

MBS REALbasic OCR Plugin Documentation

Christian Schmitz

January 17, 2011

0.1 Introduction

This is the PDF version of the documentation for the REALbasic Plug-in from Monkeybread Software Germany. Plugin part: MBS REALbasic OCR Plugin

0.2 Content

- 1 List of all topics 3
- 2 All items in this plugin 5
- 3 List of all classes 19

Chapter 1

List of Topics

• 2 OCR	5
– 2.1 class OCRCharacterMBS	5
* 2.1.1 Blanks as Integer	6
* 2.1.1 Bottom as Integer	6
* 2.1.1 CharCode as Integer	6
* 2.1.1 Confidence as Integer	7
* 2.1.1 FontIndex as Integer	7
* 2.1.1 Formatting as Integer	7
* 2.1.1 Height as Integer	7
* 2.1.1 Left as Integer	8
* 2.1.1 PointSize as Integer	8
* 2.1.1 Right as Integer	8
* 2.1.1 Top as Integer	9
* 2.1.1 Width as Integer	9
– 2.2 class OCRBlockMBS	9
* 2.2.1 Character(index as integer) as OCRCharacterMBS	10
* 2.2.2 Count as Integer	10
* 2.2.2 EndTime as UInt32	10
* 2.2.2 ErrorCode as Integer	10
* 2.2.2 MoreToCome as Boolean	11
* 2.2.2 OCRAlive as Boolean	11
* 2.2.2 Progress as Integer	11
* 2.2.2 Text as String	11
– 2.3 class OCRMBS	12
* 2.3.1 BeginPage(pic as picture) as boolean	13

* 2.3.1 BeginPage(width as UInt32, height as UInt32, data as memoryblock, bpp as integer) as boolean	13
* 2.3.1 BeginPageUpright(width as UInt32, height as UInt32, data as memoryblock, bpp as integer) as boolean	14
* 2.3.1 Constructor(lang as string, path as folderitem)	15
* 2.3.1 EndPage	15
* 2.3.1 RecognizeABlock(left as UInt32, right as UInt32, Top as UInt32, Bottom as UInt32) as OCRBlockMBS	16
* 2.3.1 RecognizeAllWords as OCRBlockMBS	16
* 2.3.2 Handle as Integer	16

Chapter 2

OCR

2.1 class OCRCharacterMBS

class OCRCharacterMBS

class, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: Description of a single character.

Notes:

The character code is defined by the character set of the current font.

Output text is sent as an array of these structures.

Spaces and line endings in the output are represented in the structures of the surrounding characters. They are not directly represented as characters.

The first character in a word has a positive value of blanks.

Missing information should be set to the defaults in the comments.

If word bounds are known, but not character bounds, then the top and bottom of each character should be those of the word. The left of the first and right of the last char in each word should be set. All other lefts and rights should be set to -1.

If set, the values of right and bottom are left+width and top+height.

Most of the members come directly from the parameters to ocr_ append_ char.

The formatting member uses the enhancement parameter and combines the line direction stuff into the top 3 bits.

The coding is 0=RL char, 1=LR char, 2=DR NL, 3=UL NL, 4=DR Para, 5=UL Para, 6=TB char, 7=BT char. API users do not need to know what the coding is, only that it is backwards compatible with the previous version.

It should be noted that the format for char code for version 2.0 and beyond is UTF8 which means that ASCII characters will come out as one structure but other characters will be returned in two or more instances of this structure with a single byte of the UTF8 code in each, but each will have the same bounding box. Programs which want to handle languages with different characters sets will need to handle extended characters appropriately, but **all** code needs to be prepared to receive UTF8 coded characters for characters such as bullet and fancy quotes.

2.1.1 Properties

Blanks as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The number of spaces before this character.

Notes: (Read and Write property)

Bottom as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The bottom position of the character.

Notes:

Value is -1 if unknown.

(Read and Write property)

CharCode as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The char code of the character itself.

Notes: (Read and Write property)

Confidence as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The confidence for this character.

Notes:

Range from 0 to 100.

0 is perfect, 100 is reject.

(Read and Write property)

FontIndex as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The font index.

