Referenzdaten für die computerassistierte Diagnose in der Mammographie

M. Elter[1], A. Horsch[2], R. Schulz-Wendtland[3], H. Sittek[5], M. Athelogou[4], G. Schmidt[4], T. Wittenberg[1]

[1]Fraunhofer-Institut für Integrierte Schaltungen IIS, Erlangen
[2]Inst. für Medizinische Statistik und Epidemiologie, TU München
[3]Definiens AG, München
[4]Radiologisches Institut der Universität Erlangen-Nürnberg
[5]Diagnostisches Mammazentrum München
Email: matthias.elter@iis.fraunhofer.de

Zusammenfassung. Die Computerassistierte Diagnose (CAD) in der Mammographie hat durch den europaweiten Aufbau von nationalen Screeningprogrammen in den letzten Jahren stark an Bedeutung gewonnen. Für die Evaluierung und vor allem für den Vergleich von CAD Algorithmen sind öffentlich zugängliche Referenzdaten nötig. Am Fraunhofer IIS wird derzeit eine Referenzdatenbank für die Mammographie aufgebaut, die die veralteten bestehenden Datenbanken ersetzen bzw. ergänzen soll. Dabei wurde auf dem Stand der Gerätetechnik entsprechende Bildqualität und zeitgemäße ikonische und textuelle Annotation geachtet. Die Veröffentlichung dieser Referenzdaten wird es Forschungsgruppen erlauben, CAD Algorithmen für die Mammographie auf umfangreichen und aktuellen Referenzdaten zu evaluieren und ihre Leistungsfähigkeit miteinander zu vergleichen.

1 Einleitung

Am Fraunhofer Institut für Integrierte Schaltungen (IIS) wird im Kontext eines mehrjährigen Projektes zur Computerassistierten Diagnose für die Mammographie eine Referenzdatenbank von digitalen Mammographien mit ikonischer und textueller Annotation aufgebaut. Die öffentliche Verfügbarkeit dieser Referenzdaten wird es Forschungsgruppen aus aller Welt erlauben, CAD Algorithmen für die Mammographie auf umfangreichen und aktuellen Referenzdaten zu evaluieren und ihre Leistungsfähigkeit miteinander zu vergleichen. Die Referenzdaten werden derzeit im Radiologischen Institut der Universität Erlangen-Nürnberg und im Diagnostischen Mammazentrum München erfasst, annotiert und kreuzvalidiert. Über die Projektlaufzeit von drei Jahren soll die Datenbank von derzeit 250 auf insgesamt 1000 Fälle erweitert werden.

2 Stand der Forschung und Fortschritt durch den Beitrag

Seit der Erstbeschreibung eines Systems zur computerassistierten Analyse von Mammographien durch Winsberg im Jahr 1967 [1] haben sich zahlreiche Arbeits-

gruppen mit dieser Problematik beschäftigt. Obwohl seither unzählige Lösungsansätze für diese Problemstellung veröffentlicht wurden (z.B. [2, 3, 4]), stehen weltweit nur zwei öffentliche Referenzdatenbanken zur Evaluation und zum Vergleich der verschiedenen Ansätze zur Verfügung. So hat die *Mammographic Image Analysis Society* bereits 1994 eine 320 Fälle umfassende Referenzdatenbank für die Mammographie veröffentlicht [5]. Die zweite Referenzdatenbank stammt von der University of South Florida [6] und umfasst zrika 2.500 Fälle. Beide Referenzdatenbanken stammen aus den frühen neunziger Jahren. Durch die rasante Weiterentwicklung der Akquisitionsgeräte und vor allem durch den Wechsel von analoger auf digitale Technik gilt die Bildqualität der enthaltenden Mammographien mittlerweile als veraltet. De facto steht derzeit weltweit keine Referenzdatenbank mit Mammographien, die dem aktuellen Stand der Gerätetechnik entsprechen, zur Verfügung.

Neben der, aus heutiger Sicht, unzureichenden Bildqualität der bestehenden Referenzdaten, ist auch der Umfang und die Qualität der Annotation der Daten nicht mehr zeitgemäß. So hat das *American College of Radiology* (ACR) mit dem *Breast Imaging Reporting and Data System* (BI-RADS) [7] in der Zwischenzeit einen weltweit angewandten Standard für die Befundung und damit auch Annotation von Mammographien entwickelt bzw. weiterentwickelt. Dieser umfasst neben der Klassifizierung von Läsionen wie Herdbefunden und Mikrokalzifizierungen auch eine standardisierte Beschreibung ihrer konkreten Eigenschaften wie der Form, Verteilung oder Begrenzung.

