Extraktion von Informationen aus binären Dateien: Ein SAS-Makro zur Erfassung von strukturierten Meta-Informationen aus Bilddateien aktueller Digitalkameras

Thomas G. Grobe
Institut für Sozialmedizin, Epidemiologie
und Gesundheitssystemforschung
ISEG
30159 Hannover, Lavesstr. 80
grobe@iseg.org

Zusammenfassung

Projektbezogen kann die automatisierte Erfassung von Meta-Informationen zu Dateien wünschenswert sein. Hierzu bietet SAS über Optionen zum Lesen nahezu beliebiger binärer Daten sowie die Makroprogrammierung sehr flexible Möglichkeiten. Diese sollen an einem konkreten Beispiel, der Extraktion von Meta-Informationen aus JPEG-komprimierten Bildern moderner Digitalkameras, demonstriert werden.
Da ein entsprechendes Vorgehen bei einer großen Zahl unterschiedlicher Dateitypen prinzipiell realisierbar sein dürfte, kann die Programmierung eines entsprechend angepassten SAS-Makros immer erwogen werden, sofern Informationen aus Meta-Daten a) relevant erscheinen, b) ausreichende Informationen zu deren Struktur vorliegen und c) eine Erfassung nicht nur sporadisch vorgesehen ist. Über entsprechende Makro-Programme können so Informationen ohne Rückgriff auf externe Programme in SAS-Datendateien verfügbar gemacht werden, die über vorkonfektionierte Importfilter primär nicht zugänglich sind.

Keywords: Datenaufarbeitung, binäre Daten, SAS-Makro-Programmierung, JPEG, EXIF 2.1,

1 Hintergrund

Im Rahmen einer projektbezogenen Datenerfassung können Informationen in den unterschiedlichsten (Datei-) Formaten anfallen, wobei nicht alle dieser Informationen über ein und dieselbe Programmumgebung primär bzw. unter Verwendung vorkonfektionierter Importfilter zugänglich sind. Die Erfassung von Informationen unter einer einheitlichen Programmumgebung erscheint für automatisierte Erfassungsvorgänge jedoch oftmals erstrebenswert. SAS bietet hierfür recht flexible Möglichkeiten, die weit über einfache Einlesevorgänge strukturierter ASCII-Daten hinausgehen und Zugriffe auch auf Dateien in nahezu beliebigen binären Kodierungen erlauben.
Entsprechende Möglichkeiten zur Extraktion von Informationen aus binären Dateien sollen an dieser Stelle beispielhaft an der Extraktion von Meta-Informa-

tionen aus Bilddateien moderner Digitalkameras im JPEG-Format demonstriert werden, die in einem standardisierten Datei-Header umfangreiche bildspezifische Informationen zur Aufnahmetechnik sowie zum Aufnahmezeitpunkt und in Sonderfällen auch zum Aufnahmeort beinhalten können. Diese Informationen werden im Rahmen des vorgestellten SAS-Makros zum automatisierten Aufbau einer Bilddatenbank in einer SAS-Datendatei erfasst.

2 Meta-Informationen in Bilddateien von Digitalkameras

Aktuelle Digitalkameras bieten zum Teil unterschiedliche Formate zur Speicherung von Bildern an. Von nahezu allen Kameras wird dabei eine Speicherung im JPEG-Format, zumeist in unterschiedlichen Qualitätsstufen bezüglich Auflösung und Kompression, angeboten. Abbildung 1 zeigt ein unbearbeitet Originalbild einer Digitalkamera.

Abbildung 1: Originalaufnahme einer Digitalkamera

Das JPEG-Format besitzt eine festgelegte Struktur, wobei bestimmte Bildinformationen in einem Header zu Beginn der Datei und vor den eigentlichen bildgebenden Daten abgelegt werden. Im JPEG-Header können zusätzlich Bereiche für Informationen definiert werden, die Elemente außerhalb der eigentlichen JPEG-Formatspezifikation enthalten. Zur Ablage erweiterter Aufnahmeinformationen wird dabei von aktuellen Digitalkameras nahezu ausnahmslos das EXIF 2.1-Format verwendet. Genauere Formatspezifikationen sind (1) zu entnehmen.

Bestimmte Informationselemente sind bereits bei einer Betrachtung der Bilddatei in einem Hexadezimaleditor ohne weitere Kenntnis der Spezifikation erkennbar, wie der Screenshot in Abbildung 2.

