International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
International Conference & workshop on Advanced Computing 2014 (ICWAC 2014) – www.ijais.org

Generation of Web Pages from Document Image

Aparna Halbe
Department of Computer
DJSCOE
Mumbai, India

Abhijit R. Joshi, Ph.D
Head of the Department,
Information Technology
DJSCOE
Mumbai, India

ABSTRACT
The development of any project in software industry begins with Requirement specification followed by User Interface [UI] design. Normally UI design is drawn on paper first. Web designers then design the web pages as per the design on the paper. Various Mark Up languages such as HTML/XML are used to design and publish web pages on the internet.

In this paper a novel approach is proposed that will do the job of web designer. This system will convert the UI design drawn on paper to HTML page. A scanned image of UI design will be provided as an input to the system and it generates the output which will be a HTML page of that UI. To do this, system requires the conversion of paper document image into hyper documents. Currently, the work done in this area is restricted only to the conversion of images and text into hyper document. Here, an idea of converting document image of UI design into actual HTML page, is proposed.

Also work done so far in this area is restricted only to the text and images on documents. It does not consider various HTML controls like textbox, radio button, checkboxes, button etc. Therefore, existing system just converts the paper document into hypertext document and does not identify each HTML control as a separate component, which is (a primary requirement) required while designing UI. Given a UI design with different HTML controls, the existing system would just convert it to HTML page without providing any functionality. The generated HTML page will have an image of the UI design rather than actual HTML controls. The proposed work is addressing all these issues and will be considering most of the HTML controls those are required for designing static pages.

Keywords
Image Processing for GUI, rapid web development, GUI design, processing document images, automatic web page generation

1. INTRODUCTION
The purpose of user interface design is to enable people to interact with application. If people can't figure out how the application works or where to go on the website they will get confused and frustrated. To design a good User interface the designer requires a better understanding of user needs. There are several phases and processes in the user interface design, some of which are more dependent upon the end user. In common practice, designers use pen and paper to draw UI design and then start developing application on the computer as per the design drawn on the paper. Nearly all software applications have a graphical user interface, through which user carry out interactions. It means that the program code includes graphical controls, which the user can select using a mouse or keyboard for interactions. Commonly used graphical controls in the designing of UI are button, textbox, checkbox, radio button etc. A web page is a document on the World Wide Web (WWW), consisting of Hyper Text Markup Language (HTML) controls and any related scripts and graphics, and often hyperlinked to other documents on the Web. Today with the help of web pages we can easily share the data across the world.

Generation of web pages from GUI design involves following basic steps.

1. Draw UI design on the paper.
2. User approval for UI design drawn on the paper.
3. Web developer generates HTML pages as per UI design drawn on the paper.
4. User approval for UI/ HTML pages developed by web developer.

2. MOTIVATION
Considering rapid growth of internet, rapid development of WebPages is necessary. To develop WebPages, web designers generally use available tools. In order to use these tools, one needs the knowledge of HTML. In some organizations generally, sign off from client is mandatory before proceeding to the next stage of the project. Communication with client usually happens via project manager. If at all any changes are suggested by the client in the UI design then project manager has to communicate those changes to the web developer. This process requires more time.

Many a times what looks good on paper may not look good on a screen. In that case UI design has to be redrawn on the paper.

Work done so far in this area is restricted only to the text on the document image. It does not take care of various HTML controls. The proposed software can be used by any non-technical person to generate the web pages automatically from a UI design document. This will save developer's time and will help in rapid web development.

3. MAJOR CHALLENGES
UI designs are extensively used in software industry as a first step towards designing web pages. UI designs are simple and intuitive way for expressing look and feel of web pages. Some of the major challenges in the development of this application are

1. GUI design can be drawn with ball pen, marker pen or even by a pencil. The thickness is a major issue which needs to be considered.
2. Image continuity is another challenge which needs to be handled by our application. For *example*, in case of checkbox, even if all four sides are not perfectly connected. It is to be identified as a square and then as a checkbox.
3. GUI design is drawn without any help of graphical tools and hence radio button may not be an exact circle but could be an ellipse. This irregular circle is to be identified as a radio button.
4. A textbox and button drawn on a paper look same. Shape of textbox and button is a rectangle. In order to distinguish between two shapes and recognize correct one is a challenge.

