

Unstructured Invoice Recognition and Processing with **LEADTOOLS**

LEADTOOLS[®]
THE WORLD LEADER IN IMAGING SDKs

 **LEAD
TECHNOLOGIES**
I N C O R P O R A T E D


Introduction

There is no doubt that imaging technologies such as forms recognition and OCR have revolutionized the way we do business and have freed up countless man-hours from the tedious and error-prone task of data entry. However, these technologies still rely on the assumption that the data will be located in the same place on every document. Unstructured forms and scanned documents such as invoices, bills and tabs are notoriously difficult, if not impossible, to process on a large scale within an automated system.

This is no longer the case after the release of LEADTOOLS Version 19 and its advanced Invoice Recognition and Processing SDK technology. Complex, unstructured and semi-structured documents with tabular data can be processed even when each row and cell have different heights, or when the table spans across multiple pages. In this white paper we will show how to set up a master form template then extract the data from a filled invoice.

Setting up the Master Form Template

The first requirement is to create a template which defines where the various fields are located and what kind of data to expect. This should be a blank or redacted version of the form to recognize. Some fields may have static locations such as the customer number, invoice number, addresses, etc., but tables found on invoices and other unstructured form types can wreak havoc on recognition accuracy since the master template and filled invoice differ in size due to the table area.




CUSTOMER #:
INVOICE #:
DATE:

SOLD TO:
SHIP TO:

ATTENTION _____ PHONE: (555) 555 - 5555
 SHIP VIA: _____ TERMS: _____ SHIP DATE: _____
 CUSTOMER PO#: _____ SALESPERSON: _____ DUE DATE: _____

ITEM	QTY	DESCRIPTION	PRICE	AMOUNT

Inv Total (USD):



CUSTOMER #: 66T3000
INVOICE #: 0090152-IN
DATE: 12/20/2014

SOLD TO: John Doe
13420 Johnny Cake Avenue
Houston, TX 77058
SHIP TO: John Doe
13420 Johnny Cake Avenue
Houston, TX 77058

ATTENTION _____ PHONE: (555) 555 - 5555
 SHIP VIA: _____ TERMS: CREDIT CARD SHIP DATE: 12/20/2014
 CUSTOMER PO#: _____ SALESPERSON: ECOM DUE DATE: 12/20/2014

ITEM	QTY	DESCRIPTION	PRICE	AMOUNT
DOCIMG19	1	LEADTOOLS Document Imaging V19	2995.00	2995.00
AMDOCIMG19	1	Annual Maintenance LEADTOOLS Document Imaging V19	599.00	599.00
OCRADV19	1	LEADTOOLS OCR Module Advantage V19	995.00	995.00
AMOCRADV19	1	Annual Maintenance LEADTOOLS OCR Module Advantage V19	199.00	199.00

Inv Total (USD): \$4788.00

Figure 1: Master Form Template compared to a Filled Invoice

LEADTOOLS overcomes this major obstacle with the Table field type. To implement a table field, simply define the column headers and the table data will be recognized regardless of the number of rows, height of each cell, or the number of pages. Traditional fixed location fields can also coexist underneath the table (e.g. Invoice Total) and will be discovered based on its location in proportion to the end of the table.

