International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

Graphics Extraction - Vector Processing XSLT and XML Transformation

[1] Irudaya Ratnam S [2] Dr. Nalini Chidambaram, [3] A. Punitha
[2] Bharath University [3] MNM Jain
[1] irudayams@gmail.com [2] drnalinichidambaram@gmail.com
[3] sweetpunitha@gmail.com

***Abstract*:-- Software application has reduced many manual related repeated works. Traditional sparse image models treat color image pixel as a scalar, which represents color channels separately or concatenate color channels as a monochrome image (Sharpening the images done by designing team).**

In this project, we propose a vector art representation model for images using quaternion matrix analysis. End user will view these images from DOCX or any other form of documentations for example: PDF file, these document files are for example: prepared by various Doctors using chemical theories and published for student reference. If user wants information regarding study materials, student will have links to the website then they need to download the document files and they will refer the images for their studies. To reduce these manual work process called manual image processing (MIP). In this project new approaches got implemented that is vector art graphics extraction. Using this algorithm user will create images and published for the end user for example students. Once published end users will refer these images and user will use it for their own educations study material preparation purpose. Also by separating these images it will reduce the size of storage in the local system disk space. Since if we store the document files it would occupy more space when compare to individual image files with captions.

The images will be in the following order: Horizontal, Vertical and Image inside the image that is small part of the image over the part of main image. Current system will provide only sharpen the image. That is Vector sparse representation, quaternion matrix analysis, color image, dictionary learning, and image restored in a location. Only related images user can extract and prepare with their study materials, instead of depending on the document. These approaches reduce the disk space.

***Keywords*: DOCX represents Latest Microsoft Office word document Files, DOC is for Microsoft Office word document file, MIP stands for manual image processing, PDF is Portable Document File, IPTS is Image Processing Transformation System, IPGE is Image Processing Generation, XSLT/XSL Extensible Style sheet Language and XML is Extensible Markup Language**

I. INTRODUCTION

Traditional system does the image sharpening which represents color channels separately or concatenate color channels as a monochrome image. In this project the main aim is to reduce the manual image extraction from the Microsoft word document i.e. sample.doc or sample.docx file(s). Due to educational portal and it's related to chemical study material. Doctors will specify the document files with images and hints of an image as a caption of the images. In this project the goal is to transform the xslt or xsl file from document.xml and generate XML file to create the image extraction path information. Generate the images with caption and stored in the designated folder location from website end user can view the images directly instead of downloading the document files and doing the sharpening using the existing system/application.

Consider a format study material with image(s) and caption if the input file is DOC. i.e. 'sample.doc' file. Covert the document file to latest version of Microsoft office document file i.e. DOCX file. For example 'sample.doc' files to 'sample.docx' file. Rename the selected document file to zip format file. For example 'sample.docx' files to 'sample.zip' file. Unzip a zipped file. For e.g.: 'sample.zip' files.

Transformation: In Unzip folder find a word folder and there will be a document.xml file. Using this document.xml file create an XSLT file as document.xsl file. Using saxon9pe.dll file Transform it to data.xml file. Generate IPGE: From the data.xml file find the Graphics node for image path details. From the data.xml file find the Caption node for caption about the image(s). Using the .Net framework graphics name space find one or more images along with the caption and generate single format image as sample.tif file with dpi resolutions. Zooming Image: By selecting an image, user can view the full zoom of an image. View from Website: Once images generated, user can view the images for their study

purpose from website. Also in this stage storage of file size got reduced

In this project the main aim is to reduce the manual image extraction from the Microsoft word document i.e. sample.doc or sample.docx file(s).

Due to educational portal and it's related to chemical study material. Doctors will specify the document files with images and hints of an image as a caption of the images.

In this project the goal is to transform the xslt or XSL file from document.xml and generate XML file to create the image extraction path information

II. AUTHENTICATION AND AUTHORIZATION

1. Authorization:

The process of authorization is distinct from that of authentication. Whereas authentication is the process of verifying that "you are who you say you are", authorization is the process of verifying that "you are permitted to do what you are trying to do". Authorization thus pre-supposes authentication. For example, a client showing proper identification credentials to a bank teller is asking to be authenticated that he really is the one whose identification he is showing. A client whose authentication request is approved becomes authorized to access the accounts of that account holder, but no others. However note that if a stranger tries to access someone else's account with his own identification credentials, the stranger's identification credentials will still be successfully authenticated because they are genuine and not counterfeit, however the stranger will not be successfully authorized to access the account, as the stranger's identification credentials had not been previously set to be eligible to access the account.

