 70$
4	Block	$isBlock \bigwedge width > 70 \bigwedge height > 70$

Table 5.1: Heuristics based on the statistics.

5.3.6 Nested Components Detection

In the previous stages, we do not inspect into components to exam their contents, but some elements, such as button and input box, might be superimposed on an image. Therefore, the images detected are required to be further processed to check whether there are other interface components on them.

Observation about such nested components indicates that it is impossible to view a button inside another button or a input box nested in another input box. Also, technically, it is in no sense in putting a button inside another button, this behaviour is not allowed by the front-end language either. On the other hand, the common scenario is some interactive elements are put upon an image background as shown in Figure 5.11.

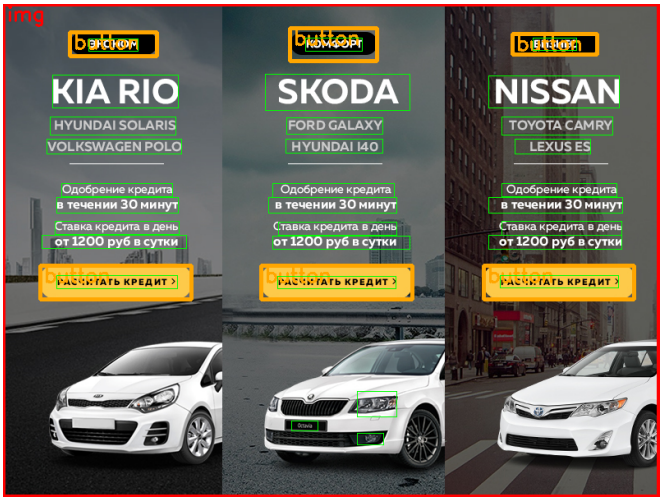


Figure 5.11: Example of figure that some interactive elements on a complicated image background

As elaborated in the pre-processing section 5.2, in order to overlook the detailed

texture and contents in the given image, I adopt a sensitive gradient threshold to try to incorporate all adjacent foreground points into individual connected components. But the drawback of this method is that the possible interface elements on an image are also integrated into this image component and be treated as a part of it.

To conquer this problem and separate the nested elements, a novel trick is proposed. I observe that the functional elements usually have solid monochromatic background, which makes the gradient of their background zero. So I reverse the binary map to generate an opposite image, which means all the background points whose pixel values are zero are assigned value 255, and vice versa, all white points are transformed into black.

(a) Binary map (b) Opposite map

Figure 5.12: The demonstration of the binary map and its opposite image.

Figure 5.12 shows the binary map that integrates some buttons into the background and identify them as an entire image component. But we can observe form the binary map 5.12(a) that a black hole in shape of a button appears surrounded by white foreground points. While this binary map is reversed, all the background became foreground, and the black holes turn to white components. By this means, the nested elements are extracted and transferred to the beginning of the UI components detection pipeline to conduct the same processes again.

As the aforementioned observation, this task cares more about the situations that interactive elements superimpose upon background images. Thus, when selecting the potential components inside images, I only leave those rectangular objects that are more likely to be interface elements rather than contents of the image.

5.4 Classification

After all the possible human-computer components are picked up in the previous steps, a classifier is built to identify their classes in order to generate the proper code at the very end of this system. Prior to any technical details, the categories and classes of interface elements in this task are defined as Table 5.2.

5.4.1 Categories and Classes of UI Components

The ultimate purpose of the UI components detection pipeline is to segment the potential interface elements from the input image and, based on their expected features, label them with human-computer interface tags, such as <img> or <button> in HTML, for code generation. Therefore, I define three categories of the UI components on the ground of the functional attributes those elements should have in real interface, and I define six classes according to the related objects and tags that perform particular functionalities in the front-end languages.

Category	Class Name
Interactive Elements	Button
	Input Box
Static Resource	Image
	Icon
	Text
Layout Structure	Block

Table 5.2: The categories and classes of UI components.

Interactive Elements: Those who are explicitly associated with some actions, such as page jumps and close the window, and can gather instructions from users are grouped in this category. For instance, the buttons are always related to some functions that will be performed after clicking, and the input boxes are the places where the user can feed their information into the application.

Static Resource: This category indicates that the elements in this group focus on displaying contents in stead of interaction with the user, just as the images and text on a webpage. Icon here is specifically useful for mobile application, it can be regarded as tiny image but it exists independently in Android development, so I separate this class from image. Although those resource can sometimes be implemented as functional elements, typically as hyperlinks, I still treat them as static resource at this stage for convenience. But at the code generation stage, I will give them the chance to expand their functions if need be.

Layout Structure: A special class distinguished from previous individual elements is Block, it is a layout structure that could contain multiple other kinds of elements. In other words, it presents a section of graphic interface consisting of various components. The meaning of this category is to find out some obvious hierarchies and layers for future code generation.

The Figure 5.13 demonstrate the visualized examples of the human-computer interface components' classes in the real application. This is also a glimpse of the final result of the UI components detection pipeline that the input image is semantically