4. OUTLINE OF THE DOCUMENT

The rest of the report is organized as follows. Section 5 covers overview of existing work done in generating digital pages from document images. Section 6 outlines 'The proposed approach of converting UI design document to static web page'. In this section we discuss series of steps proposed to analyze various algorithms used for detecting HTML controls. Finally the paper ends with conclusion.

5. RELATED WORK

[Hassan, 2007] proposed wrapper-based approach for detecting tables in PDF documents. He has proposed an algorithm to convert a wide variety of tabular presentations into HTML for information extraction purposes. The algorithm detects tables in PDF files, and correctly identifies their respective rows and columns. The algorithm also explains how to recognize spanning rows and columns, and multi-line rows. [Jiang and Yang, 2009] proposed a method to convert PDF document to HTML document with the same layout format. The information extraction and browsing online is easy in case of HTML file. [Ji-YeonLee, 2000] proposed a method of converting UI design document, into Hypertext document. He explained two methods for converting complex multi-column document images into HTML documents, and a method for generating a structured table of contents (ToC) page based on the logical structure analysis of the document image. The proposed method can generate structured table of contents page, with the hierarchically ordered section titles hyperlinked to the contents. [Leo, 2012] proposed a system, which converts a block diagram drawn on a paper into a machine-readable format using image-processing technique. In the block diagram all diagrams are drawn with pencil and hence are perfect geometric shapes. [Priyadharshini, Vijaya, 2013] proposed a method for Document Segmentation and Region Classification Using Multilayer Perceptron. They demonstrated the modeling of document segmentation as a here are various algorithms in shape recognition. But most of them work on mathematical formulas for detecting circle, square, rectangle etc. UI design is generally drawn without scale and hence it does not contain perfect shapes. Thus, existing circle detection algorithm may not detect a radio button as a circle since it is drawn with hand and may not be a perfect circle. So instead of using existing shape detection The work done so far in converting paper document to digital image considers text, images, and drawings like block diagrams. None of these projects have considered actual HTML controls. Even though few papers consider UI design document, it does not consider the HTML controls like textbox, radio button, checkbox etc.

There are various algorithms in shape recognition. But most of them work on mathematical formulas for detecting circle, square, rectangle etc. UI design is generally drawn without scale and hence it does not contain perfect shapes. Thus, existing circle detection algorithm may not detect a radio button as a circle since it is drawn with hand and may not be a perfect circle. So instead of using existing shape detection algorithms, it is required to implement new algorithms for detecting shapes.

A. EXISTING SYSTEM

Figure 1 shows Existing System Flowchart. This is an exiting methodology followed when taking approval for final web page design from the client.

