Leistungsprädiktion von Algorithmen zur Mammographie-Bildauswertung

P. Klausmann, K. Kroschel

Fraunhofer-Institut für Informations- und Datenverarbeitung (IITB)
Fraunhoferstr. 1, 76131 Karlsruhe
Email: { klm, kro }@iitb.fhg.de

Zusammenfassung Der vorliegende Ansatz zeigt neue Methoden zur Vorhersage der Verfahrensleistung bei der Detektion von Mikrokalzifikationen in Mammogrammen auf. Dazu werden für ein gegebenes Mammogramm-Auswertungsverfahren einerseits für die zu detektierenden Mikrokalzifikationen, andererseits für den verbleibenden Hintergrund-Bildbereich separate Modelle erstellt, aus denen verfahrensspezifische Bildqualitätsmaße abgeleitet werden können. Um die Qualität dieser Prädiktionsmodelle nachzuweisen, wird die Verfahrensleistung aufgrund der extrahierten Bildqualitätsmaße geschätzt und mit der tatsächlich erzielten Verfahrensleistung verglichen.

Schlüsselwörter: Leistungsprädiktion, verfahrensspezifische Bildqualität, Bewertung, Mammographie

1 Einleitung

Die wachsende Zahl und zunehmende Komplexität verfügbarer Algorithmen zur automatischen und interaktiven Bildauswertung in der Medizin erfordern eine Einordnung und Bewertung, die dem Anwender eine gezielte Auswahl ermöglicht und ihm gleichzeitig detaillierte Informationen über Einsatzmöglichkeiten und Grenzen der Verfahren bereit stellt.

Der vorliegende Ansatz skizziert eine Strategie zur Modellierung von Verfahren zur Detektion von Mikrokalzifikationen in Röntgen-Mammogrammen. Aus einem solchen Modell können verfahrensspezifische Maße für die Bildqualität, von der die Verfahrensleistung maßgeblich beeinflusst wird, extrahiert werden. Solche Bildqualitätsmaße sind insbesondere bei begrenzter Bilddatenmenge und gleichzeitig hohen Anforderungen an die Robustheit und Zuverlässigkeit der Systeme von größter praktischer Relevanz [1][2]. Verfahrensmodelle liefern umfangreiche Zusatzinformationen über ein gegebenes Verfahren und ermöglichen beispielsweise die Prädiktion der Verfahrensleistung auf neuen Bilddaten. Weiterhin können damit Bildaufnahmeparameter eines Sensors so eingerichtet werden, dass er die für ein Verfahren optimale Bildqualität liefert. Bei bekannter Bildqualität wird die Auswahl eines optimalen Verfahrens erleichtert.

Die wichtigsten Leistungskenngrößen für Detektionsalgorithmen sind die Detektions- und die Falschalarmrate. Um diese beiden Größen für ein gegebenes

Detektionsverfahren zu bestimmen, müssen für die betrachteten Bilddaten geeignete Referenzdaten (sogenannte Truthdaten) vorhanden sein, die angeben, an welchen Stellen im Bild sich vom Verfahren zu detektierende Strukturen befinden. Dann können die Verfahrensergebnisse mit den Referenzdaten verglichen und die entsprechenden Leistungskenngrößen bestimmt werden. Im vorliegenden Fall wurden alle Strukturen, die als mögliche Kandidaten pathologischer Mikrokalzifikationen vom Arzt einer näheren Betrachtung unterzogen werden sollten (im folgenden einfach nur noch als Mikrokalzifikationen bezeichnet), von einem menschlichen Betrachter durch umschreibende Polygone (Truth-Polygone) in Bildkoordinaten markiert.