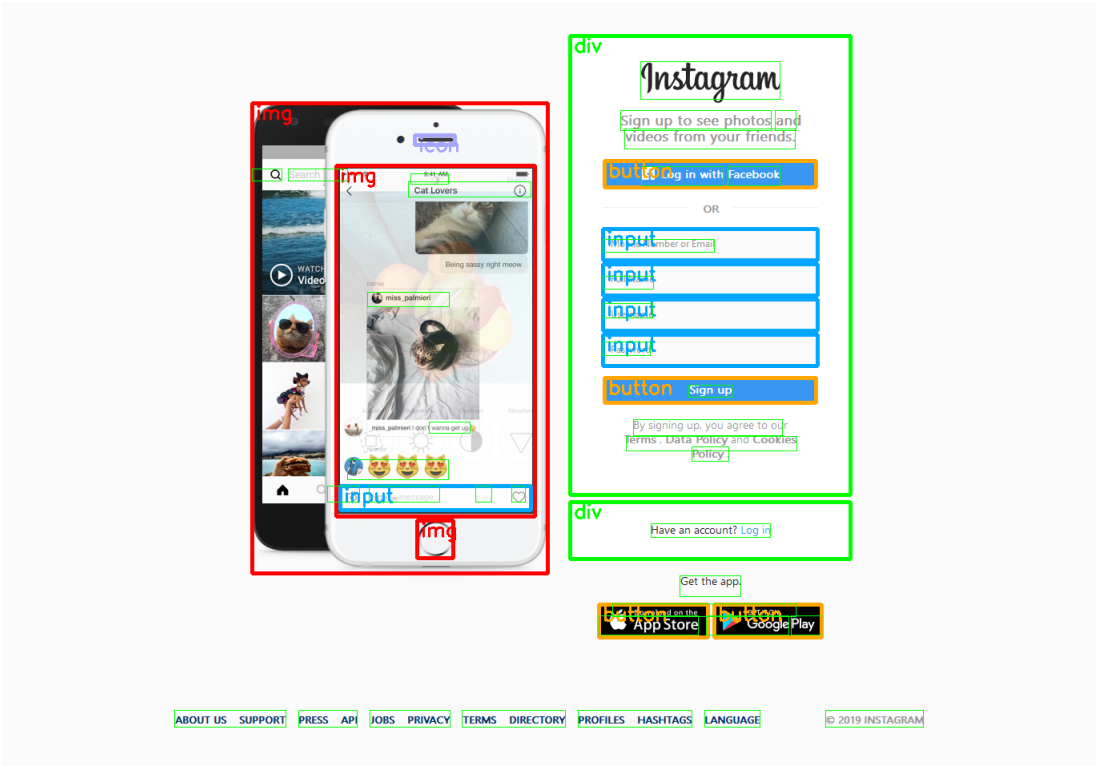


Figure 5.13: Demo of a labeled web page screenshot. Various classes are tagged with different colours of bounding box; the slim green boxes in this picture are the results from the CTPN showing the text recognition.

segmented into components with tags in which we know the locations and classes of those detected objects.

5.4.2 Classifier Model

To perform components classification, I recur to machine learning methods. Several techniques have been tried in the process including Super Vector Machine and Neural Network, and finally a simple four-layer convolutional neural network is adopted for its best performance.

Since the emphasis of this part is not on neural network, I only simply introduce the model's structure and evaluate the performance of various classifiers to sustain the choice of CNN.

There are three different methods implemented in this section, the Scale-invariant Feature Transform (SIFT) [Lowe, 2004, 1999], the Histogram of Oriented Gradients (HOG) [McConnell, 1986; Freeman and Roth, 1994] combined with Support Vector Machine (SVM) [Cortes and Vapnik, 1995] and the Convolutional Neural Network (CNN). Again, this part does not elaborate the technical details of those techniques, instead, it just introduces the basic theories of them and state the experimental re-

sults.

5.4.2.1 HOG + SVM

Another popular feature extraction technique is the Histogram of Oriented Gradients (HOG), it is usually be utilized as a feature descriptor in computer vision and image processing [Freeman and Roth, 1994]. This pipeline has some unique advantages compared to others. First, HOG processes image on a dense grid of uniformly spaced cells, which endows it the favourable ability to keep the optical and geometric invariant of the image. Besides, this algorithm is more robust and can be less sensitive for small changes [Dalal and Triggs, 2005].

This pipeline consists of five steps: gradient computation, orientation binning partition, descriptor blocks generation, block normalization and SVM object recognition.

Gradient computation: The first pre-processing step is similar to the UI components detection pipeline, the gradient of this image is calculated at the beginning. Basically, the way to compute the gradient is same as formulae 5.3, but the author implement it by the following filter kernel:

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{5.4}$$

$$g(x,y) = w * f(x,y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(x,t) f(x-s, y-t) \tag{5.5}$$

The kernel 5.4 is convoluted on image to compute the gradinet, the general expression of a convolution is stated in 5.5, where the $g(x,y)$ is the filtered image, $f(x,y)$ is the original input image and w is the kernel. Every element of the kernel is in range where $-a \leq s \leq a$ and $-b \leq t \leq b$.

Orientation binning: This step create cell histograms for the gradient result. The cells can be either rectangular of radial. In the radial shape, for example, the degree of each bin depend on the total number of channels, the degree of a bin in a night-channel histogram is $360/9 = 40^\circ$. The gradients cast a weighted vote for those bins, for instance, if a pixel's gradient direction is 12°, then the first bin (range from 0° to 40°) increases by x, where x is the magnitude of the gradient.

Descriptor blocks generation: The gradients vary greatly because of the illumination and contrast, so the strength of gradients should be normalized to acquire accurate result. To this end, the cells are grouped into larger spatially connected blocks, and those blocks are concatenated to a descriptor.

Block normalization: As mentioned in the previous paragraph, HOG conduct the

normalization to mitigate the huge variation among gradients' lengths. I adopted the L2-norm in my implementation, the expression of it is shown below:

$$f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}} \tag{5.6}$$

where: v is the vector, $\|v\|_k$ is the k-norm of v which can be 1 or 2 to show L1-norm or L2-norm, and e is a small constant.

Object recognition: The previous steps generate a normalized feature vector of an image or object, then this vector can be feed into some classifiers to recognize its class. A widely used technique to combine with the HOG is the Supper Vector Machine (SVM) [Dalal and Triggs, 2005], a supervised learning models based on learning algorithm to perform classification and regression [Patel, 2017].

The basic idea of SVM is that we try to separate p-dimensional vectors with $(p-1)$-dimensional hyperplanes. There can be multiple hyperplanes that might partition the data. The one representing the largest separation or margin between the two classes is selected as the maximum-margin hyperplane. The distance from it to the nearest data point on each class is maximized [Asa Ben-Hur, 2001].

