A STUDY OF DECODING COMPRESSED IMAGE TIFF FILES OF CANDIDATES RESPONSE SHEETS IN OBJECTIVE TYPE ENTRANCE TESTS

Sapna Mittal [1], Hema Arora [2]

[1 ,2]Computer Science , RSD College , Ferozepur

I. INTRODUCTION

Recent advances in communication and computer technologies have made possible the exchange and retrievals of information through, the electronics media. To store and transmit digital data efficiently, variety of data compression algorithms has been developed. Yet, information in compressed data cannot be easily retrieved or modified without decompression. While the research body on compression schemes is quite rich, algorithms for retrieving and modifying information in compressed data have not been widely integrated.

The objective of this research is to identify operations that can be applied directly and efficiently to digital information encoded by a given compression algorithm. A formal method of analysis is planned. This project mainly focuses on lossless compression techniques and images processing and decoding algorithms. This project mainly relies on two technologies viz. Optical Mark Recognition (OMR) andi mage Compression/Decompression.

Under the system, first the images of candidates' responses are scanned and stored in compressed format. The file type used is TIFF as it supports lossless compression techniques and is portable across machine architectures. Presently Bi-level images are supported as their storage requires very little amount of storage space, i.e. one bit per pixel as compared to gray scale or RGB images, which take a lot of storage space.

After scanning and storing the images the next step is to decode the images. For this purpose first the image is decompressed in memory and then decoding algorithms are applied, which convert the Optical Marks into ASCII text.

II. OPTICAL MARK RECOGNITION

OMR refers to the technique of converting a handwritten mark into an ASCII value. Filling a circle or a box or a special form, which is designed as per the given specifications, makes a mark. The presence or absence of a mark in a specific location is then converted into a value such as a selection in multi-choice question, the selection of one item in a list of several or even to code a specific numeric or alphanumeric value. Thus, if a mark by it self can only generate a binary choice (presence or absence), the combination of several marks arranged on a form can provide intelligent answers to questions and be used as a fast and accurate replacement to data entry in many applications. The form data represented by mark positions are translated to ASCII text records.

Advantages of OMR versus manual Key-entry are as follows:-

- **Accuracy: -** The source document is the input document, eliminating the risk of manually key entering wrong information.
- **Speed: -** Depending on the scanner model and OMR reader can read documents at speeds up to 8,000 forms per hour. A typical full sheet OMR reader for example will have up to 3840 possible mark-character interpretations per side, and if a form used only 250 marks, this would equal to 2,000,000 data characters per hour.
- **Cost: -** Time and money are saved by increasing accuracy and decreasing the time and numbers of personnel needed to complete a data entry project.
- **Types: -** OMR readers typically come into varieties, Full Sheet Readers and Card readers. While there are differences in the physical characteristics of each type, they both serve the several purpose, to collect and interpret data in a low cost, high volume .

Since the data for the evaluation of the results even if compressed should not be lost hence we use lossless compress technique. TIFF file format is one such format, which handles data using lossless compression technique.TIFF file format is one such format, which handles data using lossless compression technique.

III. SCANNING

Scanning is the first step towards decoding of candidates' responses For this purpose, response sheets are fed into scanner, which captures the images of the response sheet and stores the image into a specified image file format. For this project the file format opted for storage of images, is TIFF. While storing the image certain choices are to be made pertaining to type of image (Bi-level, gray scale, RGB) compression techniques and image resolution etc. These factors are largely responsible for determining the size of the image.

IV. BI-LEVEL IMAGE

Bi-level image types are used when we want to store images with 2 colors only i.e. Black and White. Bi-level images require only one bit per pixel for representation of image and therefore require least amount of storage space.

V. RESOLUTION

Resolution determines the level of detail recorded by the scanner, and is measured in dots per inch (dpi).The greater the dpi number, the higher resolution is and the higher is the storage space required for the image.

In the present project recommended is 75dpi as it was found to be significant representing images clearly. However, higher resolutions, in multiple of 75(i.e. 150,225,300) etc. are also supported but their use is not encouraged as they only increase the storage space required for the image, without being efficient for decoding purposes.

VI. RESOLUTION

Resolution is the number of dots available to represent graphic detail in a given number area:-

On a Computer Screen	The number of pixels per linear inch –ppi (72 to 96 ppi is the Maximum a monitor displays).
On a Printer	The number of dots printed in a Linear inch –dpi.
On a Scanner	The number of pixels scanned per linear inch of the

	scanned image–ppi.
Image Resolution	It is measured in pixels per inch- ppi
Printer Resolution	It is measured in dots per inch-dpi

If an image a resolution of 72ppi, this means it contains 5184 pixels in a square inch(72 pixels wide *72 pixels high= 5184)

Because the number of pixels in an image is fixed, when resolution decreases, the image size increases. Conversely, if we increase the resolution of an image size (dimension) will decrease.

High resolutions allow for more details and subtle color transitions in an image.

Image processing

An image consists of a two dimensional array of number (f(x, y)). The color gray scale displayed for a given picture element (Pixel) depends on the number of bits stored in the array for that pixel. The simplest type of image data is black and white i.e. bi-level. It is a binary image since each pixel is either 0 or 1.

TIFF SPECIFICATIONS

TIFF stands for the **Tagged Image file Format**. TIFF is used to describe the image data that typically comes from scanners, frame grabbers, and paint and photo retouching programs. The **TIFF** format was designed to be portable across machine architectures. It is governed by a specification maintained by Aldus Corporation and revised occasionally to reflect advances in scanner and digitizer technology.

