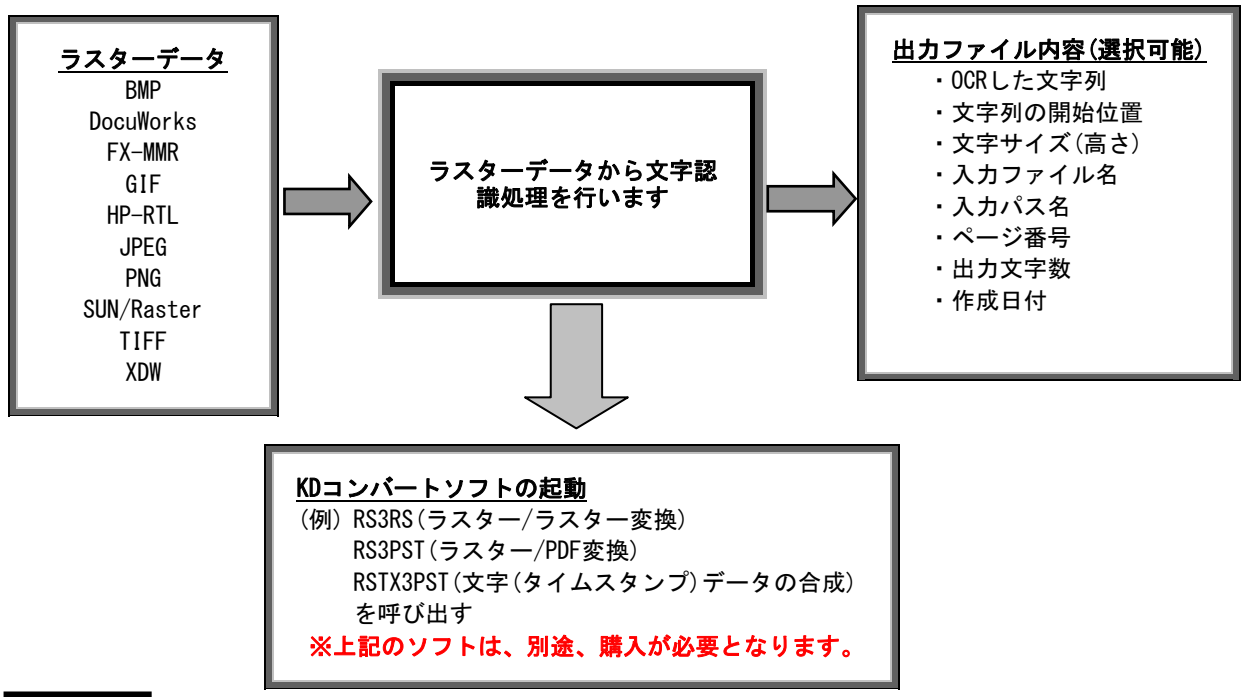
RS2OCR 簡易マニュアル

RS2OCRは、イメージデータからOCR(文字認識)処理を行ない、テキスト文書(TEXT,RTF,CSV形式等)へ変換を行なうソフトウエアです。
さらに、OCR認識により取得した情報を利用し、他のモジュール(別売)を呼び出し、入力データを変換することも可能です。

KCS Kernel Computer System
カーネルコンピュータシステム株式会社

本社：パッケージ販売部
〒221-0056
横浜市神奈川区金港町6-3 横浜金港町ビル
Tel：045-442-0500　　Fax：045-442-0501
URL:http://www.kernelcomputer.co.jp/