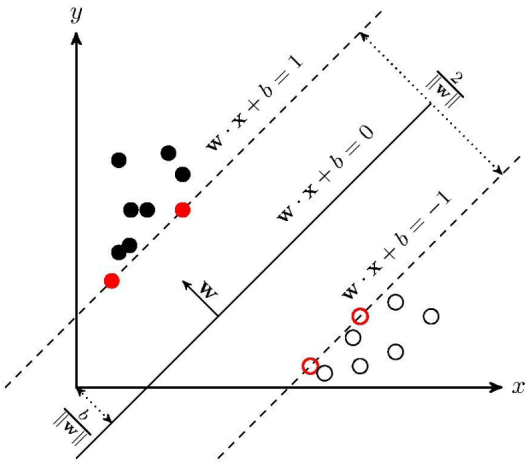


Figure 5.14: Hyperline partitioning two groups of points

Figure 5.14 [Yu, 2019] presents the demonstration of the linear SVM, where w is the normal vector and b is a constant parameter; the parameter $\frac{2}{\|\vec{w}\|}$ defines the offset of the hyperplane from the origin along the normal vector $\vec{w}$. All possible hyperplanes should satisfy the expression $\vec{w} \cdot \vec{x} + b = 0$ for the set of points $\vec{x}$. In this case, the $\vec{x}$ is the set of HOG feature vectors computed in the preceding steps.

5.4.2.2 SIFT

The Scale-invariant Feature Transform is a classic feature detection method widely used in computer vision tasks to describe the local features in images. It is adopted

in many applications such as objection recognition and action recognition [Lowe, 1999], the robust performance in recognition tasks is the motivation that I try this technique.

As its name suggests, the core concern this algorithm addresses is to keep the features of objects invariant in various scales [Mikolajczyk and Schmid, 2005]. It search for the extreme points in the space and extracts their location, scale and orientation, and then all those features are stored in database through Hash Table. An object from a new image is recognized by comparing and matching its features to the database to find the candidates based on Eucidean distance of their feature vectors. Mainly four steps involved in SIFT algorithm, this section is just a brief summary of it [Mordvintsev and Revision, 2013].

Scale-space extrema detection: This step extracts a large set of feature vectors from the input image. Those vectors are theoretically invariant to image scaling and rotation and robust to local geometric distortion. To this end, the author proposed the scale-space extrema detection by using Gaussian filter in various σ scales, where the σ decides the smoothness of an image. Small σ reflects the details of the image while large σ presents the blurred picture.

SIFT uses Difference of Gaussians (DoG) to calculate the d Gaussian Pyramid. DoG are obtained as the difference between two images blurred by different Gaussian filters with different σ, and this process is done for all layers in Gaussian Pyramid as illustrated below.

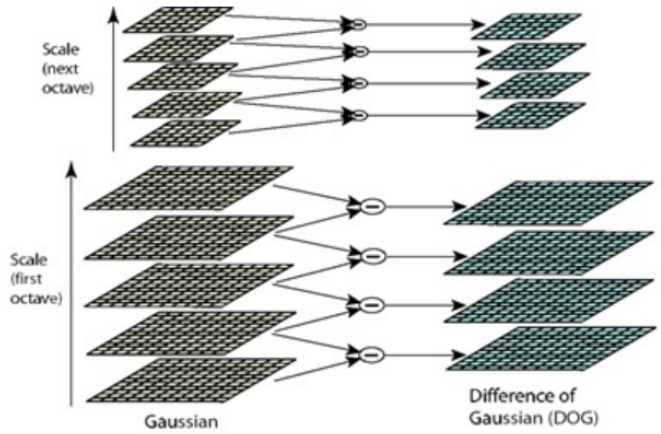


Figure 5.15: DoG are computed in all layers in Gaussian Pyramid

After the DoG are calculated, each pixel in an image is compared with its eight neighbours as well as corresponding nine pixels in the next scale and the previous scale to find the local extrema that can be a potential key point representing the feature in its scale.

Keypoint localization: For all potential keypoints, SIFT uses Taylor series expansion of scale space to get more accurate location, but if the intensity at a point is less

than a specific threshold, it is rejected.

Orientation assignment: The keypoints are assigned orientation in this step by calculating the gradients with its neighbours on this scale. The gradients' direction and magnitude is stored in an orientation histogram with thirty six bins. This process is helpful to keep invariance to image rotation.

Keypoint descriptor: SIFT store the keypoints in a descriptor which consists of multiple bins of orientation histogram. In detail, a 16x16 neighbourhood around each keypoint is taken and divided into 16 4x4 sub-blocks. Then an eight-bin histogram is created for each sub-block to collect the gradients. So a total of 128 bins are created and vectorized to store the information of this keypoint.

Eventually, the features of an image are extracted and can be used to recognize others in the same category either by matching the Euclidean distance or by utilizing this vector as an encoder and feed into machine learning techniques such as SVM.

5.4.2.3 CNN

Compared with aforementioned techniques, the Convolutional Neural Network (CNN) is a more end-to-end method [Jeeva, 2018]. Because of the popularity of CNN, I will not iterate its basic mechanism and principle in this thesis. The emphasis here is on the model's structure and its effectiveness.

