Langfassung des Beitrags für die 11. KSFE

Abstract + Kommentierte Macros am Ende

Titel
Automatische Texterkennung (OCR) in Ultraschallbildern der A. carotis - SAS & Open Source Software im Team

Namen aller Autoren *(Referent bitte unterstreichen)*
Dietrich Alte, André Werner

Einrichtung(en) und Ort(e)
Institut für Epidemiologie & Sozialmedizin an der Medizinische Fakultät der EMA-Universität Greifswald

eMail Adresse(n)
alte@uni-greifswald.de

Kurze Beschreibung des Inhalts *(½ A4 < Umfang < 1 A4; bei Times New Roman 12pt)*

Einleitung / Hintergrund
Der Einsatz bildgebender Verfahren in epidemiologischen Studien erfordert mitunter die Extraktion von Text-Informationen aus z.B. Bildschirmkopien eines Ultraschallgeräts. Die Informationen sind bei einfachen Bildformaten wie BMP, TIF, JPG nicht als Text zum Bild gespeichert (wie bei DICOM), sondern befinden sich als Pixelstruktur im Bild. Bei großen Studien kann die Anzahl dieser Bilder in die Tausende gehen, was keine manuelle Extraktion der Daten erlaubt. Ein ähnliches Problem stellte sich in SHIP (Study of Health in Pomerania).

Material und Methoden
Von 4310 Probanden in SHIP-0 (davon bei ~2500 Sonographie Aa. carotis) und > 3000 Probanden in SHIP-1 liegen bis zu fünf Aufnahmen der A. Carotis vor (~45.000 TIF Bilder, 736x556 Pixel). Bei der Vermessung zur Bestimmung der Intima-Media-Dicke (IMT) ist der am Ultraschallgerät eingestellte Zoom-Faktor für die Kalibrierung relevant. Die Bilder liegen in >5000 Verzeichnissen nach Probandennummer sortiert vor. Zur Automatisierung wurde ein SAS-Macro entwickelt, welches die Verzeichnisinformationen einliest, alle Bilder in diesen Verzeichnissen durchläuft und dabei OpenSource Programme aufruft: XNVIEW (1) schneidet den relevanten Teil des Bildes aus und konvertiert ihn ins PCX-Format. GOCR (2) erkennt daraus per Texterkennung/Optical Character Recognition (OCR) die relevanten Informationen und schreibt diese in eine Textdatei, die SAS dann einliest und in einer SAS-Datei speichert und weiterverarbeitet.

Ergebnisse
In einem Testlauf gelang die Text-Erfassung trotz geringer Bildauflösung in 98% der Fälle (der Bildausschnitt hat 170x50 Pixel für eine Textzeile mit ~10 Zeichen). Verarbeitungszeit der Bilder: <5min je 1000 Bilder. Nichterkennen von Bildern kommt teilweise durch mangelnde Bildqualität zustande. Manuelle Nacherfassung/Prüfung muss erfolgen.

Diskussion / Schlussfolgerungen
Die Methode ist flexibel auf Probleme der Texterkennung in Ultraschall- oder anderen

Bildern anwendbar. Die geringe Rate nicht erkannter Bilder erlaubt einen Einsatz auch bei sehr großen Bildmengen. Die Erkennungsrate sollte bei höherer Auflösung steigen. Eine Umsetzung in anderen Sprachen wie R (3) oder perl (4) ist möglich.

Literatur

(1) www.xnview.com
(2) jocr.sourceforge.net
(3) www.r-project.org
(4) www.perl.org

Postanschrift des Erstautors bzw. Einreichenden

Dr. Dietrich Alte
Institut für Epidemiologie & Sozialmedizin
Walther-Rathenau-Str. 48
D-17487 Greifswald

Art des Beitrags:

☐ Vortrag (30 Min)
☐ Tutorium mit ☐ 30 Min., ☐ 60 Min., ☐ 90 Min.
☐ Softwaredemonstration
☒ **Poster**

Benötigte technische Ausstattung:

☐ Beamer
☐ PC (Win XP, Office 2003, SAS 8.2, SAS 9)
☐ Overheadprojektor
☐ Nutzung eines eigenen Notebooks

Thema zum Schwerpunkt:

☒ **Anwendungen in Biometrie, Statistik und Informatik**
☐ Data Mining, Web Mining, Text Mining
☐ Anwendungen im Gesundheitswesen
☒ **Datamanagement und Data Warehousing**
☐ Anwendungen in der pharmazeutischen Forschung
☐ Ausbildung mit und in SAS
☒ **Tipps & Tricks**
☐ Freies Thema

Zielpublikum

☒ **SAS-Programmierer**
☒ **Wissenschaftler/Statistiker**
☐ SAS-Anwender
☐ Sonstige und zwar _____