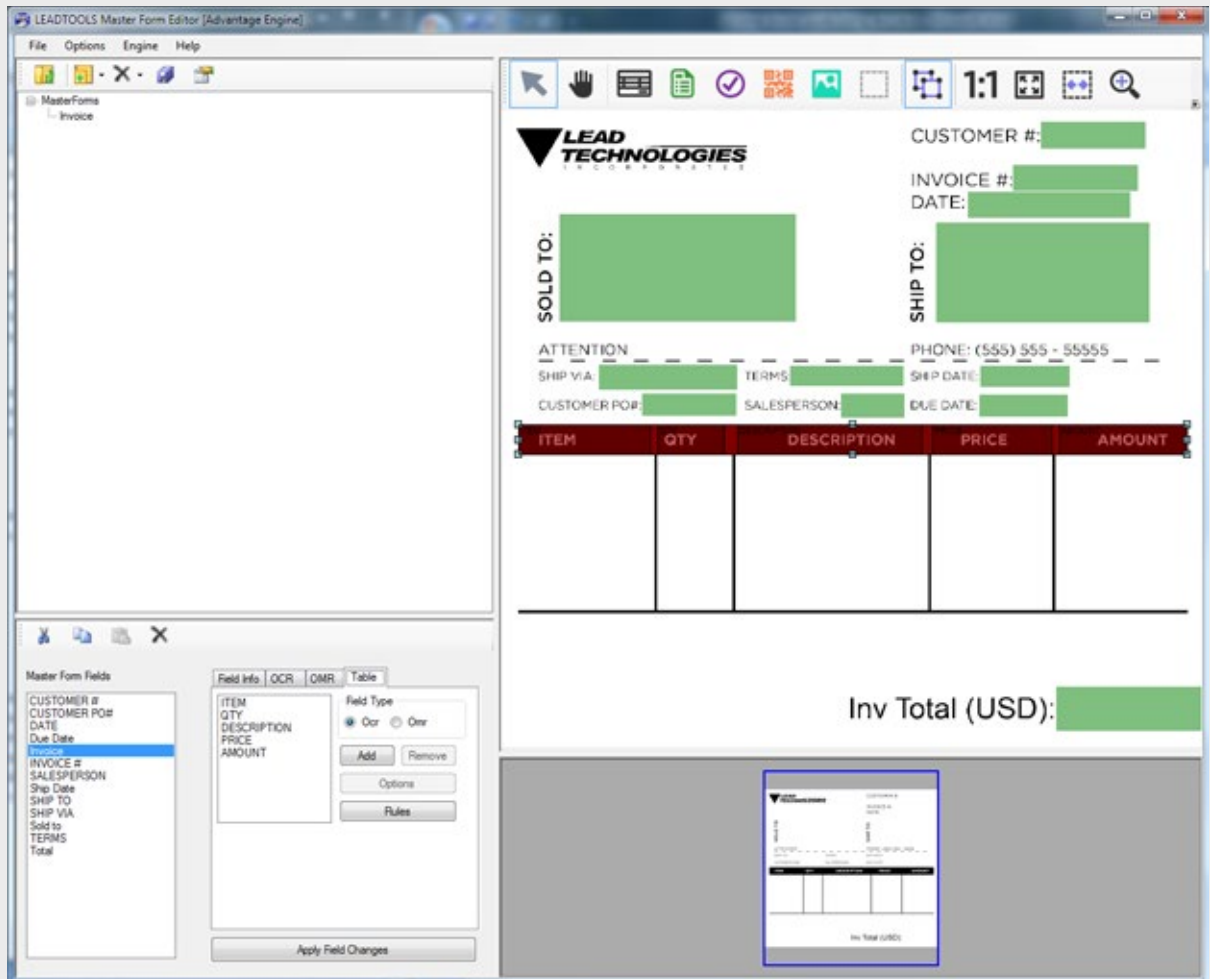


Figure 2: LEADTOOLS Master Form Editor Demo editing a master form with a table field in addition to traditional static fields

Recognizing and Processing a Filled Invoice

After setting up the master form templates, you are ready to start recognizing (i.e. classifying) and processing filled out documents. LEADTOOLS' advanced forms recognition algorithm will first determine what kind of form the filled document matches. There is no limit to the number of master form templates within a repository so it is possible to create a single, automated solution for all of the forms and invoices within an enterprise. The [AutoFormsEngine](#) does all of the work of aligning, orienting, classifying and processing the form regardless of its resolution. Best of all, setting up the [AutoFormsEngine](#) is quick and easy, requiring little more than telling it where to find the master form templates and passing it the filled invoice image.

```
// Set up the master forms repository based on a folder path
DiskMasterFormsRepository workingRepository =
    new DiskMasterFormsRepository(rasterCodecs, repositoryPath);

// Use the Ocr recognition manager. Other options include Default and Barcode
AutoFormsRecognitionManager managers = AutoFormsRecognitionManager.Ocr;

// Create the forms engine
AutoFormsEngine autoEngine = new AutoFormsEngine(workingRepository, ocrEngines[0],
    barcodeEngine, managers, 30, 80, _menuItemRecognizeFirstPageOnly.Checked);
autoEngine.UseThreadPool = true;

// Create the filled form we want to identify/regognize/classify
// against the repository
FilledForm form = new FilledForm();
form.FileName = Path.GetFileName(imagePath);
form.Name = Path.GetFileNameWithoutExtension(imagePath);
form.Image = rasterCodecs.Load(imagePath, 0, CodecsLoadByteOrder.Bgr, 1, 1);

// Run the recognition and processing on the filled form
AutoFormsRunResult result =
    autoEngine.Run(form.Image, MyPageRequestCallback, form, null);

if (result != null && !canceled)
{
    // Get the master form (recognize/classify)
    MasterForm master = new MasterForm();
    master.Properties = result.RecognitionResult.Properties;
    form.Master = master;

    // Populate our form with the recognition and processing results
    // and alignment information
    form.Result = result.RecognitionResult.Result;
    form.Alignment = result.RecognitionResult.Alignments;

    //We have successfully recognized and processed a form
}
}
```