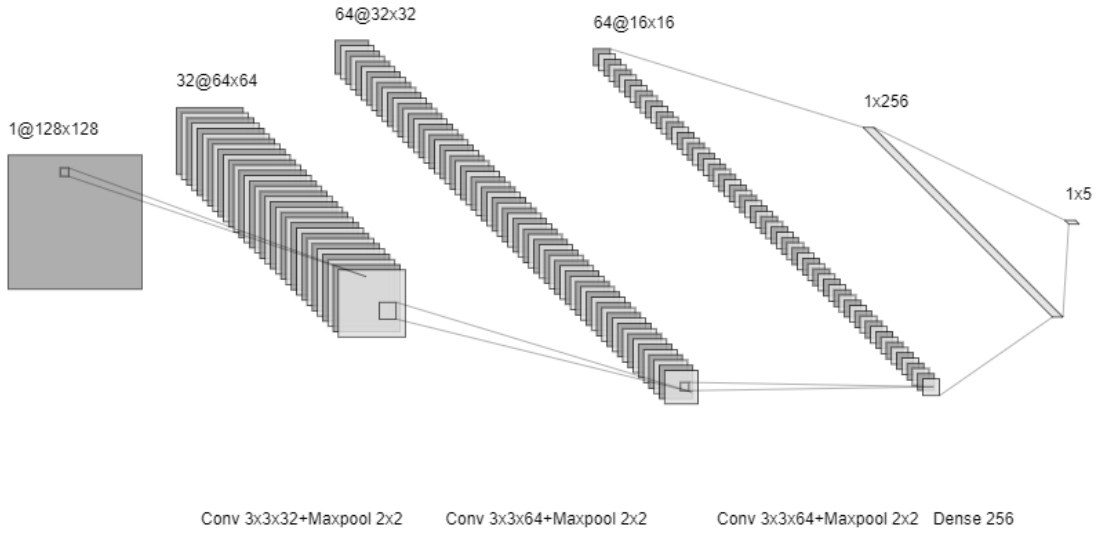


Figure 5.16: The structure of the four-layer network. A 3x3 sliding window is adopted to move through the original image that is resized into size of 128x128. All convolutional layers are followed by a 2x2 max-pooling layer. The output layer is a five-class softmax to classify the input into *image*, *button*, *input box*, *icon* and *text*.

As the classic use case, this technique is utilized to recognize the class of image in

the UI components detection pipeline. The input now is the individual component. An observation on the datasets of human-computer interface components is that the buttons and input boxes are less variant. In other words, components in those two classes are always in a similar pattern. For example, the buttons are always in the form that one or few words locate in the center of a bordered region. On the contrary, the image elements are always variegated and colourful. So the difference among various classes can be rather obvious and a shallow neural network might be sufficient in this case.

Therefore, I build a four-layer model to handle this task, and the srtucture is presented in Figure 5.16. In order to offset the effect caused by the insufficient parameters in a shallow network, each layer is created more broad.

5.4.3 Performance

As mentioned above, I implement and compare those three techniques to select the best one as the classifier. The experiment is conducted on the components dataset consisting of image of web and mobile application. Although there are some slight variations among interface elements of web and mobile apps, those nuances does not impose considerable influence on the model's performance. Therefore, all components in the same class from different sources (web and mobile application) are collected together and fed into the training models. The table 5.3 shows the experimental performance of those three classifiers.

Model	Accuracy	Recall	Balanced Accuracy
SIFT + SVM	0.9166	0.9061	0.9007
HOG + SVM	0.9326	0.9112	0.9065
CNN	0.9566	0.9128	0.9226

Table 5.3: The performance of models. The balanced accuracy is taken into account as a criterion because of the imbalanced datasets where image components are far more than others. The experiments presents that the CNN model is relative better than the other two in all aspects.

5.5 Text Processing

The text processing is achieved by a popular text detection method, the Connectionist Text Proposal Network (CTPN) [Tian et al., 2016]. The CTPN was originally designed for accurately localizing the text lines in nature scene, leveraging the vertical anchor regression and connectist proposals to accurately detect text lines. This technique achieves a high performance on the ICDAR 2013 and 2015 benchmarks at a fast processing speed (0.14s/image).

5.5.1 Introduction

The CTPN refers to the state-of-the-art object detection methods, especially the Faster RCNN [Ren et al., 2015], but the authors proposed some novel improvement to adapt the network to the natural scene text line detection based on the visual properties of sentence. In this process, several following contributions are made:

First, the authors transformed the OCR problem into localizing a sequence of fine-scale text proposals, which could apply some mechanisms of object detection means [Girshick, 2015; Ren et al., 2015]. For instance, anchor regression that widely used in recent object detection methods is utilized to acquire the position of the targets. But CTPN takes the morphological characters of text line into consideration and deploys vertical anchors to produce the region proposals, which makes a significant progress to the localization accuracy. This novelty departs from the Region Proposal Network in the Faster RCNN, which attempts to predict a whole object, hence is difficult to provide a satisfied localization accuracy in the case of text detection.

Second, in view of the consecutiveness of sentence, the authors incorporated an in-network recurrence mechanism that connects sequential proposals in the convolutional feature maps into the model. Thereby, the network is capable of exploring the meaningful context information.

Third, this method is scalable and robust in various environments. An end-to-end trainable network is produced, able to handle multi-scale and multi-lingual text in the same image without and revision and filtering. This robustness is achieved by the diversity of the training dataset as well as the anchor regression mechanism.

Attribute to the aforementioned novelties, the CTPN refreshed a series of benchmarks, significantly improving results of preceding methods. (e.g., 0.88 F-measure over 0.83 in [Gupta et al., 2016] on the ICDAR 2013, and 0.61 F-measure over 0.54 in [Zhang et al., 2016] on the ICDAR2015).

5.5.2 Technical Details

Three critical novelties are proposed in the CTPN: detecting in fine-scale proposals, recurrent connectionist text proposals and side-refinement. Figure 5.17 shows the architecture of the CTPN.

Fine-scale proposals: This technique adopts the fully convolutional network to allows an input image of arbitrary size. Referring to FRCNN, it leverages slide windows to detect the text areas in the the convolutional feature maps, then generates a series of fine-scale text proposals.

The sliding-windows methods used to facilitate the region proposal. Most classical sliding-windows approaches define several fixed-size anchors to detect objects of similar size. The CTPN extends this efficient mechanism by means of revising the scale and aspect ratio of anchors to fit the properties of text. One peculiarity of text line is that it is a sequential region where it does not have an obvious closed boundary. Multi-level components, such as stroke, character, word, text line and paragraph are involved in the process. The text detection in this case, however, focuses on the