特 徴

- 非常に多くのラスターデータ(モノクロ、カラー)の入力をサポート
  DocuWorksデータも可能ですが、別途、DocuWorksが必要となります。
- 文字の認識を行なう領域の指定が可能(全景も可能)。最大サイズはA3サイズ。
- 入力がマルチページファイルの場合、処理するページを指定する事が出来ます。
- 認識した文字列毎の開始位置や文字の大きさ(高さ)を出力する事が出来ます。
- 認識した文字数を出力する事が出来ます。
- ページ番号、作成日付や入力ファイル名などを出力する事が出来ます。
- 認識したイメージ情報を利用し、他のKDコンバートを読み出し、入力データを変換する事も可能です。
  他の弊社製品と連携し、OCRした内容に応じてファイルの変換処理などを行う事が可能です。
  例えば、OCRで認識されたテキストからキーワードを検索し、検索したキーワードを出力ファイル名として変換します。
  又は文字列の方向によって、入力データを正立して、新しいファイルを作成します。
- 認識された文字データはテキスト形式、HTML形式、RTF形式、XML形式に出力できます。
- A3サイズを超える大きいサイズの図面はクリッピングを行い、A3サイズに収まるサイズにすることで、OCR処理する事が出来ます。

操作説明

変換を実行するには、次のような形式でコマンドラインに入力します。

RS2OCR［イメージファイル名］-O[出力テキストファイル名］-[各オプション]

操作例１(ファイル「sample.tif」の3ページ目をOCRして、出力ファイル名「sample.txt」で出力)

```
A>RS2OCR sample.tif -Osample.txt -Zocr_out.atr -N3
Raster(sample.tif) -> TEXT(sample.txt)ファイル　コンバータ

ただいま，ファイル変換中です。


RS2OCR 変換終了
```

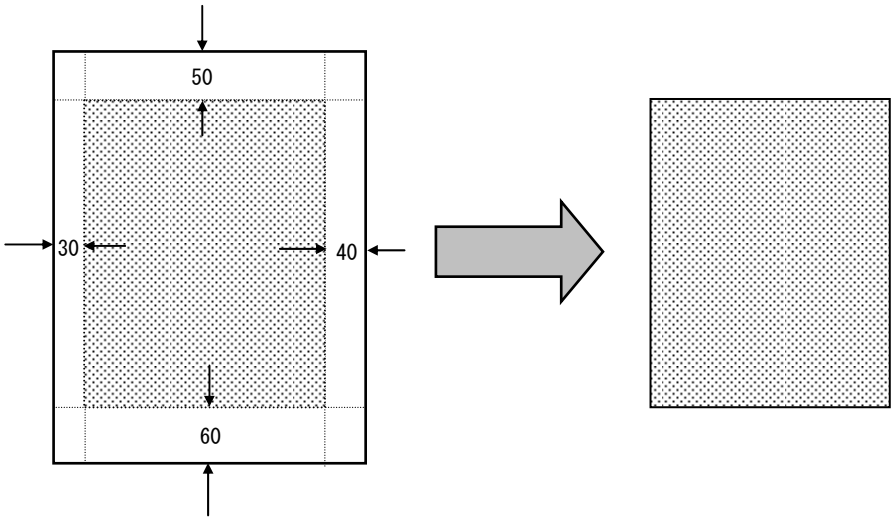
（注） 入力ファイル指定なしの場合標準入力となります。
出力ファイル指定なし(-Oオプションなし)の場合標準出力となります。

変換オプション

RS2OCRには、変換時に指定できる様々なオプションが用意されています。

変換オプションの概要：

- 出力ファイル名や入出力属性ファイル名を指定します。
- OCR処理のテキストを出力する時、前のファイルに重ねて(後に追加)出力する事が出来ます。
- 入力がマルチページファイルの場合、認識結果はページ毎にファイルで出力する事が出来ます。
- 入力がマルチページファイルの場合、処理ページを指定します。
- OCRの処理を行う際の領域(エリア)を定義するテキスト形式のファイルを指定します。
  (テキストファイルによって指定可能な領域は512個までとなります)
- イメージをクリッピングします。OCR処理対象の大きさはA3までしか出来ませんので、A3を越えた図面に対して、クリッピングで図面をA3以下に切り出すと、OCR処理が可能になります。
  (例) 左30ピクセル、右40ピクセル、上50ピクセル、下60ピクセルを切り落とす場合



- ログファイルや履歴ファイルの出力を指定します。
- ログファイルや履歴ファイルの形式はテキストやXMLを指定する事が出来ます。
- バッチファイルで指定されたファイル又はディレクトリ内のファイルを一括処理します