After a form is successfully classified, we can parse the results and extract data from the fields defined by the template.

```

// Create a DataGridView control for displaying the table data
DataGridView _tableResults;
_tableResults = new DataGridView();

TableFormField tableField =
    form.ProcessingPages[pageIndex][fieldIndex] as TableFormField;
if (tableField.Result.Status == FormFieldStatus.Success)
{
    // Add the columns
    foreach (TableColumn column in tableField.Columns)
        _tableResults.Columns.Add(column.OcrField.Name, column.OcrField.Name);

    // Add the data, looping first by the rows, then for each field (cell)
    // within the row
    TableFormFieldResult results = tableField.Result as TableFormFieldResult;
    for (int i = 0; i < results.Rows.Count; i++)
    {
        TableFormRow row = results.Rows[i];
        _tableResults.Rows.Add();

        for (int j = 0; j < row.Fields.Count; j++)
        {
            // Get the text from the field!
            OcrFormField ocrField = row.Fields[j];
            TextFormFieldResult txtResults = ocrField.Result as TextFormFieldResult;
            _tableResults.Rows[i].Cells[j].Value = txtResults.Text;
        }
    }
}
}

```

Each field contains a plethora of information including the recognized data, confidence information, font characteristics, bounding rectangle and more. As you can see in Figure 3, the entire table is detected as a field, and is further broken down into each cell. LEADTOOLS' flexible and highly accurate processing routine is able to detect and account for variances in cell height, properly identifying the bounding rectangle of each cell and using optical character recognition to extract the data and detailed character information.

The screenshot displays the LEADTOOLS software interface for invoice recognition and processing. The main window shows a list of items with their quantities, descriptions, prices, and amounts. A detailed table results window is open, showing character recognition data. The shipping information and total invoice amount are also visible.

ITEM	QTY	DESCRIPTION	PRICE	AMOUNT
DOCIMG19	1	LEADTOOLS Document Imaging V19	2995.00	2995.00
AMDOCIMG19	1	Annual Maintenance LEADTOOLS Document Imaging V19	599.00	599.00
OCRADV19	1	LEADTOOLS OCR Module Advantage V19	995.00	995.00
AMOCRADV19	1	Annual Maintenance LEADTOOLS OCR Module Advantage V19	199.00	199.00

Character	Second Queue	Font Style	Font Size	Bounding Rectangle
D	0	Regular	42	D=42 1+3 Width=36 Height=42
C	0	Regular	42	C=42 1+3 Width=42 Height=42
I	0	Regular	42	I=42 1+3 Width=42 Height=42
M	0	Regular	42	M=42 1+3 Width=42 Height=42
G	0	Regular	42	G=42 1+3 Width=42 Height=42
L	0	Regular	42	L=42 1+3 Width=42 Height=42
B	0	Regular	42	B=42 1+3 Width=42 Height=42

Customer Information:
 CUSTOMER #: 66T3000
 INVOICE #: 0090152-IN
 DATE: 12/20/2014
 SHIP TO: John Doe, 13420 Alvin, Houston, TX 77058
 PHONE: (555) 555-5555
 TERMS: CREDIT CARD
 SHIP DATE: 12/20/2014
 SALES PERSON: ECOM
 DUE DATE: 12/20/2014

Inv Total (USD): \$4788.00

Figure 3: Displaying the Forms Recognition and Processing results

Want to see more? [Click here for a video demonstration of the LEADTOOLS Invoice Recognition and Processing SDK.](#)

Conclusion

Recognizing and processing data from unstructured forms such as invoices and bills is just one of many real-world solutions you can tackle with LEADTOOLS. Its state-of-the-art Invoice Recognition and Processing SDK makes it possible to create automated accounting, billing and invoicing solutions. Furthermore, LEADTOOLS provides a bountiful collection of document cleanup functions including deskew, line remove, despeckle, hole punch remove, etc., to filter out noise prior to recognition. LEADTOOLS offers an incredible value with its comprehensive family of toolkits for raster, document, medical and multimedia imaging. For more information on how LEAD Technologies can image-enable your application and boost your ROI, visit www.leadtools.com to download a free evaluation, or give us a call at +1-704-332-5532.

Sales: (704) 332-5532
sales@leadtools.com

Support: (704) 372-9681
support@leadtools.com



LEAD Technologies, Inc.
1927 South Tryon Street
Suite 200
Charlotte, NC 28203

About LEAD Technologies

With a rich history of nearly 25 years, LEAD has established itself as the world's leading provider of software development toolkits for document, medical, multimedia, raster and vector imaging. LEAD's flagship product, LEADTOOLS, holds the top position in every major country throughout the world and boasts a healthy, diverse customer base and strong list of corporate partners including some of the largest and most influential organizations from around the globe.

LEADTOOLS[®]
THE WORLD LEADER IN IMAGING SDKs