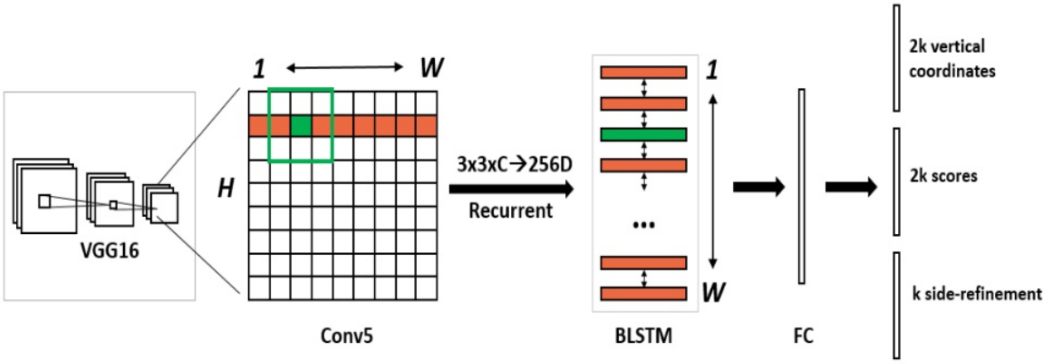


Figure 5.17: The overall structure of the Connectionist Text Proposal Network (CTPN). A 3x3 sliding window is applied through the last convulotional map of the base network (VGG16). Then the sequence of windows in each row are recurrently connected by a Bio-directional LSTM to gather the sequential context information. At the end, the RNN layer is connected to a 512D fully-connected layer and the output layer where the text/non-text score and y-coordinate are predicted, and the k anchors are offset by the side-refinement.

text line and region level, so the accuracy might not be satisfied if it still treat the targets as single objects.

Therefore, the authors proposed a vertical mechanism that predicts a text/non-text score and y-axis location of each proposal. The experiments prove that detecting text line in a sequence of fix-width proposal is more effective and accurate than recognizing individual characters. Moreover, the fixed-width proposals also work well text of various scales and aspect ratios.

In detail, the CTPN utilizes k vertical anchors in fixed width of 16 pixels. Those k anchors have same horizontal location, but they vary in k different heights vertically. Thus, each predicted proposal has a bounding box with size of h x 16, where the h is the predicted height. The network's output are the text/non-text scores and the predicted y-coordinates that represents the height of this anchor for k anchors of each window.

Recurrent Connectioist Text Proposals: The fine-scale proposals results from splitting the text line into a sequence of slender regions, and those proposals are predicted independently. However, it is obviously not robust to process each part of text regions separately. Therefore, the authors leveraged the recurrent mechanism to make use of the sequential nature of text. Furthermore, the CTPN integrates the context information into the convolutional layer.

To this end, the CTPN, the long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] is used for the recurrent neural network (RNN) layer. The authors adopted some tricks to address vanishing gradient problem. Moreover, they use a bi-directional LSTM [Graves and Schmidhuber, 2005] to extend the RNN layer to collect the context information in both directions.

Side-Refinement: After acquiring the fine-scale proposals, the CTPN straightforward connects those whose text/non-text score is greater than 0.7 to construct the text line. The text regions are now divided by a sequence of equal 16-pixel width proposals of different heights.Such variations can lead to inaccuracy of text line detection. To address this problem, the authors proposed a side-refinement approach to estimate the offset for each proposal horizontally.

5.6 Merge

The UI components detection and text detection are done independently by the two branches. In the end, the pipeline cross checks the correctness and integrates those results.

One drawback of the image processing based method is that the noises would be selected by mistake. The pre-processing stage attempts to incorporate all the variegated contents into individual components, and plenty of objects that are unlikely to be interface components are filtered out based on their sizes and aspect ratios according to the heuristics in table 5.1. But there are still some isolated regions where are wrongly selected as component candidates. The observation on those false positives is that they always come from two source, the small isolated parts of the image components and the text regions.

As stated at the beginning of this chapter, the detection of interface components and recognition of text regions are separate and performed in two specified pipelines. Those two branches are expected to focus on their won tasks for the sake of accuracy. Therefore, the text regions are not desired in the UI components detection pipeline and should be filtered out as noises. To this end, the CNN in the this branch is also trained to be able to recognize the text to avoid mixing it up with other interface elements.

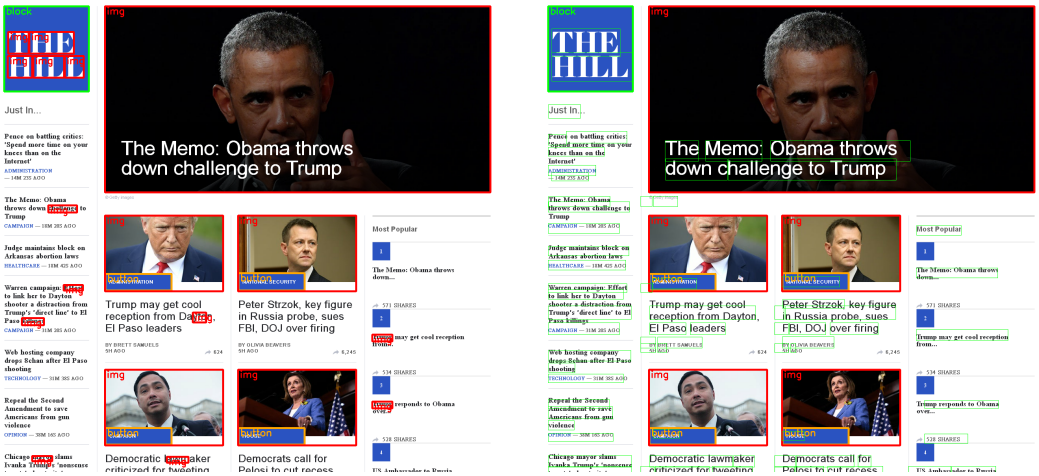
But the real text regions still have chance to be improperly recognized as buttons or images because of the false prediction of the CNN. Thus the system applies the results of CTPN to double check the results of components detection and discard those misidentified human-interface elements. The process is visualized in Figure 5.18.

To achieve this, the resulting UI components on the left hand side are filtered by computing the overlap with the text lines. The overlap computation is on the ground of their *Intersection over Area* (IoA), as $IoA(a,b) = \frac{Area_a \cap Area_b}{Area_a}$. A valid UI component i should satisfy the two requirements:

$$\forall j \in Text(IoA(i,j) < 0.7) \tag{5.7}$$

$$(\sum_{j \in Text} Inter(j,i))/Area_i < 0.85 \tag{5.8}$$

The requirement 5.7 means the intersection area of a single text region with the component i should smaller than the 70 percentage of component i's area; and the



(a) Result of the UI components detection

(b) Result after merging with text detection

Figure 5.18: A section of web application interface. The 5.18(a) demonstrates the detecting result of the UI components detection pipeline, in which several text regions are wrongly recognized as image elements (marked with red bounding boxes). The merged result is shown in the 5.18(b), the green slim lines in this figure are the text areas detected by the CTPN. After double-checking by the CTPN, those false positive image elements that are actually text are discarded.

expression 5.8 stipulates that the total text area of an interface element should less than 85 percentage of the its area.