属性パラメータの設定方法

RS2OCRの変換における様々なパラメータを設定する入出属性パラメータファイルが用意されています。
これらの変換パラメータ内容を変更する事で種々な変換が可能となります。
属性パラメータファイルで入出力データフォーマット、OCRの処理条件、認識結果の出力書式や
他のモジュールの呼び出し方法などを指定する事が出来ます。

入力属性パラメータファイルの概要：

- 入力データのフォーマットを指定します。
  入力フォーマットとしては、TIFF(非圧縮,Packbits,JPEG,CCITT-1D,MMR,MR,MH)， JPEG、BMP や PNG など
  多数の種類が指定できます。指定しない場合は自動判定も可能です。
- 入力データのヘッダ情報を与える
  通常、ラスターデータの幅高さなどの情報は、フォーマット毎に規定された形式のヘッダとして
  ラスターデータに付加されています。しかし一部のフォーマットではこのようなヘッダ情報がないので、
  外部から情報を与える必要があります。
- 入力データの色を反転する事が可能です。

出力属性パラメータファイルの概要：

この属性ファイルの内容に従ってOCR処理及び出力を行います。

- 出力形式のフォーマットを指定します。
  出力フォーマットは TEXT, RTF, HTMLやXMLが指定できます。
  - 入力データの全頁をOCRする時、認識されたテキストはOCRするブロック順で出力されます。
    (OCR処理は画面を自動的にブロックに分割して処理しています)
  - テキストフォーマットの出力サンプルは5ページの「テキストサンプル」を参照してください。
  - RTF(リッチテキスト)フォーマットの出力サンプルは5ページの「RTF(リッチテキスト)出力サンプル」を参照してください。
- 出力文字フォントの設定 (出力フォーマットがRTFの場合のみ有効)
  指定可能の文字フォントは以下の種類が有ります。
  ＨＧｺﾞｼｯｸE-PRO, HG正楷書体-PRO, HG丸ｺﾞｼｯｸM-PRO,
  ＭＳ 明朝, ＭＳ Ｐ明朝, ＭＳ ゴシック, ＭＳ Ｐゴシック
- 出力文字サイズの設定 (出力フォーマットがRTFの場合のみ有効)
- 出力文字色の設定 (出力フォーマットがRTFの場合のみ有効)
- カラーテーブル(RGB)の設定 (出力フォーマットがRTFの場合のみ有効)
- 読み込んだイメージのノイズ除去強度を設定します。
- 文字に罫線が接触している場合、その罫線を削除するかどうかを設定します。
- 認識言語を設定します。
- 日本語の知識処理を行うかどうかを設定します。(認識言語が日本語の場合のみ有効)
- 出力するOCR処理情報の書式を指定することが可能です。
  検出したテキスト、文字列の開始位置(X,Y)、文字高さや文書情報などを出力する事が可能です。

(例1) 検出したテキストのままで出力

> 検出したテキストの情報は、1行につき1件ずつ、この書式にしたがって出力されます。
> 書式指定の中の「%」で始まるシーケンスは、表1にしたがって、それぞれ対応する情報に
> 置き換えられます。それ以外の文字はそのまま出力されます。

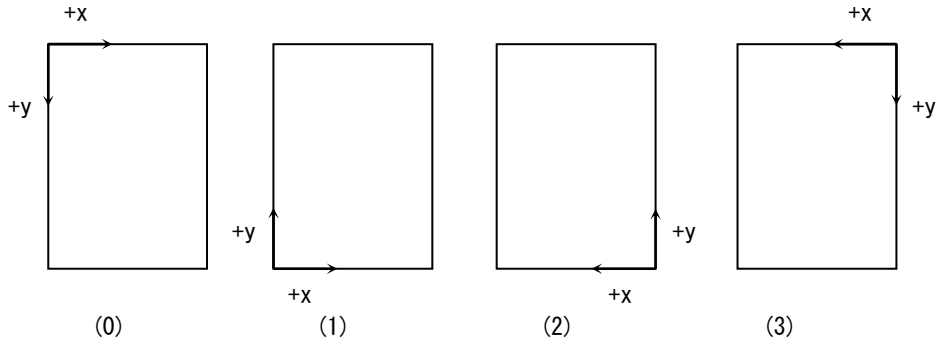
(例2) 文字列の開始位置(X,Y)、文字の高さ、文字列の順で出力します。

```
20.19,23.69,3.37,″ OCR(文字認識)概要″
27.31,32.38,3.24,″イメージの文字は、そのままでは選択したり、コピーしたりできません。
                    (略)
```