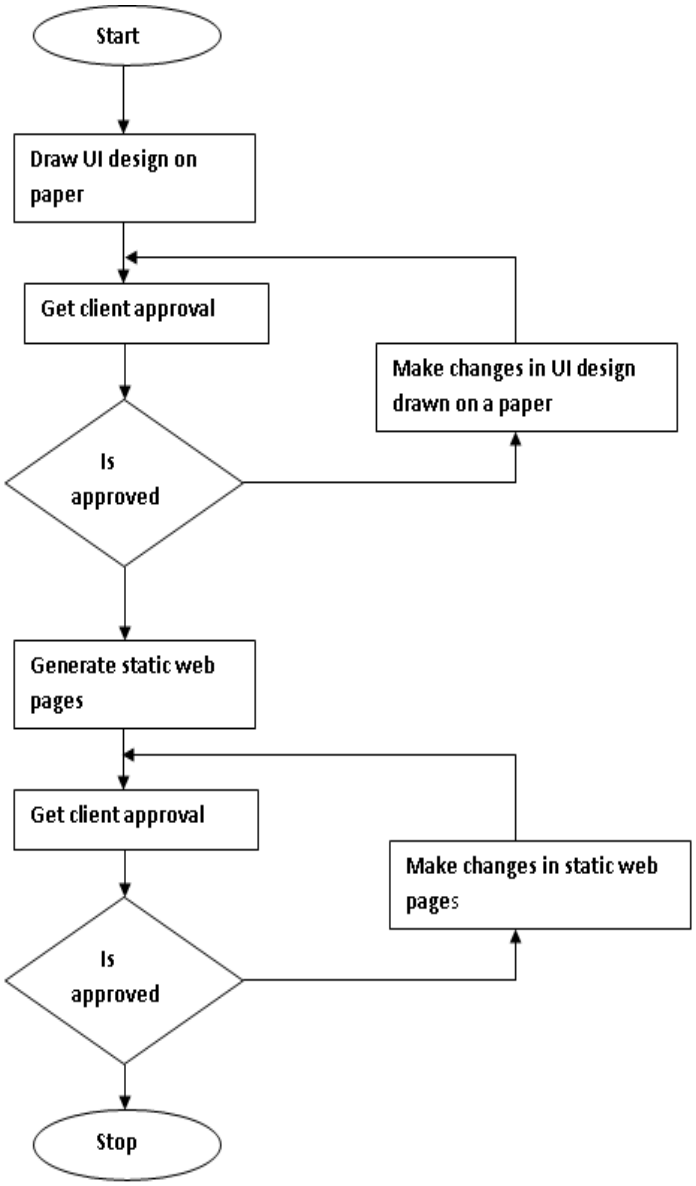

Figure 1: existing system flowchart

Classification task and describe the implementation of machine learning approach for segmenting the document into various regions.

The work done so far in converting paper document to digital image considers text, images, and drawings like block diagrams. None of these projects have considered actual HTML controls. Even though few papers consider UI design document, it does not consider the HTML controls like textbox, radio button, checkbox etc.

6 THE PROPOSED APPROACH

Figure 2 shows the internal blocks of the proposed software. There are 3 separate databases one for original scanned

VARIOIS GUI CONTROLS DETECTION ALGORITHM

Algorithm isButton (centerX, centerY)

Scan the image at center coordinates centerX, centerY.