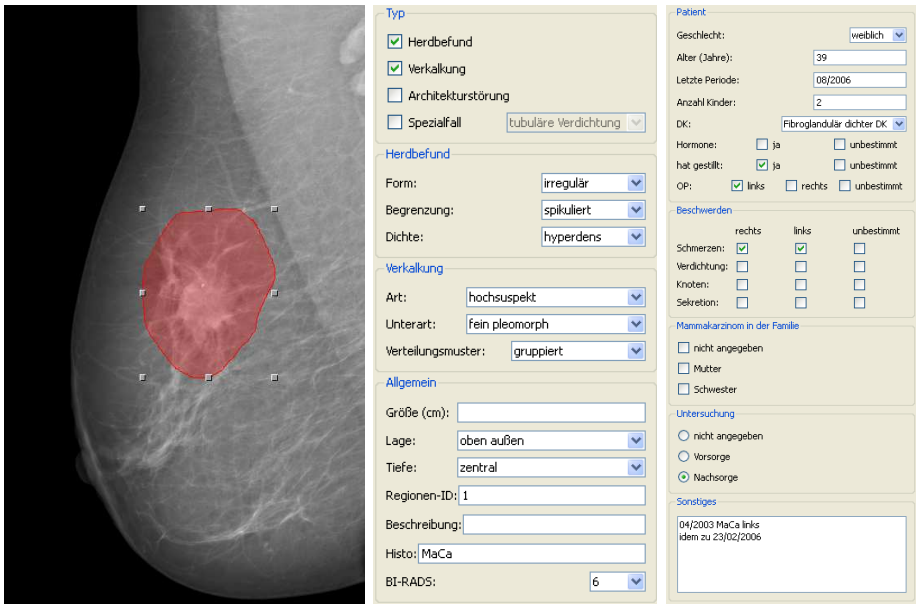
Um mögliche Risikofaktoren und Symptome des Mammakarzinoms zu erfassen, lassen Radiologen umfangreiche Anamnesedaten, wie zum Beispiel Angaben zu Krebserkrankungen in der Familie, in ihre Diagnose miteinfließen. Um eine höhere Erkennungsrate zu erzielen werden derartige Zusatzinformationen zunehmend auch von CAD Systemen berücksichtigt. Die beiden bestehenden Referenzdatenbanken enthalten aber leider keine bzw. nur sehr wenig Anamnesedaten.

Eine zeitgemäße Referenzdatenbank sollte also neben einer ikonischen Annotation von Läsionen auch eine auf der BI-RADS Nomenklatur basierende textuelle Annotation der Eigenschaften dieser Läsionen sowie textuelle Anamnesedaten enthalten. Daher wird am Fraunhofer IIS eine Referenzdatenbank für die Mammographie aufgebaut. Sie enthält im Gegensatz zu den bestehenden Referenzdatenbanken ausschließlich mit modernen volldigitalen Geräten gewonnenes Bildmaterial. Die Annotation enthält neben BI-RADS konformen und kreuzvalidierten Diagnosen auch umfangreiche Anamnesedaten.

3 Methoden

Die im Aufbau befindliche Referenzdatenbank enthält Mammographien die im Radiologischen Institut der Universität Erlangen-Nürnberg und im Diagnostischen Mammazentrum München erfasst werden. Jeder Datensatz umfasst in der Regel zwei Mammographien (die Standardansichten medio-lateral oblique und cranio caudal) je Brust. Am Radiologischen Institut der Universität Erlangen-

Abb. 1. Ikonisch und textuell annotierte Mammographie (rechte Brust medio-lateral oblique) aus einem typischen Datensatz der Referenzdatenbank



Nürnberg kommen dazu zwei volldigitale Geräte (Siemens Mammomat Novation DR) und am Diagnostischen Mammazentrum München ein Speicherfoliengerät von Agfa zum Einsatz. Neben den Mammographien im DICOM Format enthält jeder Datensatz textuelle sowie ikonische Annotation. Zur einfachen Weiterverarbeitung ist die gesamte Annotation im XML Format gespeichert.

3.1 Ikonische Annotation

Die ikonische Annotation von Läsionen erfolgt mittels einer Annotationssoftware die am Fraunhofer IIS speziell für die Annotation von Mammographien entwickelt wurde. Läsionen werden dazu vom Befunder mit der Maus bzw. am Touchscreen eingezeichnet. Diese eingezeichneten Regionen dienen zusammen mit einer zugeordneten Klassifizierung nach BI-RADS als Ground-Truth Daten. Sie werden in den XML basierten Annotationsdateien sowohl vektoriell als Polygonzug, wie auch rasterorientiert mittels einer Binärmaske gespeichert.

3.2 Textuelle Annotation

Jede ikonische Annotation einer Läsion ergänzen zusätzliche, der BI-RADS Nomenklatur entsprechende, textuelle Annotationen. Diese umfassen bei Mikrokalzifizierungen die Art, Unterart und das Verteilungsmuster des Kalks und bei

Herdbefunden die Form, Begrenzung sowie Dichte des Herdes. Sowohl bei Mikrokalzifizierungen als auch bei Herden umfasst die textuelle Annotation zusätzlich die Größe und Lage der Läsion sowie den histologischen Befund.