(例3) 文書情報はページの最後に付けます。

```
20.19,23.69,3.37,″ OCR(文字認識)概要″
27.31,32.38,3.24,″イメージの文字は、そのままでは選択したり、コピーしたりできません。
                    (略)
入力ファイル名:rs2ocr
入力パス名:C:¥exec¥rs2ocr.tif
ページ番号:1
文字数:876
作成日付:2009/01/16 19:09:14
```

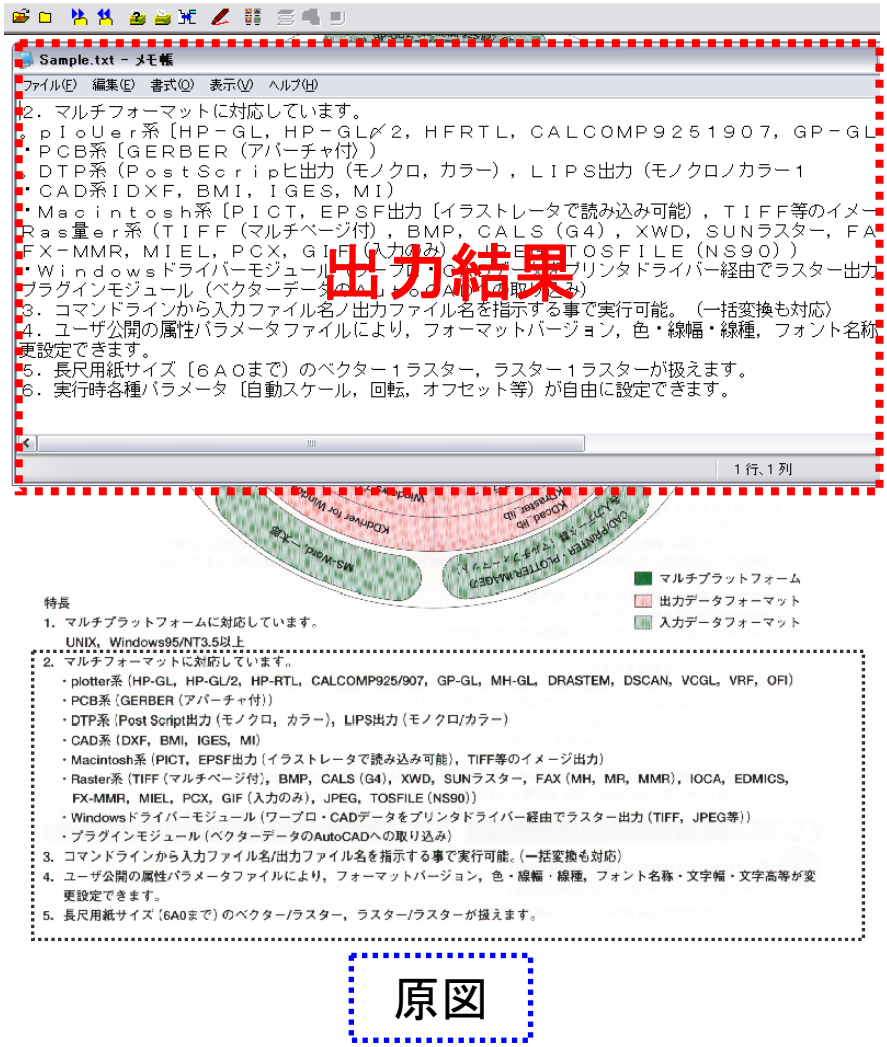
・ 検出したいキーワードを指定します。
  指定したキーワードを検出し、そのキーワードをファイル名の一部とし、
  他のモジュールを呼び出して入力データを変換して出力する事が出来ます。