If data is present at the center then Return true

Else Return false

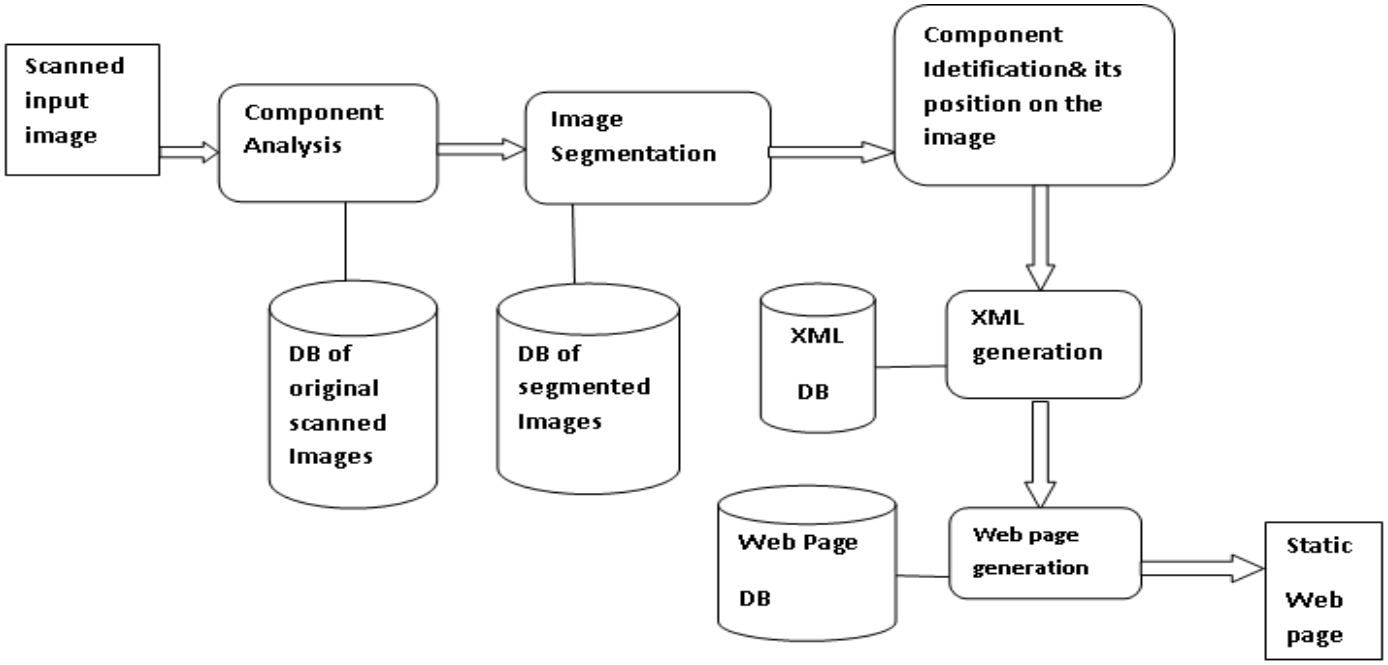

Figure 2 System Architecture

images, second database is of segmented images and third database is of generated static web pages. Component analysis block uses scanned image database. The output of this block is the different components which need to be segmented first. The segmentation happens in the next block, Image segmentation. Segmented images are kept in a separate database. Once these components are segmented, each component needs to be identified as a type of HTML control. Prposed system uses XML DB to store identified HTML controls with their position on the image. Information from XML database is retrieved using XML parser and corresponding static web page is generated. This page is stored in a separate database.

Algorithm isCheckBox (heightY, widthX, centerX, centerY)

If (widthX = heightY OR difference between width and height is NOT greater than 10) then

Call flag = isRadioButton (heightY, widthX , centerX, centerY)

If flag is true thenReturn radio button is detected.

ElseReturn checkbox is detected.

Algorithm isRadioButton (centerX, centerY)

Scan the image and find the length of first horizontal line detected as HL.

Using center coordinates find the length of diameter as DiLgth.

If HL = DiLgth OR

Difference between HL and DiLgth is not greater than 10 then Return true

ElseReturn false

Algorithm isTextBox(heightY, widthX, centerX, centerY)

If (widthX>heightY and difference between width and height is greater than 20) then

Call flag = isButton(centerX, centerY)

If it is not a button then

return textbox is detected.

else return button is detected

7. CONCLUSION

This pape describes analysis of the document Images of GUI designs drawn by hand and have created functional web page from those images. It describes the implementation of various HTML control algorithms for generating HTML page from the document image. Proposed software takes care of all types of GUI images for e.g. image drawn by hand, image drawn using scale, image drawn using different thickness pencil. HTML basic controls such as Radio button, checkbox, textbox, button, are implemented successfully. So far the work done was restricted only to the document layout and the text on the document. The proposed software not only considers the layout of the document but also various HTML controls and their position on the image. The proposed software can be used by any non-technical person to generate the web pages automatically from a UI design document. This will save developer's time and will help in rapid web development.

REFERENCES

[1] Hassan, T., Baumgartner, R. "Table Recognition and Understanding from PDF Files" International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil(2007)1143-1147

[2] Jiang, D., Yang, X "Converting PDF to HTML Approach Based on Text Detection" 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human.ACM New York, NY, USA, Seoul, Korea (2009)

[3] Ji-Yeon Lee, Jeong-Seon Park, HyeranByun, JongsubMoon, Seong-Whan Lee, Pattern Recognition Society. Elsevier Science Ltd, December 2001

[4] Klink, S., Dengel, A., Kieninger, T. "Document Structure Analysis Based on Layout and Textual Features" International Workshop on DocumentAnalysis Systems, Rio de Janeiro, Brasil (2000)41-52.

[5] Leo G Vailati, "block diagram detection", EECS 741 - Computer Vision, EECS - Dept. of Electrical Eng. & Computer Science KU - The University of Kansas (2012)

[6] Oro, E., Ruffolo, M.: PDF-TREX "An Approachfor Recognizing and Extracting Tables from PDF Documents" 10th International Conference on Document Analysis and Recognition 2009. IEEE ComputerSociety,Barcelon

[7] Priyadharshini N1, Vijaya MS "Document Segmentation and Region Classification Using Multilayer Perceptron", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013

[8] Rosmayati Mohemad, Abdul RazakHamdan, Zulaiha Ali Othman and Noor MaizuraMohamad Noor, "Automatic Document Structure Analysis of Structured PDF Files", International Journal on New Computer Architectures and Their Applications (IJNCAA) 1(2): 404-411, The Society of Digital Information and Wireless Communications, 2011

[9] Sneha Sharma, Dr. Roxanne Canosa, advisor "Extraction of Text Regions in Natural Images" Rochester Institute of Technology, 2007

[10] Yildiz, B., Kaiser, K., Miksch, S. "pdf2table: A Method to Extract Table Information from PDF Files" Indian International Conference on Artificial Intelligence, India (2005) 1773–178512