Similarly when someone tries to log on a computer, they are usually first requested to identify themselves with a login name and support that with a password. Afterwards, this combination is checked against an existing login-password validity record to check if the combination is authentic. If so, the user becomes authenticated. Finally, a set of pre-defined permissions and restrictions for that 3 particular login name is assigned to this user, which completes the final step, authorization. Even though authorization cannot occur without authentication, the former term is sometimes used to mean the combination of both.

2. Authentication:

Authentication is the act of confirming the truth of an attribute of a single piece of data claimed true by an entity. In contrast with identification which refers to the act of stating or otherwise indicating a claim purportedly attesting to a person or thing's identity, authentication is the process of actually confirming that identity. It might involve confirming the identity of a person by validating their identity documents, verifying the authenticity of a website with a digital certificate, determining the age of an artifact by carbon dating, or ensuring that a product is what its packaging and labeling claim to be. In other words, authentication often involves verifying the validity of at least one form of identification.

Authentication is relevant to multiple fields. In art, antiques and anthropology, a common problem is verifying that a given artifact was produced by a certain person or in a certain place or period of history. In computer science, verifying a person's identity is often required to allow access to confidential data or systems. Authentication can be considered to be of three types:

- The first type of authentication is accepting proof of identity
- The second type of authentication is comparing the attributes of the object itself to what is known about objects of that origin.
- The third type of authentication relies on documentation or other external affirmations

In this project the first type of authentication is used as proof of identity. Following is the screenshot from this project.

III. MODULES DESCRIPTION

There are three modules in this project. They are

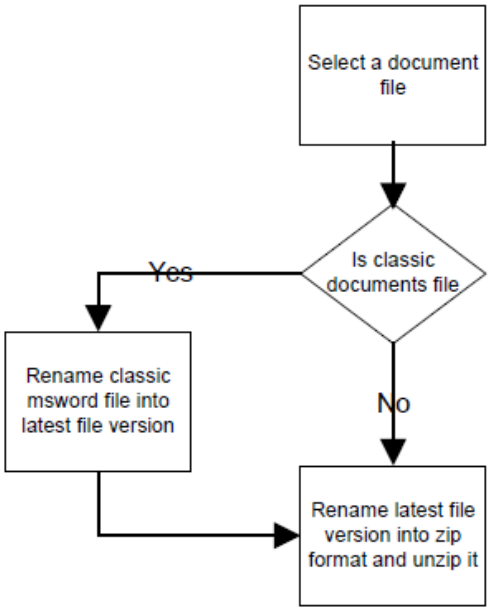
- Conversion logic.
- Transformation Algorithm.
- Generate Stitched Image or Images

1. File Conversion Logic

Once authentication and authorization process done, useris able to access the application as per

the association role. From document file need to extract images. This was in manual process for sharpening images in existing system. In this project we make this process in to compute system application. Using this approach system will the check the upload file is which kind of document files? Current approach will check for only Microsoft windows word document files (classical and latest files). If it is classical old version of files, this approach will convert into latest version of file.

Once it renamed, by using this new approach system will rename this new latest version of file into zip format file. Once it renamed. Through application using this proposed method, it will unzip the zip file into to a specific folder location. From this approach now we have all the images in specific folder location and other related files that will help to generate images in proper format as it in document file. Following figure will show this approach:

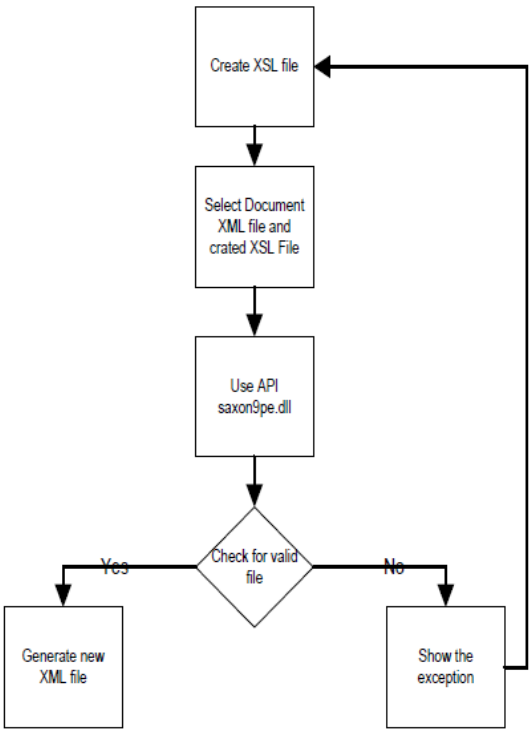


2. *Transformation Algorithm*

In this module, main process is Image Processing transformation System– Algorithm. For this transformation algorithm, the API / DLL used are saxon9pe.dll. For this step, first need to create XSL/XSLT file, named as document.xsl. This file needs to create using document.xml file, which was extracted from the zip file of document from the first module as conversion logic. Input will be the xml and xsl files for this transformation. API will check and compare each nodes and attributes of the XML and XSL files and will generate new XML file called data.xml.