Eventually, the merged result is presented to user as the visualized output of this whole detection system, and it is also transferred to the code generation pipeline to produce the corresponding front-end code.

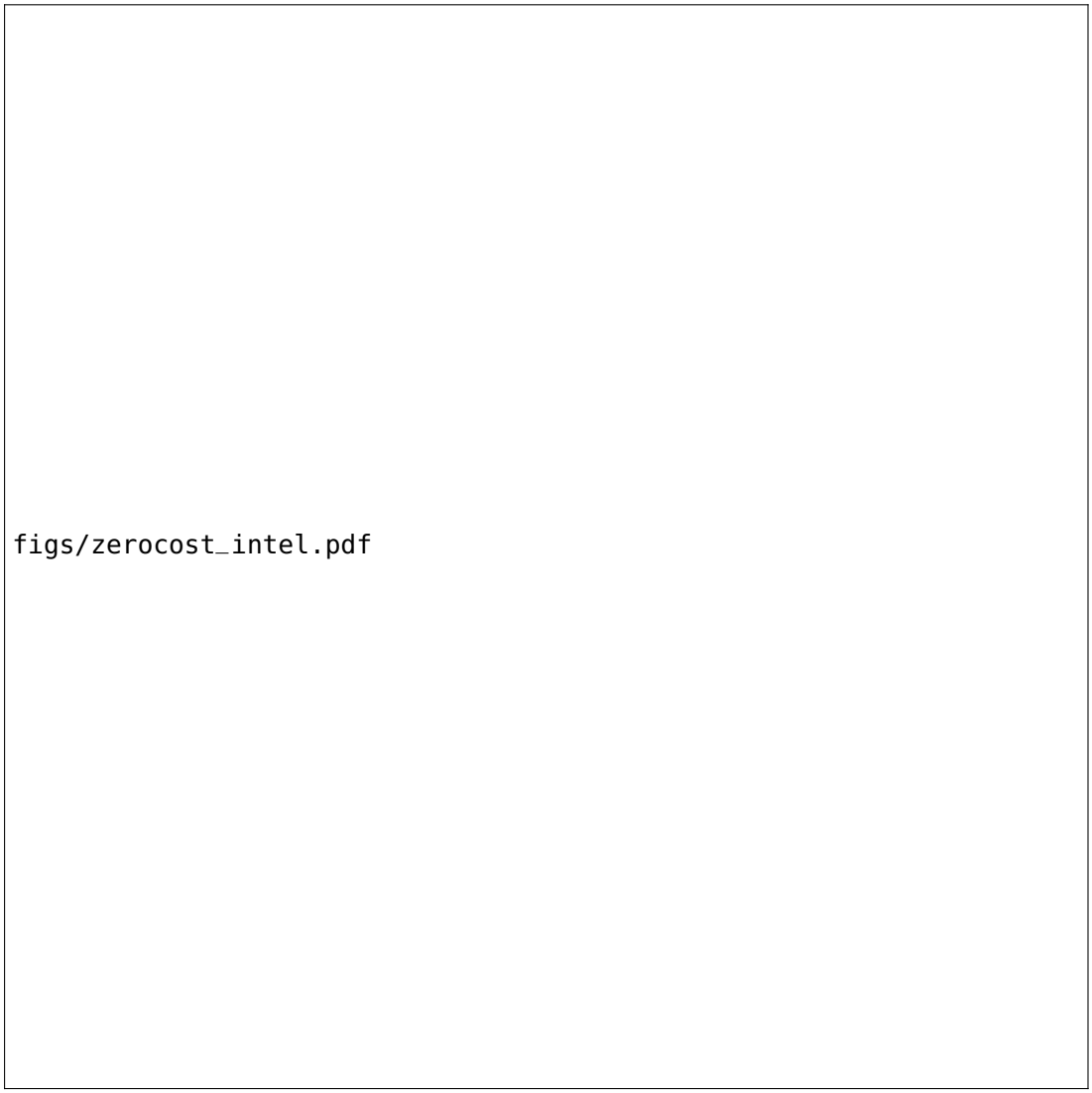
Chapter 6

Results

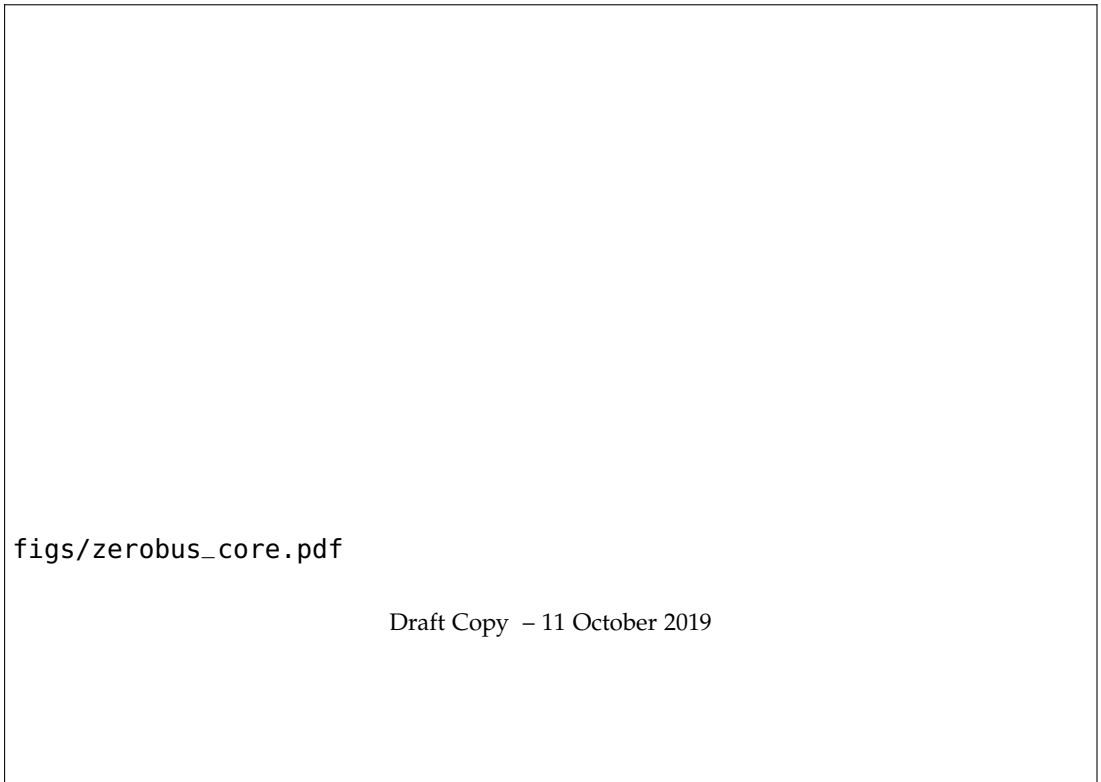
6.1 Direct Cost

Here is the example to show how to include a figure. Figure 6.1 includes two subfigures (Figure 6.1(a), and Figure 6.1(b));

6.2 Summary



(a) Fraction of cycles spent on zeroing



Chapter 7

Conclusion

Summary your thesis and discuss what you are going to do in the future in Section 7.1.

7.1 Future Work

Good luck.

Bibliography

Asa Ben-Hur, H. T. S. V. V., David Horn, 2001. Support vector clustering. *Journal of Machine Learning Research*, (2001), 125–137. http://www.jmlr.org/papers/v2/horn01a.html. (cited on page 37)

Canny, J., 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8, 6 (Nov 1986), 679–698. doi:10.1109/TPAMI.1986.4767851. (cited on page 19)

Canny, J., 1987. A computational approach to edge detection. *Readings in Computer Vision*, (1987), 184–203. doi:10.1016/b978-0-08-051581-6.50024-6. (cited on page 9)

Cepelewicz, J., 2019. Where we see shapes, ai sees textures. https://www.quantamagazine.org/where-we-see-shapes-ai-sees-textures-20190701. (cited on page 17)

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20, 3 (Sep 1995), 273–297. doi:10.1007/BF00994018. https://doi.org/10.1007/BF00994018. (cited on page 35)

Dalal, N. and Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. In *International Conference on Computer Vision & Pattern Recognition (CVPR '05)*, vol. 1, 886–893. IEEE Computer Society, San Diego, United States. doi:10.1109/CVPR.2005.177. https://hal.inria.fr/inria-00548512. (cited on pages 36 and 37)

Dillencourt, M. B.; Samet, H.; and Tamminen, M., 1992. A general approach to connected-component labeling for arbitrary image representations. *J. ACM*, 39, 2 (Apr. 1992), 253–280. doi:10.1145/128749.128750. http://doi.acm.org/10.1145/128749.128750. (cited on page 24)

Douglas, D. and Peucker, T., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, (1973), 112–122. doi:https://utpjournals.press/doi/10.3138/FM57-6770-U75U-7727. (cited on page 28)

Duda, R. O. and Hart, P. E., 1972. Use of the hough transformation to detect lines and curves in pictures. *Comm. ACM, Vol. 15,*, (01 1972), 11–15. (cited on page 28)

EGGLESTON, P., 2015. Understanding oversegmentation and region merging. https://www.vision-systems.com/non-factory/security-surveillance-transportation/

article/16739494/understanding-oversegmentation-and-region-merging. (cited on page 11)

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D., 2010. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 9 (Sep. 2010), 1627–1645. doi:10.1109/TPAMI.2009.167. (cited on page 8)