Notes: (Read and Write property)

Formatting as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The formatting information for this character.

Notes: (Read and Write property)

Height as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The height of the character.

Notes: (Read and Write property)

Left as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The left position of the character.

Notes:

Value is -1 if unknown.

(Read and Write property)

PointSize as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The font size.

Notes:

72 is one inch.

(Read and Write property)

Right as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The right position of the character.

Notes:

Value is -1 if unknown.

(Read and Write property)

Top as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The top position of the character.

Notes:

Value is -1 if unknown.

(Read and Write property)

Width as Integer

property from class OCRCharacterMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The width of the character.

Notes: (Read and Write property)

2.2 class OCRBlockMBS

class OCRBlockMBS

class, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: Description of the output of the OCR engine.

Notes:

This class is used as both a progress monitor and the final output header, since it needs to be a valid progress monitor while the OCR engine is storing its output to shared memory.

During progress, all the buffer info is -1.

Progress starts at 0 and increases to 100 during OCR. No other constraint.

2.2.1 Methods

Character(index as integer) as OCRCharacterMBS

method from class OCRBlockMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: Returns the character with the given index.

Notes: Returns nil on any error.

2.2.2 Properties

Count as Integer

property from class OCRBlockMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The number of characters found.

Notes: (Read and Write property)

EndTime as UInt32

property from class OCRBlockMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: This is the time to stop if not zero.

Notes: (Read and Write property)

ErrorCode as Integer

property from class OCRBlockMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The last error code.

Notes: (Read and Write property)

MoreToCome as Boolean

property from class OCRBlockMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: True if more data is to be reported.

Notes: (Read and Write property)

OCRAlive as Boolean

property from class OCRBlockMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: Set to one by the OCR engine to report that it is still running.

Notes: (Read and Write property)

Progress as Integer

property from class OCRBlockMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The progress value.

Notes:

Increases from 0 to 100.

(Read and Write property)

Text as String

property from class OCRBlockMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The text recognized.

Notes:

This is a plugin convenience function which builds a string from all the character objects.

Spaces are added as needed and UTF8 characters are collected correctly. If character is more left than character before, a new line character is inserted.
(Read and Write property)

2.3 class OCRMBS

class OCRMBS

class, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: This is a class for optical text recognition based on the Tesseract OCR engine.

Example:

```
dim o as new OCRMBS("deu", GetFolderItem("tesdata"))
dim f as FolderItem = GetFolderItem("phototest.tif")
dim p as Picture = f.OpenAsPicture
```

```
if o.BeginPage(p) then
dim t as OCRBlockMBS = o.RecognizeAllWords
```

```
MsgBox t.text
end if
```

```
Exception e as OCRFatalErrorExceptionMBS
MsgBox "You have the tesdata folder in the project folder?"
```

Notes:

You can find the Tesseract OCR engine on this website: <http://code.google.com/p/tesseract-ocr/>

The Tesseract OCR engine was one of the top 3 engines in the 1995 UNLV Accuracy test. Between 1995 and 2006 it had little work done on it, but it is probably one of the most accurate open source OCR engines available. The source code will read a binary, grey or color image and output text.

In order to use it you need the language files which you can find on the MBS and the google website above.

2.3.1 Methods

BeginPage(pic as picture) as boolean

method from class OCRMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe in REAL Studio 2010r3 or newer, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: Starts text recognition on a picture.

Example:

```
dim o as new OCRMBS("deu", GetFolderItem("tessdata"))
dim f as FolderItem = GetFolderItem("phototest.tif")
dim p as Picture = f.OpenAsPicture

if o.BeginPage(p) then
dim t as OCRBlockMBS = o.RecognizeAllWords

MsgBox t.text
end if
```

Exception e as OCRFatalErrorExceptionMBS
MsgBox "You have the tessdata folder in the project folder?"

Notes:

internally picture is converted to grayscale if needed.

Returns true on success and false on failure.

See also:

- 2.3.1 BeginPage(width as UInt32, height as UInt32, data as memoryblock, bpp as integer) as boolean
13

BeginPage(width as UInt32, height as UInt32, data as memoryblock, bpp as integer) as boolean

method from class OCRMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: Starts text recognition on a picture.