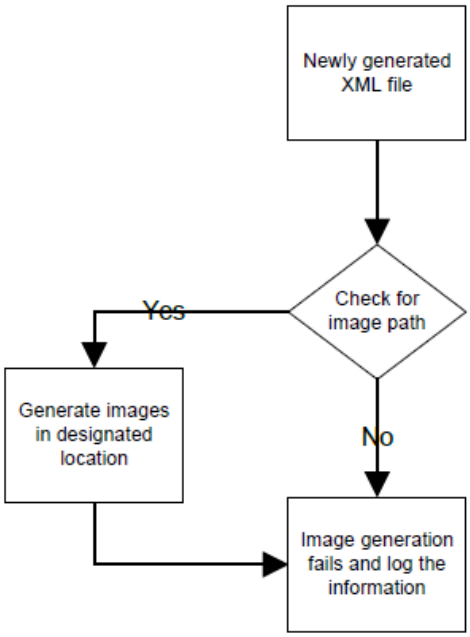
IV. TRANSFORMATION LOGIC:

Here first we created xsl or xslt file as document.xsl. All word documents must take the same value as same document.xsl file. Here word documents are created with predefined image caption method. Following figure will show this step:



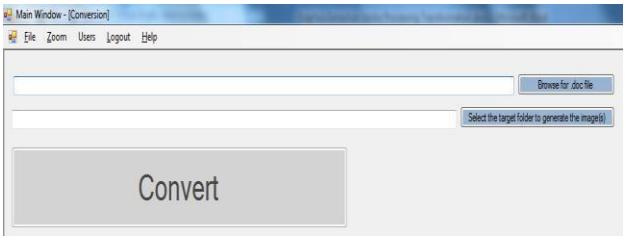
3. *Image generation*

In this module images will get generate from the document file using the newly created XML file created using previous module. Check for the newly created XML file. In this file the information of the image path location and caption of the images also available. All will be in the form of attributes, check for image path if available generate the stitched image if it is more than one images otherwise it's a single image. More than one image will be in any order that is horizontal, vertical and images inside the image. Following figure shows this approach:

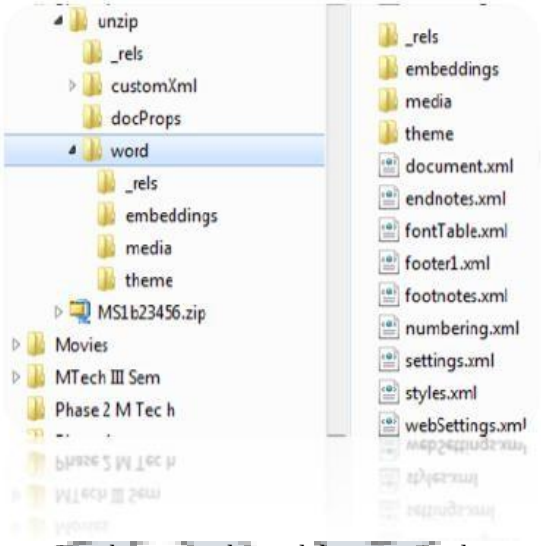


V. MODULE WISE REQUIRED INPUT AND OUTPUT

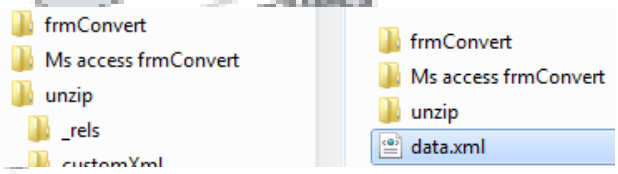
In the first module document file is the input and the document file is in the format of doc or docx files. Here based on the input format the conversion logic will do the needful. If the input file is in MSWord old format for example doc file then the conversion logic will have one step more action that is converting the doc file to docx file. This step is to have a new format of docx file also all docx files are in zip format. Due to this approach the first conversion logic is applied. Also all document files whether it is in doc or docx format files, the predefined format is that the images inside the documents and the text should be in the form of image caption format. This is the based pre-defined format requirement for this image extraction approach. Also through this module the zip file will be unzipped into the respected folder with all inner file formats of document files. The main file to get all the information is document.xml. Following is the screenshot for conversion logic:



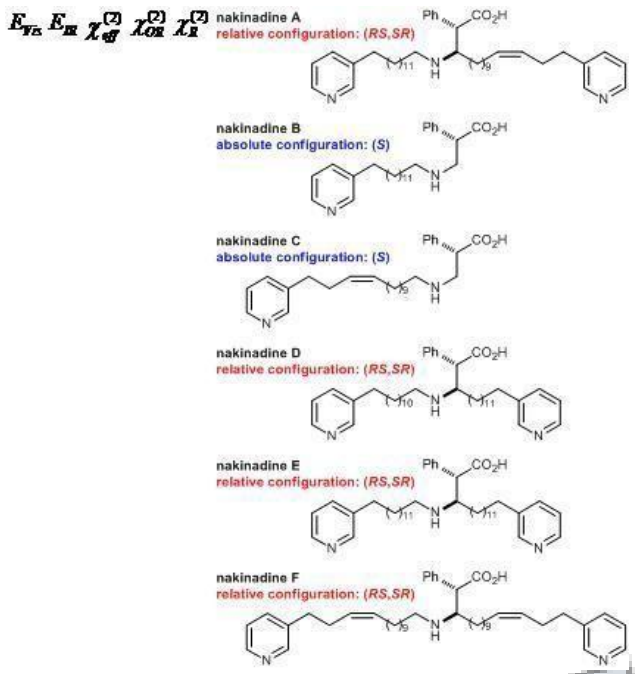
Unzip folder created after conversion logic applied as shown is below screenshot:



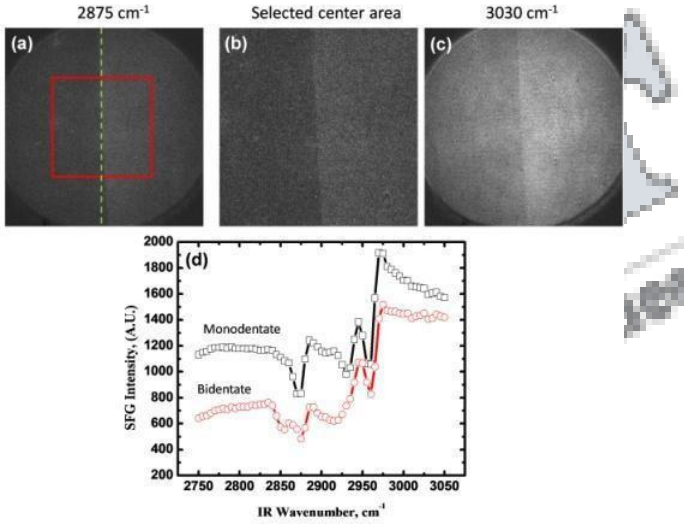
In the second module we implement the transformation algorithm. In this using the document.xml file and we created xml style sheet file based on document.xml file named as document.xsl. For transformation logic these two are input files and by using the saxon.dll the transformation algorithm got applied and produces a new file called data.xml file. This is the main file which will have all the information related to the images and its caption or text.



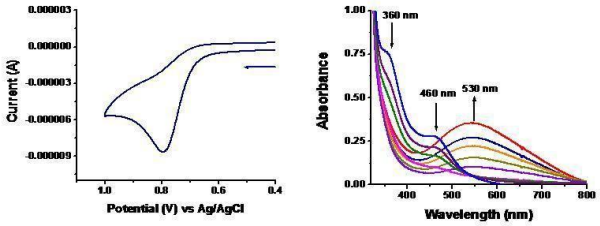
In the third module we implement the logic of generating images that is extracted from the document files. Images generated in different format that is it will be in the form of vertical, horizontal and images over the other images. Following are some of the generated images:



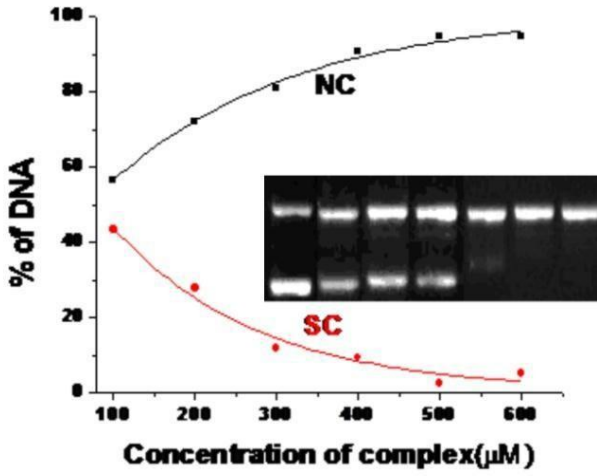
Above figure shows the equations of images extracted from the document file including 300 dpi image qualities.