Freedman, D., 2012. *Statistical models: theory and practice*. Cambridge University Press. (cited on page 6)

Freeman, W. T. and Roth, M., 1994. Orientation histograms for hand gesture recognition. Technical Report TR94-03, MERL - Mitsubishi Electric Research Laboratories, Cambridge, MA 02139. https://www.merl.com/publications/TR94-03/. (cited on pages 35 and 36)

Gandhi, R., 2018. R-cnn, fast r-cnn, faster r-cnn, yolo - object detection algorithms. https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e. (cited on pages 5 and 8)

Geirhos, R., 2018. Out of shape? why deep learning works differently. https://blog.usejournal.com/why-deep-learning-works-differently-than-we-thought-ec28823bdbc. (cited on page 17)

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W., 2019. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bygh9j09KX. (cited on page 17)

Girshick, R., 2015. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*. (cited on pages 5 and 41)

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J., 2014a. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on page 5)

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J., 2014b. Rich feature hierarchies for accurate object detection and semantic segmentation. (cited on page 6)

Graves, A. and Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18, 5 (2005), 602 – 610. doi:https://doi.org/10.1016/j.neunet.2005.06.042. http://www.sciencedirect.com/science/article/pii/S0893608005001206. IJCNN 2005. (cited on page 42)

Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; and Chen, T., 2018. Recent advances in convolutional neural networks. *Pattern Recognition*, 77 (2018), 354 – 377. doi:https://doi.org/10.1016/j.patcog.2017.10.013. http://www.sciencedirect.com/science/article/pii/S0031320317304120. (cited on page 8)

Gupta, A.; Vedaldi, A.; and Zisserman, A., 2016. Synthetic data for text localisation in natural images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on page 41)

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B., 2017. Mask R-CNN. *CoRR*, abs/1703.06870 (2017). http://arxiv.org/abs/1703.06870. (cited on page 5)

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 9, 8 (1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735. https://doi.org/10.1162/neco.1997.9.8.1735. (cited on page 42)

Huizinga, D. and Kolawa, A., 2007. Automated defect prevention best practices in software management. https://www.amazon.com/Automated-Defect-Prevention-Practices-Management/dp/0470042125. (cited on page 14)

Jacobs, D., 2005. Image gradients. *Class Notes for CMSC*, (9 2005). http://www.cs.umd.edu/~djacobs/CMSC426/ImageGradients.pdf. (cited on page 20)