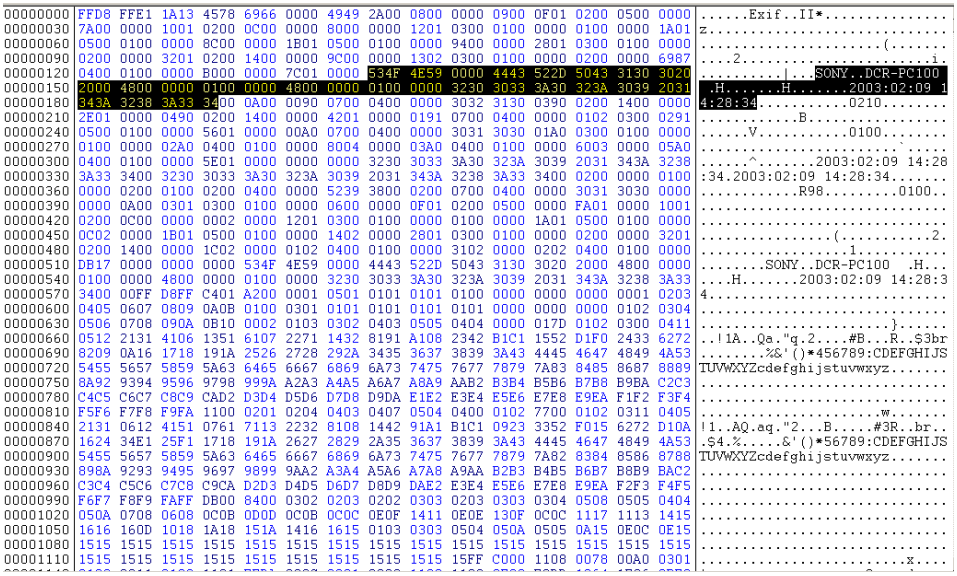

Abbildung 2: Bilddateidaten in kombinierter Hexadezimal- und Textdarstellung

Beim Einlesen der Informationen sind verschiedene grundsätzliche Formen der Informationsablage im Bild-Header zu unterscheiden, aus denen sich die Anforderungen an ein entsprechendes SAS-Makro ableiten lassen.
Wenige grundlegende Informationen befinden sich an einer fest definierten Byte-Position in der Datei. Ein überwiegender Teil der Informationen wird über spezifische Schlüsselwörter bzw. entsprechende Byte-Folgen zugänglich gemacht, die die Art und Form der vorhandenen Merkmale beschreiben. Diese Schlüsselwörter (Tags) können an unterschiedlichen Positionen innerhalb der Datei abgelegt sein und sind nur obligat, sofern eine bestimmte Information überhaupt in der Datei abgelegt wird. Die eigentlichen Informationen bzw. Merkmalsausprägungen folgen entweder direkt in den Bytes nach dem Schlüsselwort, alternativ können die nachfolgenden Bytes aber auch lediglich eine Positionsangabe der gewünschten Information innerhalb der Bilddatei enthalten.

3 Anforderungen an ein SAS-Makro zur automatisierten Erfassung von Bildinformationen

Vor dem Hintergrund der bisherigen Darstellungen lassen sich für eine automatisierte Bilderfassung folgende Anforderungen an ein SAS-Makro formulieren:

- Das Makro soll relevante Bilddateien in vorher spezifizierten Bereichen von Datenträgern (Verzeichnissen auf Festplatten) identifizieren können.

- Das Makro soll Abschnitte der identifizierten Dateien einlesen.
- In den eingelesenen Abschnitten soll die Verfügbarkeit von zuvor festgelegten Informationen überprüft werden. Sind entsprechende Merkmale vorhanden, sollen diese ausgelesen werden.
- Die ausgelesenen bildspezifischen Informationen sollen gemeinsam mit Informationen zur Identifikation der jeweiligen Bilddatei strukturiert in einer SAS-Datendatei erfasst werden. Bei Bedarf können die Informationen nachfolgend auch für automatisierte Dateimodifikationen, z.B. die Umbenennung der Bilddateien, genutzt werden.

4 Umsetzung der Makroprogrammierung

Nachfolgend soll die Umsetzung der genannten Anforderungen in einem SAS-Makro dargestellt werden, wobei nur grundsätzliche Elemente des Programm-Codes dargestellt werden. Das vollständige Makro ist auf der Tagungs-CD enthalten und kann beim Autor per Email angefordert werden.