Above images is a single image type.



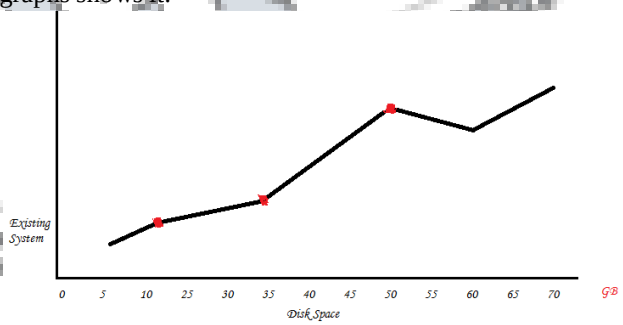
Above image is a horizontal position images that contains two images and the output is stitched images into a single image.



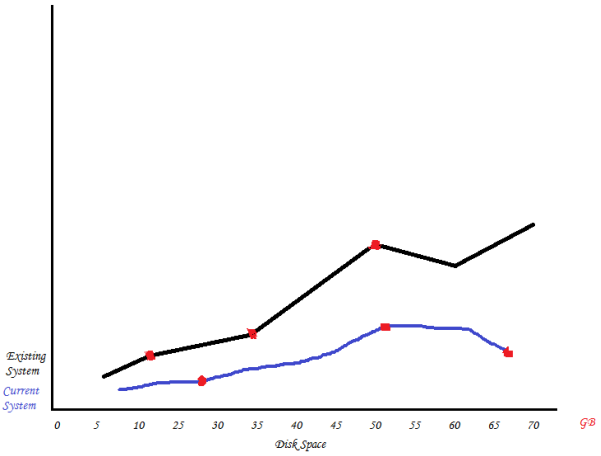
Above images shows the image side other images.

Advantages and Disadvantages:

Before this approach user will have the document files and some of the portion process was in manual like Photoshop painting and graphics related jobs. This approach will occupy more memory spaces. Following graphs shows it:



In this project, the image generation process will reduce the image separation manual process not all some of the manual process and it is reducing the memory space too. Following graphs will shows it:



VI. CONCLUSION

Here we conclude that we present new ideas to reduce the manual image processing from the document files that is the images are cropped from the document files and with the help of Photoshop graphics designers will generate the images as shown in previous pages. Also by using this project instead of downloading the entire document file users can only download the respected images. This will save disk space too. This is one of the important pro of this project.

FUTURE ENHANCEMENT

In furthermore, apart from MSWord document need to implement the same approach for other document format files for example .pdf files. Likewise even in MSWord document files images over the other images and images inside the tables are not able to generate. However further studies are needed to upgrade the performance.

REFERENCE:

1] Graphics extraction from heterogeneous online documents with hierarchical random fields – 2015 - Published: Friday, 25 December 2015 - Graphical objects are important elements of freely handwritten notes but their segmentation from the document is challenging due to their irregular properties. This project introduces an original solution for automatically segmenting diagrams and drawings from unstructured online documents (Basic logic of this project got from this project approach.)

2] Context-Aware Patch-Based Image In-painting Using Markov Random Field Modeling – 2015 - Published: Friday, 25 December 2015 - where textural descriptors are used to guide and accelerate the hunt for well-matching (candidate) patches. A completely unique high-down splitting procedure divides the image into variable size blocks consistent with their context, constraining thereby the rummage around for candidate patches to nonlocal image regions with matching context. (This part of logic used to split images and its text context)

3] Semi supervised Biased Maximum Margin Analysis for Interactive Image Retrieval – 2012 - Published: Thursday, 27 September 2012. - With many potential practical applications, content-based image retrieval (CBIR) has attracted substantial attention during the past few years. A variety of relevance feedback (RF) schemes have been developed as a powerful tool to bridge the semantic gap between low-level visual features and high-level semantic concepts, and thus to improve the performance of CBIR systems. (Using this approach we implement content based 8 images extraction logic using xml and xslt/xsl)

4] Image Super resolution Using Support Vector Regression - Published: Monday, 25 June 2012 - Support vector machine (SVM) is a statistical learning algorithm that is capable of estimating high-dimensional functions. Recently, support vector regression (SVR) - the use of SVM for regression - has been used to generate super-resolution images. In this paper, we propose to apply the SVR algorithm on edge pixels only so as to reduce the emboss effect that would appear in the edge region of an enlarged image if the SVR is applied on the entire input image. Image vector processing principle and logic used by this approach.