Die Aufgabe des Verfahrens ist damit exakt spezifiziert und besteht darin, die Schwerpunkte dieser Strukturen im Bild zu ermitteln. Die Schwerpunkte werden nach Anwendung des Verfahrens auf einen Bilddatensatz untersucht, ob sie sich innerhalb oder außerhalb eines Truth-Polygons befinden. Im ersten Fall handelt es sich um eine Detektion, andernfalls um einen Falschalarm. Truth-Polygone, die keinen Schwerpunkt enthalten, werden als Nichtdetektionen gewertet. Abb. 1 zeigt einen Ausschnitt eines Mammogramms mit eingeblendeten Truth-Polygonen und Ergebnissen des am IITB entwickelten Verfahrens MAMDET [3].

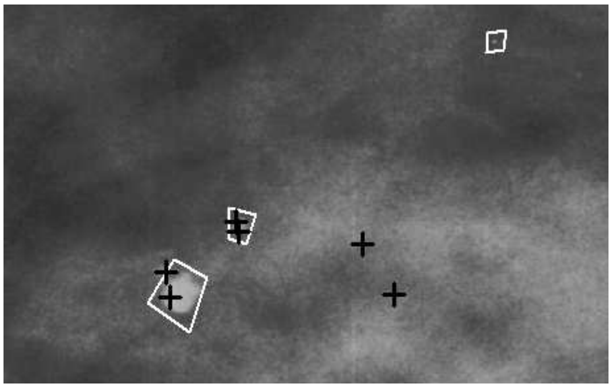

Abbildung1. Ausschnitt eines Mammogramms mit eingeblendeten Truth-Polygonen (weiße Polygone) und Verfahrensergebnissen (schwarze Kreuze). Man erkennt zwei (mehrfach) detektierte Mikrokalzifikationen, zwei Falschalarme und eine Nichtdetektion.

2 Modellierung des Detektionsverhaltens

Damit eine Mikrokalzifikation vom Detektionsverfahren detektiert werden kann, muss sich deren Signatur möglichst deutlich vom Bildhintergrund unterscheiden.

Da nicht nur die Signaturen der Mikrokalzifikationen, sondern auch der verbleibende Bildhintergrund starken Schwankungen unterliegen kann, werden im Folgenden Maße für die *lokale Detektierbarkeit* definiert, die dann zur Erstellung eines Modells für das Detektionsverhalten des Verfahrens und zur Prädiktion der Detektionsleistung verwendet werden können. Es wird also eine kleine Umgebung um die Mikrokalzifikationen herum betrachtet, die als *lokaler Hintergrund* bezeichnet wird. In diesem Fall werden als lokaler Hintergrund achsenparallele, rechteckige Umgebungen um die Truth-Polygone verwendet. Dann werden aus den Signaturen der Mikrokalzifikationen und dem lokalen Hintergrund folgende Merkmale x_i, $i = 1, 2, 3$ berechnet:

- x_1: Differenz der Anzahl der Kantenpixel der Mikrokalzifikationen und der Anzahl der Kantenpixel bezogen auf die jeweilige Bildfläche. Die Kantenpixel werden mit dem Sobel-Operator berechnet.
- x_2: Differenz des mittleren Grauwertes der Mikrokalzifikationen und des lokalen Hintergrunds.
- x_3: Größe der Mikrokalzifikationen in Pixel.

Der Zusammenhang des Merkmalsvektors $x = (x_1, x_2, x_3)^t$ und der Wahrscheinlichkeit P_D einer Signatur, vom Verfahren detektiert zu werden, kann durch eine logistische Funktion geschätzt werden [4]:

$$\hat{P}_D = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{1}$$

mit

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Die Koeffizienten β_i, $i = 1, 2, 3$ sind mittels eines Maximum-Likelihood Verfahrens aus den auf dem Datensatz erzielten Ergebnissen zu schätzen [5]. Dies bedeutet, dass die Merkmale mit dem dichotomen Ergebnis (Mikrokalzifikation vom Verfahren detektiert oder nicht detektiert) in Verbindung gebracht werden müssen, so dass die Wahrscheinlichkeit, die gegebenen Daten zu erhalten, maximiert wird [4]. Sind die logistischen Regressionskoeffizienten erst einmal geschätzt, steht mit dem Merkmalsvektor x und Gl. (1) ein *verfahrensspezifisches* Modell zur Verfügung, das angibt, mit welcher Wahrscheinlichkeit eine gegebene Signatur mit gegebenem Merkmalsvektor vom betrachteten Verfahren detektiert wird. Der Mittelwert der Einzeldetektionswahrscheinlichkeiten dient dann als Schätzwert für die Detektionsleistung des Verfahrens.