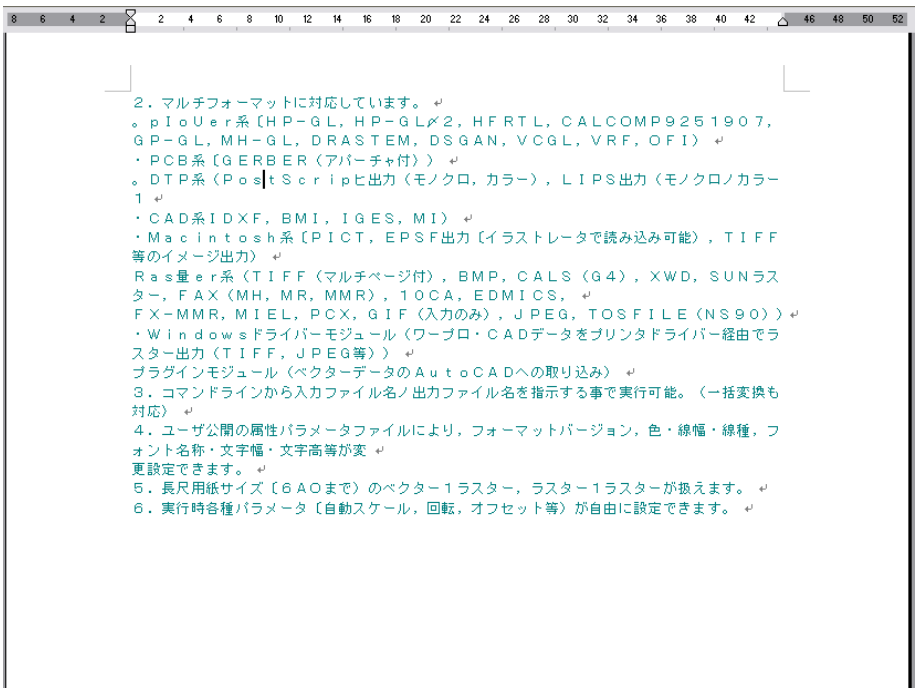
・ 座標原点位置を指定します。抽出される座標値はこの原点位置からの値となります。
・ 検出したいキーワードの検索範囲を設定します。
・ 検出したテキスト情報を利用して他のモジュールを呼び出す事が可能です。
  RS2OCRで検出したテキストの情報を使って、他のモジュールを呼び出す事が出来ます。
  この機能を利用すれば、他の製品と連携して、テキストの内容に応じてファイルの変換処理などを
  行う事が可能です。
・ 他のモジュールを呼び出すタイミングの指定が可能です。
・ OCR処理範囲の設定が可能です。
・ 認識させたい文字の種類を指定することが可能です。

出力サンプル

テキストサンプル(原図の下の枠をOCRする領域を指定し、テキストに出力したサンプルです。)



RTF(リッチテキスト)出力サンプル



制限事項

・ 動作環境
Windows XP (SP3以降)(32bit版)
Windows Vista (32bit版)
Windows 7 (32/64bit)
Windows 8 (32/64bit)
Windows 8.1 (32/64bit)
Windows 10 (32/64bit)

Windows Server 2003 (SP1以降)(32bit版)
Windows Server 2008 (32bit版)
Windows Server 2008 R2
Windows Server 2012
Windows Server 2012 R2
Windows Server 2016

・ 入力データの最大サイズはA3サイズとなります。
入力データがA3サイズを超える場合には、領域(A3サイズ以内)を指定する必要があります。

・ デフォルトの学習文字ファイル(MdtOcr.upt)が認識辞書のあるフォルダに存在する場合、削除してください。

価格

RS2OCR　　:　30万円(税抜き)