Notes:

BeginPage assumes the first memory address is the bottom of the image.

The memoryblock contains width*height pixels. bpp is the number of bits per pixel.

Returns true on success and false on failure.

See also:

- 2.3.1 BeginPage(pic as picture) as boolean

13

BeginPageUpright(width as UInt32, height as UInt32, data as memoryblock, bpp as integer) as boolean

method from class OCRMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: Starts text recognition on a picture.

Example:

```

dim f as FolderItem= GetFolderItem("eurotext.tif")
dim eurotext as Picture = f.OpenAsPicture

dim o as new OCRMBS("deu", GetFolderItem("tessdata"))
dim p as Picture = NewPicture(eurotext.Width, eurotext.Height, 32)

p.Graphics.DrawPicture eurotext,0,0

// We pass data in a memoryblock. 8 bit per pixel in grayscale from the Red channel

dim buf as MemoryBlock = NewMemoryBlock(p.Width*p.Height)
dim r as RGBSurface = p.RGBSurface
dim n,x,y as integer

for y=0 to p.Height-1
for x=0 to p.Width-1
dim c as color = r.Pixel(x,y)

buf.Byte(n)=c.Red
n=n+1
next
next

if o.BeginPageUpright(p.Width, p.Height, buf, 8) then
dim t as OCRBlockMBS = o.RecognizeAllWords

MsgBox t.text
end if

```

Exception `e` as `OCRFatalErrorExceptionMBS`
MsgBox "You have the tessdata folder in the project folder?"

Notes:

`BeginPageUpright` assumes the first memory address is the top of the image.
The memoryblock contains width*height pixels. `bpp` is the number of bits per pixel.

Returns true on success and false on failure.

Constructor(lang as string, path as folderitem)

method from class `OCRMBS`, `OCR`, `MBS REALbasic OCR Plugin (tesseract)`, Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The constructor.

Notes:

`lang` is the code of the language for which the data will be loaded.
(Codes follow ISO 639-3.) If it is "", english (eng) will be loaded.

`path`: the location of the tessdata folder. Note data this folder must be named tessdata.

EndPage

method from class `OCRMBS`, `OCR`, `MBS REALbasic OCR Plugin (tesseract)`, Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: Ends the current page.

RecognizeABlock(left as UInt32, right as UInt32, Top as UInt32, Bottom as UInt32) as OCRBlockMBS

method from class OCRMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: This function allows you to extract one word or section from the picture.

Notes:

Limited to 32000 characters.

Returns nil on any error.

RecognizeAllWords as OCRBlockMBS

method from class OCRMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: This function allows you to extract all text from the picture.

Notes:

Same as RecognizeABlock(0,0,0,0)

Limited to 32000 characters.

Returns nil on any error.

2.3.2 Properties**Handle as Integer**

property from class OCRMBS, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: The internal reference to the OCR engine.

Notes:

If this value is not zero, the constructor was successful.

(Read and Write property)

2.4 class OCRFatalErrorExceptionMBS

class OCRFatalErrorExceptionMBS

class, OCR, MBS REALbasic OCR Plugin (tesseract), Plugin version: 9.5, console safe, Mac OS X: Works, Windows: Works, Linux x86: Works.

Function: A fatal error occurred.

Example:

```
dim o as new OCRMBS("deu", GetFolderItem("tessdata"))
dim f as FolderItem = GetFolderItem("phototest.tif")
dim p as Picture = f.OpenAsPicture
```

```
if o.BeginPage(p) then
dim t as OCRBlockMBS = o.RecognizeAllWords
```

```
MsgBox t.text
end if
```

```
Exception e as OCRFatalErrorExceptionMBS
MsgBox "You have the tessdata folder in the project folder?"
```

Notes:

Typical error is that the data files can't be found or have the wrong format.
Subclass of the RuntimeException class.

Chapter 3

List of all classes

• OCRBlockMBS	9
• OCRCharacterMBS	5
• OCRFatalErrorExceptionMBS	17
• OCRMBS	12