Neben der textuellen Beschreibung von Läsionen umfasst ein Referenzdatensatz auch textuelle Anamnesedaten. Diese beinhalten das Alter, Geschlecht und Gewicht des Patienten, sowie das Datum der letzten Periode, die Anzahl der Kinder, Angaben über verabreichte Hormonpräparate und frühere Brustoperationen. Dazu kommen Angaben zu Beschwerden wie Schmerzen, tastbaren Knoten oder Sekretion der Brust sowie Angaben zu Mammakarzinomen in der Familie.

3.3 Pseudonymisierung

Um die persönlichen Daten der Patienten zu schützen ist es erforderlich alle personenbezogenen Daten, die Rückschlüsse auf die Identität der Patienten liefern können, aus den Mammographiedaten zu entfernen (Anonymisierung). Um einzelne Patientengeschichten langfristig (mehrere Mammographiekontrollen über Monate oder Jahre verteilt) verfolgen zu können, ist es aber gleichzeitig wichtig, Mammographiedatensätze ein und desselben Patienten, die zu verschiedenen Zeitpunkten erstellt wurden, einander zuordnen zu können. Daher werden die Patientendaten für die Referenzdatenbank pseudonymisiert.

Bei der Pseudonymisierung wird der Name oder ein anderes Identifikationsmerkmal durch ein eindeutiges Pseudonym ersetzt, um die Identifizierung des Betroffenen auszuschließen. Im Gegensatz zur Anonymisierung bleiben bei der Pseudonymisierung Bezüge verschiedener Datensätze, die auf dieselbe Art pseudonymisiert wurden, erhalten. Konkret wird für die Mammographiereferenzdatenbank aus dem Patientennamen und dem Geburtsdatum mittels der MD5 (Message Digest Algorithm 5) [8] Hashfunktion ein Hashwert gebildet und als Pseudonym verwendet. Auf diese Weise können Mammographiedatensätze ein und desselben Patienten sicher einander zugeordnet werden ohne die Annonymität der Patienten zu gefährden.

4 Ergebnisse

Am Fraunhofer IIS wird zur Zeit eine umfangreiche Referenzdatenbank für die Mammographie aufgebaut und anschließend öffentlich verfügbar gemacht. Die Datenbank umfasst derzeit 250 Fälle. Für die textuelle und ikonische Annotation der Daten wurde eine Annotationssoftware entwickelt, die eine Annotation von Mammographien anhand der Nomenklatur des BI-RADS Standards erlaubt. Art und Umfang der Annotation richtet sich neben den Vorgaben des BI-RADS Standards vor allem nach den Anforderungen moderner CAD Systeme für die Mammographie. Das Problem der Anonymisierung von Patientendaten wurde mittels eines Pseudonymisierungsansatzes gelöst. Ein typischer Datensatz der Referenzdatenbank ist in Abb. 1 dargestellt.

5 Diskussion

Die neue Referenzdatenbank für die Mammographie ist eine Alternative zu den beiden bestehenden aber veralteten frei zugänglichen Datenbanken. Durch den Einsatz von modernen volldigitalen Akquisitionsgeräten ist die Bildqualität deutlich höher als die der bisher verfügbaren Referenzdaten. Neben den Bilddaten und der ikonischen Annotation von relevanten Bildbereichen besteht ein Datensatz auch aus textueller Annotation der Eigenschaften von Läsionen sowie der Patientengeschichte. Die Referenzdaten zeichnen sich daher auch durch einen größeren Umfang an Zusatzinformationen aus. Diese Informationen können vor allem in wissensbasierten Systemen oder für das fallbasierte Schließen verwendet werden. Neben dem Einsatz zum Training, zur Evaluation und zum Vergleich von CAD Systemen für die Mammographie eignen sich die annotierten Daten auch als Fallbeispiele für die medizinische Aus- und Weiterbildung.

Danksagung

Diese Arbeit wurde von der Bayerischen Forschungsstiftung im Rahmen des Projektes Mammo-iCAD gefördert.

Literaturverzeichnis

1. Winsberg F, Elkin M, Marcy J, Bordaz V, Weymouth W. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. Radiology 1967;89:211–215.
2. Yu Songyang, Guan Ling. A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films. IEEE Transactions on medical imaging 2000; 115–126.
3. Peitgen ThomasNetsch*andHeinzOtto. Scale-Space signatures for the detection of clustered Microcalcifications in digital mammograms. IEEE Transactions on medical imaging 1999;18(9).
4. Linda J Warren Burhenne,Susan A Wood, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215(2):554–562.
5. Suckling J, Parker J, Dance D, Astley S, Hutt I, Boggis C, et al. The mammographic images analysis society digital mammogram databases. Exerpta Medica International Congress Series 1069 1994; 375–378.
6. Heath M, Bowyer K, Kopans D, Moore R, Jr PKegelmeyer. The Digital Database for Screening Mammography. In: The Proceedings of the 5th International Workshop on Digital Mammography. Madison, WI, USA: Medical Physics Publishing; 2000.
7. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS©) Atlas; 2006.
8. Rivest R. Request for comments (RFC) 1321 - The MD5 message-digest algorithm. http://www.ietf.org/rfc/rfc1321.txt; 1992.