4.1 Identifikation von Dateien in Verzeichnissen

Sollen Informationen zu einer größeren Anzahl von Bilddateien eingelesen werden, wäre eine manuelle Angabe der einzelnen Dateien zur Aufarbeitung im Makro recht mühsam und fehlerträchtig. Insofern wurde eine automatisierte Identifikation von relevanten Dateien im Makro implementiert. Allerdings musste hierbei auf Betriebsystem-spezifische Befehle zurückgegriffen werden. Getestet wurde die Makrofunktion unter Windows XP sowie Windows 2000. Bei Verwendung anderer Betriebssysteme sind ggf. Modifikationen des Makros erforderlich.
Als obligate Eingabe erwartet das Makro die vollständige Pfadangabe zum auszuwertenden Verzeichnis. Fakultativ können die vom Makro nachfolgend berücksichtigten Dateien durch weitere Angaben eingeschränkt werden. Als Resultat diese Makroabschnitts wird eine SAS-Datendatei mit Informationen zu allen im Verzeichnis befindlichen Bilddateien (insbesondere deren Dateinamen) erstellt. Die Anzahl der gefundenen Bilddateien wird in einer Makrovariablen verfügbar gemacht.

4.2 Einlesen der binären Daten

Nachfolgend durchläuft das Makro je Bilddatei dieselben Arbeitsschritte. Da die exakten Positionen der gesuchten Informationen in den Bilddateien nicht einheitlich sind und auch partiell fehlen können, empfiehlt sich als erster Schritt das komplette Einlesen potentiell relevanter Bilddateiabschnitte in eine Variable, die anschließend ohne erneuten Zugriff auf die Bilddatei weiter ausgewertet werden kann. Dies lässt sich im hier behandelten Fall recht einfach realisieren, da sich die gesuchten Meta-Informationen grundsätzlich am Anfang der Bilddatei, d.h.

in der Byte-Folge vor den eigentlichen und umfangreicheren Bilddaten befinden. Gleichzeitig bietet SAS seit der Version 8 die Möglichkeit Zeichen-Variablen mit einer Länge von bis zu 32767 Byte zu spezifizieren, was die zu erwartende Länge von Bilddatei-Headern deutlich übersteigt. Alle potentiell relevanten Informationen eines Bildes können also zunächst problemlos in einem einzigen Variablenwert einer SAS-Datendatei abgelegt werden.

SAS-Code: Einlesen von längeren Datei-Abschnitten in eine Variable

```
%LET DATALL= ;
%DO R=1 %TO &DATNO;
     data d&R;
     infile "&&DATNAME&R " lrecl=&XBYTE
                           recfm=f truncover obs=1;
     input header  $ASCII&&XBYTE.. ;
     no=&R;
     run;
     %LET DATALL=&DATALL d&R;
%END;

Erläuterungen:
&DATALL:  Makrovariable zum erfassen der Dateinamen
&DATNO:   Anzahl der Bilddateien bzw. der Durchläufe
&XBYTE:   Anzahl der auszulesenen Bytes [15000]
&&DATNAME&R:   Name der R-ten Bilddatei
```

4.3 Identifikation und Aufbereitung von Merkmalen

Nach dem Einlesen der Daten aus den Bilddateien erfolgt die Auswertung der Informationen, die zunächst je Bilddatei in einem einzigen Variablenwert in binärem Format vorliegen. Dies soll anhand einiger Merkmale beispielhaft erläutert werden. Das vollständige Makro kann unter Ausführung weiterer, prinzipiell jedoch gleichartiger, Programmschritte eine größere Zahl unterschiedlicher Merkmale erfassen, ohne allerdings dabei den Anspruch zu besitzen, alle potentiell in Datei-Headern gemäß EXIF-Spezifikationen enthalten Merkmale erfassen zu können.

Merkmale mit fester Position innerhalb der Datei
Alle Bilder gemäß JPEG-Spezifikation gleichen sich in den ersten beiden Bytes der Datei: Sie werden durch einen sogenannten Marker mit dem Hexadezimalwert FFD8 gebildet. Insofern lässt sich eine Überprüfung, ob die eingelesene Byte-Folge einer Datei überhaupt einer JPEG-Bilddatei entspringt, einfach formulieren.

SAS-Code: Identifikation des JPEG-Markers

```
pos_FFD8_1=index( header , byte(255)||byte(216) );
      *JPEG-Kennung vorhanden?;
if pos_FFD8_1=1 then do;
     *nur weiter falls JPEG-Bytes an Position 1+2;
```