Erforderliche SAS-Kenntnisse

☐ SAS-Grundlagen
☒ **gute SAS-, Makro- und Programmierkenntnisse**

Folgende SAS-Module sind Thema: **SAS/BASE**

Die unten folgenden **Macros** wurden für die Berechnungen verwendet:

```
/*-------------------------------------------------------------------------*/
/* SUBDIRINFO.SAS - Macro zum Auslesen von Unterverzeichnissen             */
/*-------------------------------------------------------------------------*/
/* Author:   D. Alte                                                       */
/* Date:     2002/10/16                                                    */
/*-------------------------------------------------------------------------*/
/* UPDATE LOG                                                              */
/*-------------------------------------------------------------------------*/
/* 20.12.2004 D.Alte                                                       */
/* - mit v9.1.3 neu kompiliert, da sonst notes im log                      */
/* 28.04.2005 D.Alte                                                       */
/* - Umstellung auf Windows XP, Anpassung der Spalten                      */
/*-------------------------------------------------------------------------*/
%macro SubDirInfo(Path, out=work.dirinfo) /store;
filename _temp pipe "dir %bquote(&path) /T:C /-C /4";
   data &out;
      infile _temp lrecl=200 missover pad;
      length bigline $200;
      input bigline $200.;
      name = substr(bigline,37,60);
      date = input (substr(bigline,1,10), ? ddmmyy10.);
      if index (bigline, "<DIR>") > 0;
      if name not in (".", "..");
      format date ddmmyy10.;
      drop bigline;
   run;
   filename _temp;
%mend;

/*-------------------------------------------------------------------------*/
/* DIRINFO.SAS - Macro zum Auslesen der Dateien in einem Verzeichnis       */
```

```
/*-------------------------------------------------------------------*/
/* Author:   D. Alte                                                 */
/* Date:     2002/10/16                                              */
/*-------------------------------------------------------------------*/
/* UPDATE LOG                                                        */
/*-------------------------------------------------------------------*/
/* 20.12.2004 D.Alte                                                 */
/* - mit v9.1.3 neu kompiliert, da sonst notes im log                */
/* 28.04.2005 D.Alte                                                 */
/* - Umstellung auf Windows XP, Anpassung der Spalten                */
/*-------------------------------------------------------------------*/
%macro DirInfo(Path, out=work.dirinfo) /store;
filename _temp pipe "dir %bquote(&path) /T:C /-C /4";
   data &out;
      infile _temp lrecl=200 missover pad;
      length bigline $200;
      input bigline $200.;
      name = substr(bigline,37,60);
      date = input (substr(bigline,1,10), ? ddmmyy10.);
      size = input (substr(bigline,27,9), ? 14.);
      if date > 0 and size GE 0;
      format date ddmmyy10.;
      drop bigline;
   run;
   filename _temp;
%mend;

/*************************************************************************/
/*** OCR.SAS - Macro für OCR Erkennung des Zoom-Faktors imm CCA-Reading    ***/
/*************************************************************************/
/* Language:                 SAS (9.1.3 SP3)                             */
/* Date Implemented:         2006/01/30                                  */
/* Program Version:          1.01                                        */
/* Program Author:           Dietrich Alte                               */
/*************************************************************************/
%macro OCR (path, out=OCR, out_miss=OCR_miss);
   options noxwait xsync xmin;
   filename gocr "D:\TEMP\gocr_bat.bat";
   filename dummy "C:\";

* Liste der Verzeichnisse erstellen;
%SubDirInfo(&path, out=work.dirlist);

* Anzahl der Unterverzeichnisse in Macrovariable schreiben;
proc sql; select count (distinct name) into :n_subdir from dirlist; quit;

* Schleife über alle Verzeichnisse;
%do I = 1 %to 2 /*&n_subdir*/;

   * Unterverzeichnis auswählen;
   data _null_; set dirlist; if _n_ = &I then do;
        call symput ('subdir', "&path\"||strip(name)||"\"); end; run;

   * Inhalt dieses Unterverzeichnisses auslesen;
   %DirInfo(&subdir, out=work.dir);

   * Konvertierung in PCX und Schreiben der Batch-Datei für OCR;
   data dir;
   set dir;
   if _n_=1 then RC_convert = system ("C:\Programme\Xnview\NConvert.exe -crop 500 122 170 50 -out pcx -
o &subdir.%.pcx &subdir.*.TIF");
   file gocr;
   postfix = index (upcase(name), ".TIF");
   if postfix then do;
      file  = substr(name, 1, postfix-1);
      command = compbl("C:\Software\GOCR\gocr -l 0 &subdir"|| strip(file) ||
        ".pcx -o &subdir"|| strip(file) || ".txt");
      textfile = compbl("&subdir"||strip(file)||".txt");
      put command;
      end;
   else delete;
   run;

   * OCR Batch laufen lassen + Bilder wieder löschen;
   data _null_;
      call system ("D:\TEMP\gocr_bat.bat");
```

```
      *call system ("del &subdir.*.pcx");
   run;

   * Einlesen der per OCR erkannten Textschnipsel;
   data ocr ocr_missing;
   set dir (keep = textfile);
   infile dummy filevar=textfile end=done;
   do while (not done);
      input line & $50.;
      line = compress(compress(line), "_0", "kN");
      Quelle = textfile;
      if indexc(UPCASE(line), "1234567890X") then output OCR;
      else output OCR_missing;
   end;
   run;

   * Textschnipsel wieder löschen;
   data _null_; call system ("del &subdir.*.txt"); run;

   * Liste der erkannten Texte anhängen an Gesamtdatei;
   Proc Append base=&out data=OCR; run;
   Proc Append base=&out_miss data=OCR_missing; run;
%end;
%MEND;
```