3 Modellierung des Falschalarmverhaltens

Genauso wichtig wie ein Modell für die Detektionsleistung ist ein Modell zur Prädiktion der Falschalarme. Um ein Qualitätsmaß für den Bildhintergrund zu erhalten, werden im vorliegenden Fall sogenannte Richtungsflecken verwendet, die in [6] auch als *line-support-regions* bezeichnet werden. Die Berechnung dieser Richtungsflecken beginnt mit der Bestimmung des Grauwertgradienten im

Bild. Alle Pixel, die eine ähnliche Gradientenrichtung aufweisen, werden zu Regionen zusammengefasst. Der 360°-Bereich der Gradientenrichtungen wird im vorliegenden Fall in vier Bereiche partitioniert. Jedes Pixel wird ensprechend seiner Gradientenrichtung einer Partition zugewiesen. Benachbarte Pixel einer Partition werden zu einfach-zusammenhängenden Regionen zusammengefaßt.

Wird die Menge der Richtungsflecken in Abhängigkeit von Attributen wie Größe und mittlerer Gradientenbetrag in Teilmengen zerlegt, so wird eine detaillierte Beschreibung der Textureigenschaften, die dem Bild zugrunde liegen, ermöglicht. Besondere Relevanz haben diejenigen Richtungsflecken, die vergleichbare Attribute haben wie diejenigen, die auch durch Mikrokalzifikationen hervorgerufen werden.

Daher wird im folgenden die statistische Verteilung dieser Richtungsflecken betrachtet. Als erstes wird die auf die Bildfläche bezogene Anzahl n dieser Richtungsflecken betrachtet (*Richtungsfleckendichte*). Zusätzlich wird die *Varianz* σ^2 der mittleren Gradientenorientierungen dieser Richtungsflecken herangezogen und für jedes Bild bestimmt. Richtungsfleckendichte und deren Orientierungsvarianz werden mittels Fuzzy-Membership-Funktionen $T_n(n)$ bzw. $T_{\sigma^2}(\sigma^2)$ in das Intervall [0, 1] transformiert. Daraus wird ein sogenannter *Schwierigkeitsgrad* $S = min[T_n(n), T_{\sigma^2}(\sigma^2)] \in [0,\ 1]$ berechnet. Für einen hohen Wert des Schwierigkeitsgrades S ist sowohl eine hohe Richtungsfleckendichte als auch eine hohe Orientierungsvarianz erforderlich. Um den Zusammenhang zwischen diesem Schwierigkeitsgrad und der Anzahl der Falschalarme pro Bild herzustellen, wurde für das Falschalarmmodell des MAMDET-Verfahrens eine lineare Regressionsfunktion verwendet. Der Schwierigkeitsgrad zusammen mit der Regressionsgeraden dient als *verfahrensspezifisches* Modell für die Anzahl der Falschalarme, die aufgrund des Schwierigkeitsgrades S eines Bildes erwartet werden kann [3].

4 Ergebnisse

Um die Prädiktionsgenauigkeit der erstellten Modelle zu untersuchen, wurde das Verfahren MAMDET auf 14 Mammogramme[1] mit insgesamt 146 annotierten Mikrokalzifikationen angewandt und ein Detektions- sowie ein Falschalarmmodell erstellt. Das Verfahren detektierte auf dem Datensatz 95 Mikrokalzifikationen und verursachte durchschnittlich 7 Falschalarme pro Bild.