Merkmale mit variabler Position innerhalb der Datei
Merkmale gemäß EXIF-Spezifikation können sich an unterschiedlichen Stellen der Byte-Folge in EXIF-relevanten Header-Abschnitten der Bilddatei befinden oder auch gänzlich fehlen, da sie nicht obligat von allen Kameras verwendet werden müssen. Ist ein bestimmtes Merkmal gespeichet, wird dies durch einen spezifischen „Tag“ angezeigt, der wie der JPEG-Marker durch zwei aufeinanderfolgende Bytes gekennzeichnet ist. EXIF-Tags lassen sich so einfach unter Verwendung der bereits im vorausgehenden Beispiel verwendeten Index-Funktion lokalisieren, welche die Position der gesuchten Tags in der durchsuchten Zeichenfolge ausgibt. Aus der Position des Tags kann wie im nachfolgenden Beispiel direkt auf die Position des gesuchten Merkmalwertes oder alternativ zumindest auf die Position einer Positionsangabe geschlossen werden, die erst in einem weiteren Schritt zum Auslesen des eigentlich gesuchten Wertes herangezogen werden kann. Werden, wie im nachfolgenden Beispiel, Positionen von Tags in Teilen einer Zeichenketten bzw. in Substrings bestimmt, ist dies selbstverständlich auch bei Positionsangaben zum Auslesen der gewünschten Werte zu berücksichtigen.

Aufbereitung der Merkmalswerte
Merkmale werden gemäß EXIF-Spezifikation in unterschiedliche Formaten abgelegt. Angaben zum Kameratyp sowie zum Aufnahmedatum werden beispielsweise, wie auch in Abbildung 2 erkennbar, als ASCII-Text gespeichert und können einfach als Substring aus Zeichenketten extrahiert werden. Numerische Merkmale sind demgegenüber in der Regel als binäre Werte gespeichert. Sie werden vom Makro Byte-weise ausgelesen und über Rank-Funktionen zunächst einzeln in Zahlenwerte umgewandelt, aus denen schließlich nach positionsabhängiger Multiplikation und Summation auch beliebige größere Zahlenwerte zusammengesetzt werden können. Dies Vorgehen arbeitet zuverlässig, führt allerdings zu relativ langem Programmcode.

SAS-Code: Einlesen binärer kodierter Zahlenwerte.

```
*Tag-Position in Substring suchen;

pos_FFC0=index( substr(header,EXIF_L,&XBYTE-exif_l) ,
byte(255)||byte(192) );

*ggf. Merkmalswerte aus ermittelten Positionen des
Substrings Byte-weise einlesen und in Zahlenwerte
umwandeln;

if pos_FFC0 gt 0 then do;
```

```
  ActualWidth=
    rank(substr(header,pos_FFC0+5+EXIF_L-1,1))*256
    +rank(substr(header,pos_FFC0+6+EXIF_L-1,1));
  ActualHight=
    rank(substr(header,pos_FFC0+7+EXIF_L-1,1))*256
    +rank(substr(header,pos_FFC0+8+EXIF_L-1,1));
end;
```

4.4 Weitere Nutzung der Informationen im Rahmen des Makros

Im Ablauf des hier vorgestellten SAS-Makros werden die Informationen nach einer Zusammenstellung in einer Datendatei genutzt, um die Bilddateien nach der Kopie in ein Zielverzeichnis automatisiert entsprechend ihres Aufnahmezeitpunktes umzubenennen. Gleichzeitig wird ein Unterverzeichnis erstellt, in dem Vorschaubilder (Thumbnails) zu den einzelnen Bilddateien abgelegt werden, welche zuvor in gleicher Art wie die übrigen Meta-Informationen aus den Daten der EXIF-Header extrahiert werden konnten. Zwangsläufig werden auch in diesem Abschnitt des Makros Betriebssystem-spezifische Befehle verwendet.

Resümee

Das hier vorgestellte (experimentelle) SAS-Makro zeigt exemplarisch, wie Meta-Informationen zum automatisierten Aufbau einer Datenbank genutzt werden können. Dies lässt sich in der SAS-Syntax zumindest im vorliegenden Beispiel mit relativ einfachen Mitteln realisieren, selbst wenn vorkonfektionierte Importfilter fehlen und Informationen in binären Formaten abgelegt sind. Da ein entsprechendes Vorgehen bei einer großen Zahl unterschiedlicher Dateitypen prinzipiell realisierbar sein dürfte, kann die Programmierung eines entsprechend angepassten SAS-Makros immer erwogen werden, sofern Informationen aus Meta-Daten a) relevant erscheinen, b) ausreichende Informationen zu deren Struktur vorliegen und c) eine Erfassung nicht nur sporadisch vorgesehen ist. Über entsprechende Makro-Programme können so Informationen ohne Rückgriff auf externe Programme in SAS-Datendateien verfügbar gemacht werden, die primär zunächst nicht zugänglich sind.

Literatur

1. Digital Still Camera Image File Format Standard
   (Exchangeable image file format for Digital Still Cameras: Exif)
   Version 2.1, June 12, 1998, Japan Electronic Industry Development Association (JEIDA); im Internet verfügbar unter
   http://www.kodak.com/global/plugins/acrobat/en/service/digCam/exifStandard.pdf