sk-spell

podpora slovenčiny v Open Source programoch

Compiling Leptonica and Tesseract-ocr with Mingw+Msys

posledná zmena: 2. April 2012

back to tesseract-ocr-en

I plan to extend qt-box-editor with some additional features (e.g. generating boxes), but I would need to have tesseract and Leptonica as a library for Windows. Thanks to great job of Tom Powers there is already Leptonica library build by VC++ and there will be tesseract C++ library for version 3.02.

But it is not suggested using library created by other compiler. And my project use MinGW… I made some test to build tesseract with cmake+MinGW, but I plan to use Leptonica too. So I decided to compile Leptonica and tesseract with MinGW. Here is a short tutorial how to do it.

MinGW, MSYS and necessary packages

You need to have installed MinGW and MSYS. If you do not have it, then please follow great tutorial about compiling OpenTTD on MinGW.

Also try to install following packages (you should receive an error message if they are installed already):

```
mingw-get install msys-wget
mingw-get install msys-bzip2
mingw-get install msys-patch
mingw-get install msys-autoconf
mingw-get install msys-libtool
```

Create directory '/usr/src' in "MinGW Shell" and go there:

```
mkdir -p /usr/src
cd /usr/src
```

This will be our "build directory".

It is expected that you will download (source) package from individual site by yourself. Please be aware there could be available other (newer) versions of libraries/programs. If you download other version of library you will need to change names of packages/directories etc.

svn

If you do not have installed svn (e.g. TortoiseSVN), please install svn as described on wiki.openttd.org. If you do not plan to test current svn code of tesseract (you will need to download the latest package), you do not need svn.

pthreads

sourceware.org provides package of pthreads-win32.

```
tar xf pthreads-w32-2-8-0-release.tar.gz
cd pthreads-w32-2-8-0-release
export CPPFLAGS="-DPTW32_STATIC_LIB"
make clean GC-static
cp -iv pthread.h semaphore.h sched.h /mingw/include/
cp -iv libpthreadGC2.a /mingw/lib/libpthread.a
cd ..
```

Build Leptonica dependencies

Leptonica support several image format. Respective libraries has to be installed before Leptonica is configured end installed.

Build zlib

zlib is a free, general-purpose, legally unencumbered — that is, not covered by any patents — lossless data-compression library. It is need by libpng.

```
tar xf zlib-1.2.6.tar.gz
cd zlib-1.2.6
make -f win32/Makefile.gcc
```

Then change (line 33) 'SHARED_MODE=0' to 'SHARED_MODE=1' in "win32/Makefile.gcc" and run:

```
BINARY_PATH=/usr/local/bin \
   INCLUDE_PATH=/usr/local/include \
   LIBRARY_PATH=/usr/local/lib \
   make -f win32/Makefile.gcc install
cd ..
```

Build xz

xz provides lzma support needed by TIFF library.

```
tar xf xz-5.0.3.tar.bz2
cd xz-5.0.3
./configure
make -j 4 && make install
cd ..
```

Build libpng

libpng is an open, extensible image format with lossless compression.

```
tar xf libpng-1.5.9.tar.xz
cd libpng-1.5.9
./configure
```

```
make && make install
cd ..
```

Build giflib

giflib is a library for reading and writing gif images. It is API and ABI compatible with libungif which was in wide use while the LZW compression algorithm was patented.

```
tar xf giflib-4.1.6.tar.bz2
cd giflib-4.1.6
./autogen.sh
LDFLAGS="-no-undefined  -Wl,--as-needed" ./configure
make -j 4 && make install
cd ..
```

Build jpeg-8d

jpeg-8d is free library for JPEG image compression.

```
tar xf jpegsrc.v8d.tar.gz
cd jpeg-8d
./configure
make -j 4 && make install
cd ..
```

Build jbigkit

JBIG-KIT implements a highly effective data compression algorithm for bi-level high-resolution images such as fax pages or scanned documents. It can be used by TIFF library. This is optional package.

```
tar xf jbigkit-2.0.tar.gz
cd jbigkit
wget http://www.sk-spell.sk.cx/file_download/99/autotools_support.patch.gz
gzip -cd autotools_support.patch.gz | patch -p1
./autogen.sh
./configure
make && make install
cd ..
```

Build libtiff

libtiff provides support for the Tag Image File Format (TIFF), a widely used format for storing image data.

You need to use 3.9.5 version – 4.0.1 did not work with tesseract/leptonica (I need to do more testing why).

```
tar xf tiff-3.9.5.tar.gz
cd tiff-3.9.5
./autogen.sh
./configure
make -j 4 && make install
cd ..
```

Build webp

WebP is an image format that does lossy compression of digital photographic images.

```
tar xf libwebp-0.1.3.tar.gz
cd libwebp-0.1.3
./autogen.sh
LDFLAGS="-no-undefined  -Wl,--as-needed" \
   CPPFLAGS=-DQGLOBAL_H ./configure
make && make install
cd ..
```

Build Leptonica 1.68

```
tar xf leptonica-1.68.tar.gz
cd leptonica-1.68
./autobuild
./configure
```

For version 1.68 you need to patch (it should be fixed in next version):

```
wget "http://leptonica.googlecode.com/issues/attachment?aid=560001000&name=zlib-include.patch&token=say6dkQyRWJp2MvoOO1hmTqXAtU%3A1332684407152" -O zlib-include.patch
patch -p1 <zlib-include.patch
```

Then you can continue:

```
make -j 4 && make install
cd ..
```

Build tesseract-ocr

If you want to test recent code from svn then you need to fetch code from svn first:

```
svn checkout \
  https://tesseract-ocr.googlecode.com/svn/trunk/ \
  tesseract-ocr
```

WARNING: Because of number and size of language data svn repository is bigger than 624M! Alternatively you can download current snapshot of svn repository WITHOUT language data files and uncompress it as other packages.

Build process will consist:

```
cd tesseract-ocr
./autogen.sh
LDFLAGS="-no-undefined  -Wl,--as-needed" \
  ./configure --disable-tessdata-prefix
make -j 4 && make install
```

Option `'--disable-tessdata-prefix'` will prevent that "TESSDATA-PREFIX" is set to installation directory (usually "/usr/local/share" or "/usr/share") and built-in. With this option it is expected to have "tessdata" directory at the same place where is executable (or library) – if environment variable "TESSDATA_PREFIX" is not set.

Maybe you can use 'CPPFLAG="-DNDEBUG"' before './configure' for release version:

```
LDFLAGS="-no-undefined  -Wl,--as-needed" \
  CPPFLAG="-DNDEBUG" ./configure \
  --disable-tessdata-prefix
```

If you get tesseract from svn, you can install all language files with:

```
make install LANGS=
```

If you want to install just English, German and Spain language files then run:

```
make install LANGS="spa eng deu"
```

If you compiled tesseract from package, then you need to download and install (uncompress and copy to tessdata directory) language files manually.

Do not mix different versions of language data!!!. E.g. you cannot use 2.0x language files with tesseract 3.0x. You can not use higher version of language file in lower version tesseract (e.g. 3.02 language file in tesseract 3.01). But you can use 3.01 language file in tesseract 3.02.

And final info:

```
$ tesseract -v
tesseract 3.02
 leptonica-1.68
  libgif 4.1.6 : libjpeg 8d : libpng 1.5.9 : libtiff 3.9.5 : zlib 1.2.6
```

Webp is not listed there, but it is supported:

```
 convert eurotext.tif eurotext.png
 cwebp -q 100 eurotext.png -o eurotext.webp
 tesseract.exe eurotext.webp eurotext-webp
```

back to tesseract-ocr-en