Jeeva, M., 2018. The scuffle between two algorithms -neural network vs. support vector machine. https://medium.com/analytics-vidhya/the-scuffle-between-two-algorithms-neural-network-vs-support-vector-machine-16abe0eb4181. (cited on page 39)

Kobayashi, M. and Takeda, K., 2000. Information retrieval on the web. *ACM Comput. Surv.*, 32, 2 (Jun. 2000), 144–173. doi:10.1145/358923.358934. http://doi.acm.org/10.1145/358923.358934. (cited on pages 13 and 15)

Lee, S.; Kwak, S.; and Cho, M., 2019. Universal bounding box regression and its applications. *CoRR*, abs/1904.06805 (2019). http://arxiv.org/abs/1904.06805. (cited on page 6)

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C., 2016. Ssd: Single shot multibox detector. In *Computer Vision – ECCV 2016*, 21–37. Springer International Publishing, Cham. (cited on page 5)

Lowe, D. G., 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1150–1157 vol.2. doi:10.1109/ICCV.1999.790410. (cited on pages 35 and 38)

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2 (Nov 2004), 91–110. doi:10.1023/B:VISI.

0000029664.99615.94. https://doi.org/10.1023/B:VISI.0000029664.99615.94. (cited on page 35)

McConnell, R. K., 1986. Method of and apparatus for pattern recognition. https://patents.google.com/patent/US4567610. (cited on page 35)

Mehmet Sezgin, B. S., 2004. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, , 13 (2004), 146 – 165. doi:https://doi.org/10.1117/1.1631315. (cited on page 21)

Mikolajczyk, K. and Schmid, C., 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 10 (Oct 2005), 1615–1630. doi:10.1109/TPAMI.2005.188. (cited on page 38)

Moizuddin, K., 2019. Components of the selenium automation tool - dzone devops. https://dzone.com/articles/components-of-selenium-automation-tool. (cited on pages 14 and 15)

Mordvintsev, A. and Revision, A. K., 2013. Introduction to sift (scale-invariant feature transform)¶. https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_sift_intro/py_sift_intro.html. (cited on page 38)

Patel, S., 2017. Chapter 2 : Svm (support vector machine) - theory. https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72. (cited on page 37)

Prewit, J., 1970. Object enhancement and extraction. In *Picture Processing and Psychopictorics*, 75–120. B.S. Lipkin. (cited on page 20)

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on pages 5 and 17)

Redmon, J. and Farhadi, A., 2018. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767 (2018). http://arxiv.org/abs/1804.02767. (cited on page 17)

Ren, S.; He, K.; Girshick, R.; and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28* (Eds. C. Cortes; N. D. Lawrence; D. D. Lee; M. Sugiyama; and R. Garnett), 91–99. Curran Associates, Inc. http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf. (cited on pages 5, 9, 17, and 41)

Richardson, L., 2015. Beautiful soup documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/. (cited on page 14)

Roberts, L., 1963. Machine perception of three-dimensional solids. (1963). https://dspace.mit.edu/bitstream/handle/1721.1/11589/33959125-MIT.pdf. (cited on page 20)

SACHAN, A., 2017. Zero to hero: Guide to object detection using deep learning: Faster r-cnn,yolo,ssd. https://cv-tricks.com/object-detection/faster-r-cnn-yolo-ssd/. (cited on page 9)

Samet, H. and Tamminen, M., 1988. Efficient component labeling of images of arbitrary dimension represented by linear bintrees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10, 4 (July 1988), 579–586. doi:10.1109/34.3918. (cited on page 24)

SEAL, H. L., 1967. Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. *Biometrika*, 54, 1-2 (06 1967), 1–24. doi:10.1093/biomet/54.1-2.1. https://doi.org/10.1093/biomet/54.1-2.1. (cited on page 6)

Shapiro, S. G. C., Linda G., 2002. In *Computer Vision*. B.S. Lipkin. (cited on page 21)

Sobel, I., 1968. An isotropic 3x3 image gradient operator. *Presentation at Stanford A.I. Project 1968*, (02 1968). (cited on page 20)

Suzuki, S. and be, K., 1985. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30, 1 (1985), 32 – 46. doi:https://doi.org/10.1016/0734-189X(85)90016-7. http://www.sciencedirect.com/science/article/pii/0734189X85900167. (cited on page 20)

Team, O. D. Structural analysis and shape descriptors. https://docs.opencv.org/2.4/modules/imgproc/doc/structural_analysis_and_shape_descriptors.html#id5. (cited on page 28)

Team, O. D., 2012. Structural analysis and shape descriptors. https://docs.opencv.org/2.4/modules/imgproc/doc/structural_analysis_and_shape_descriptors.html#structural-analysis-and-shape-descriptors. (cited on pages 20 and 26)

Tian, Z.; Huang, W.; He, T.; He, P.; and Qiao, Y., 2016. Detecting text in natural image with connectionist text proposal network. In *Computer Vision – ECCV 2016*, 56–72. Springer International Publishing, Cham. (cited on pages 3, 17, and 40)

Uijlings, J. R. R.; van de Sande, K. E. A.; Gevers, T.; and Smeulders, A. W. M., 2013. Selective search for object recognition. *International Journal of Computer Vision*, 104, 2 (Sep 2013), 154–171. doi:10.1007/s11263-013-0620-5. https://doi.org/10.1007/s11263-013-0620-5. (cited on page 8)

Vincent, L. and Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, 6 (June 1991), 583–598. doi:10.1109/34.87344. (cited on page 24)

Whitmer, R., 2009. Document object model (dom). https://www.w3.org/DOM/#what. (cited on page 13)

Yu, S., 2019. Svm - theory. https://zhuanlan.zhihu.com/p/31886934. (cited on page 37)

Zhang, Y.; Chen, Y.; Huang, C.; and Gao, M., 2019. Object detection network based on feature fusion and attention mechanism. *Future Internet*, 11, 1 (Feb 2019), 9. doi:10.3390/fi11010009. (cited on page 9)

Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; and Bai, X., 2016. Multi-oriented text detection with fully convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on page 41)