Die Bilddaten wurden dann einer JPEG-Datenkompression, weißem Rauschen sowie Kontrastreduktionen unterschiedlicher Stärke unterzogen. Anhand dieser gestörten Bilddaten wurde die Leistung des Verfahrens mittels der Modelle geschätzt. Durch Anwendung des Verfahrens und Vergleich der Verfahrensergebnisse mit der Truth wurde die tatsächlich erzielte Leistung des MAMDET-Verfahrens ermittelt. Der Vergleich der geschätzten mit den tatsächlich erzielten Detektions- bzw. Falschalarmraten ist in Abb. 2 dargestellt. Die Differenz zwischen geschätzter und tatsächlich erzielter Verfahrensleistung bei der Detek-

[1] Die Mammogramme wurden freundlicherweise vom Forschungszentrum Karlsruhe zur Verfügung gestellt.

tionsrate schwankt zwischen 1-9 %. Bei der Falschalarmrate bewegt sich der Schätzfehler im Bereich 1-2 Falschalarme pro Bild.

Die verwendeten Merkmale sowie der Schwierigkeitsgrad stellen *verfahrens-spezifische* Bildqualitätsmaße dar, die mittels der Modelle angeben, welche Leistung vom Verfahren bei gegebenen Bilddaten erwartet werden kann.

Zukünftig sollte durch eine verbreiterte Merkmalsbasis (größere Anzahl von Merkmalen und Hintergrundbeschreibungen) die Menge der Beschreibungen der Bildeigenschaften erweitert werden. Besonders von Interesse wäre eine Erprobung der Modelle in der medizinischen Praxis.

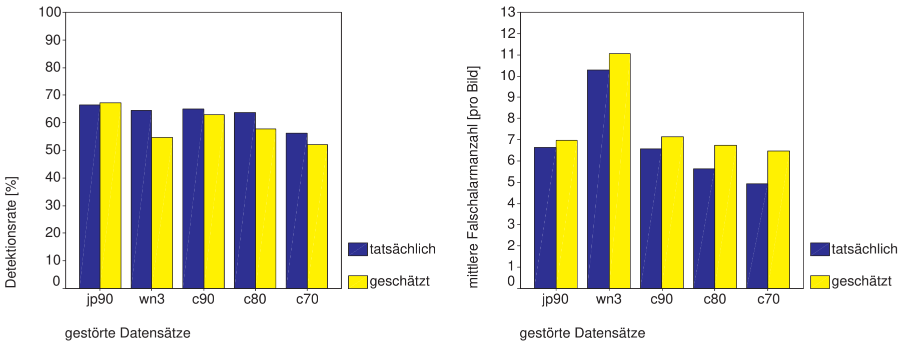

Abbildung2. Vergleich der geschätzten mit den vom Verfahren MAMDET tatsächlich erzielten Detektions- (links) bzw. Falschalarmraten (rechts) zur Untersuchung der Prädiktionsgenauigkeit der Modelle. Die gestörten Datensätze sind folgendermaßen bezeichnet: jp90 (mit Kompressionsfaktor 0.9 JPEG-komprimiert), wn3 (weißes Rauschen mit Standardabweichung 3), c90-c70 (kontrastreduziert auf 90-70 % des ursprünglichen Kontrastes).

Literatur

1. Courtney P, Thacker N, Clark AF: Algorithmic modelling for performance evaluation. Machine Vision and Applications 9:219–228, Springer Verlag, 1997.
2. Klausmann P, Kroschel K, Willersinn W: Performance prediction of vehicle detection algorithms. Pattern Recognition, Vol. 32, pp. 2063–2065, August 1999.
3. Klausmann P: Anwendungsorientierte Leistungsprädiktion und Bewertung von Detektionsverfahren. Eingereichte Dissertation, Universität Karlsruhe, Oktober 1999.
4. Hosmer DW, Lemeshow S: Applied logistic regression. John Wiley & Sons, New York, 1989.
5. McCullagh P, Nelder JA: Generalized Linear Models. Second Edition, Chapman & Hall, London, 1989.
6. Burns JB, Hanson AR, Riseman EM: Extracting straight lines. IEEE Trans. on Pattern Anal. Mach. Intell. 8 (4):425–455, July 1986.