For the purpose of this project **TIFF** has been chosen for storage of images because of the following features that are provided by the **TIFF** files:

- TIFF is capable of describing bi-level, gray scale, and palette-color and full-color image data in several color spaces.
- TIFF includes a number of compression schemes that allow developers to choose the best-space or time trace off and for their applications.
- TIFF is not tied to specific scanners, printers or computer display hardware.
- TIFF is portable. It does not favor particular operating system, file systems, compilers or processors.
- TIFF is designed to be extensible to evolve gracefully as new needs arise.
- TIFF allows the inclusion of an unlimited amount of private or special-purpose information.
- **TIFF Structure**

The **TIFF** structure that is used to store the image data is broadly divided into three main parts: **TIFF HEADER** **TIFF** file begins with 8 bytes image file header. Header is divided identifier and offset of the first IFD (image file directory). There are two variables of TIFF, which are shown in the following figures:

Byte Number	0	1	2	3	4	5	6	7
Header Contents	49	49	2A	00	08	00	00	00
	Byte Order		TIFF Identifiers		Off set of the first IFD			

Sort Order

The entries in an IFD must be sorted is ascending order bytes value. The values to which they point need not be in any particular order.2.1.3 image file directory: the first two bytes of the first IFD entry table contain the number of IFD entries that present is the IFD entry table. Each IFD entry is 12 following four fields:

- **Tag :**The first 2 bytes (i.e. 0-1) are used by the tag field. These tag fields helps to identifier as to what sort of the image information is being help by the IFD entry.
- **Count:**The next 4 bytes (i.e. 4-7) store the value for the count. The count tells us how many values are to be read of the type designed byte by the type.
- **Value /offset :**Depending on the value stored in the count- field the 4 bytes (i.e. 8-11) of this field gives the value or the offset in the file. If the value of the count is one then we know that there is only one value of the type defined by the type bytes. Hence the value gives the value for the field defined by the tag field. But if the value of the count is not one then the value gives us the offset in the file where the values can be found. The number of values to be read from the offset is given by the count field & the number of bytes to be read to correctly read the data is defined by the type field. The field contains the value only if the value fits into four bytes. If the value is shorter, it is left-justified with in the field.

Multiples Images : A TIFF file may contain several images. Each has its own IFD, which defines a sub-file. Sub-files may hold related images, such as the pages of a facsimile transmission or scanned images using ADF (Automatic Documented Feeder).

Image Types : The Base line TIFF supports four kinds of images.Bi-level, grey scale, palettes color a RGB full color. Bi-level images contain two colors: black and white. They take the least amount of space but do not allow for color or shades of grey. Grey scale images contain information about shades. Either four or eight bits are recorded for each pixel, corresponding to either 16 or 256 shades of grey. Palette color images resemble grey scale images, except that the pixel value is an index into a color map, where the red, green and blue components range between 0 and 655535.

The Required Fields for the Bi- Level Image files

Sr. no	Tag Name	Tag	Type	Value	Description
1.	Image width	256	Short / Long		This field gives us the image width i.e., the pixels in the row.
2.	Image length	257	Short / Long		This field gives us the image length i.e., the number of rows of pixels in the image.
3.	Compression	259	Short	1,2 or 32,773	This field gives us the information about the type of compression technique used. It has default value of 1 i.e. no compression.
4.	Photometric	262	Short	0 or 1	This field gives us the information about the

	Interpretation				photometric interpretation used. It does not have any default value. In TIFF file whole of the image data is broken down into strips of equal length of the image data is more than 8kb.This breaking of the image file into strips help in the better processing of the image data & the efficient I/O buffering.
5.	Strip offset	273	Short / Long		This field gives us the strip offsets for the image file.
6.	Row per Strip	278	Short / Long		This field gives us the rows per strip value for the compressed image files.
7.	Strip byte counts	279	Long / Short		This field gives us the information about the Strip byte counts.
8.	X Resolution	282	Rational		This field gives us the X Resolution for the image i.e., the number of pixels per resolution unit in the image width direction.
9.	Y Resolution	283	Rational		This field gives us the Y Resolution for the image i.e., the number of pixels per resolution unit in the image length direction.
10.	Resolution Unit	296	Short	1,2 or 3	This field tells us about the Resolution unit used for the image. The default value is 1 i.e., for the inches.

Compression Algorithms

Data compression techniques can be divided into two major areas:

The basic differences between the two different types of compression are:

LOSS LESS	LOOSY
A Compression scheme in which no bits of information are permanently lost.	A Compression scheme in which some bits of information are permanently lost during compression and decompression of an image.
The most preferred image format for storing images and that uses a lossless compression scheme is the TIFF format.	The loss is usually only minimal and hardly detectable. The most common image format that uses a lossy compression scheme is the JPEG format.
When converting an image to TIFF format, you have the option to have the image display any number of colors up to 256.	JPEG is a very efficient, true-color, compressed image format Although it is lossy, it has the capability of showing you more colors than GIF (more than 256 colors).

For most types of data, lossless compression techniques can reduce the Space needed by only about 50%.	JPEG has the ability to achieve much greater compression. It can reduce files 'size to about 5% of their normal size

Methods of Data Compression/Decompression

Following are the various algorithms that are used for the data compression used for the data compression used for the data compression used in TIFF (Tagged Image File Format):

I) Run Length encoding (RLE) algorithms

II) Lempel-Ziv-Welch (LZW) algorithms

III) CCITT

IV) Huffman algorithms

The first two algorithms, namely **Huffman** and **RLE** are based on **Statistical** method where as the third one, i.e. **LZW**, is based on **Dictionary** method.