Empirische Vergleichsmaße für die Evaluation von Segmentierungsergebnissen

Tobias Heimann, Matthias Thorn, Tobias Kunert und Hans-Peter Meinzer

Abteilung für Medizinische und Biologische Informatik,
Deutsches Krebsforschungszentrum, 69120 Heidelberg
Email: t.heimann@dkfz.de

Zusammenfassung. In diesem Beitrag entwickeln wir eine Methodik zum umfassenden und objektiven Vergleich von Segmentierungsergebnissen. Dazu wurde zuerst die Verwendung bestehender Gütemaße in der Literatur analysiert. Die unterschiedlichen empirischen Maße wurden kategorisiert und auf ihre Einsatzfähigkeit in der medizinischen Bildverarbeitung überprüft. Die erfolgversprechendsten Methoden wurden in einer klinischen Studie auf ihre Korrelation miteinander untersucht, um die kleinstmögliche Menge von komplementären Maßzahlen zu erhalten.

1 Einleitung

Im Zuge der steigenden Bedeutung der Qualitätssicherung in der Medizin darf die Validierung von Segmentierungsverfahren nicht vernachlässigt werden. Die Reproduzierbarkeit und Genauigkeit der Ergebnisse sind letztendlich mitentscheidend für die Güte der Patientenversorgung. Interessant ist hierbei nicht nur die Frage, welche Qualität automatische Segmentierungsverfahren im Vergleich zu manuellen Methoden erzielen, sondern auch, inwiefern die Ergebnisse zwischen unterschiedlichen automatischen oder manuellen Verfahren variieren. Um eine allgemeine und objektive Bewertung dieser Fragestellungen zu ermöglichen, führt kein Weg an quantitativen, mathematisch fundierten Methoden vorbei. In der Praxis sind die geeigneten Gütemaße jedoch oft unbekannt oder werden falsch eingeschätzt, da z.B. der Einfluss spezifischer Parameter übersehen wird. Dieser Beitrag stellt die unterschiedlichen Evaluationsmethoden vor, arbeitet Gemeinsamkeiten und Unterschiede heraus und präsentiert letztendlich eine Systematik zum umfassenden Vergleich von Segmentierungsergebnissen.

2 Stand der Forschung

Für die Evaluation von Segmentierungsverfahren gibt es einige grundsätzlich unterschiedliche Herangehensweisen [1]: Analytische Methoden bewerten die Segmentierungsalgorithmen direkt, indem sie Eigenschaften wie Funktionsweise, Anforderungen, Anwendungsbereiche und Komplexität untersuchen. Empirische Methoden bewerten die Algorithmen anhand von Segmentierungsergebnissen,

die auf speziellen Testbildern erzielt worden sind. Dabei gibt es zwei Möglichkeiten: Zum einen kann das Ergebnis an festgelegten Qualitätsmaßstäben gemessen werden (Goodness-Methoden), zum anderen kann die Abweichung von einem idealen Ergebnis gemessen werden (Diskrepanz-Methoden). Applikationsbezogene Methoden dagegen bewerten Segmentierungsalgorithmen anhand ihres Nutzwertes für eine bestimmte Anwendung.

Der vorliegende Beitrag konzentriert sich aufgrund der vielseitigen Einsatzmöglichkeiten und großen Popularität auf die Klasse der empirischen Diskrepanz-Methoden. In der medizinischen Bildverarbeitung wird dabei i.A. das Segmentierungsergebnis der zu überprüfenden Methode mit einem als richtig anerkannten Goldstandard verglichen.

Eine der einfachsten Möglichkeiten, zwei Segmentierungsergebnisse A und B miteinander zu vergleichen, ist die Größe der Schnittmenge $(A \cap B)$, auch bekannt als Simple Matching Coefficient [2]. Im dreidimensionalen Fall entspricht das dem von beiden Körpern geteilten Volumen, im zweidimensionalen Fall der gemeinsamen Fläche. Der Dice-Koeffizient [3], auch Sorensen- oder Czekanowski-Koeffizient genannt, normalisiert diesen Wert auf einen Bereich von 0 bis 1:

$$C_D = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

Eng verwandt damit ist der etwas konservativere Tanimoto-Koeffizient [4], der auch als Jaccard-Koeffizient bekannt ist:

$$C_T = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Der Kappa-Wert bezieht auch die zufällige oder statistisch zu erwartende Schnittmenge mit in die Berechnung ein. Dazu muss zusätzlich eine Region of Interest R definiert werden, die A und B umschließt. Aus der beobachteten Übereinstimmung P_0 und der erwarteten Übereinstimmung P_e, gegeben durch

$$P_0 = \frac{|\,(A \cap B) + (\bar{A} \cap \bar{B})\,|}{|R|} \text{ und } P_e = \frac{|A||B| + |\bar{A}||\bar{B}|}{|R|^2} \tag{3}$$

lässt sich dann der Kappa-Wert wie folgt definieren:

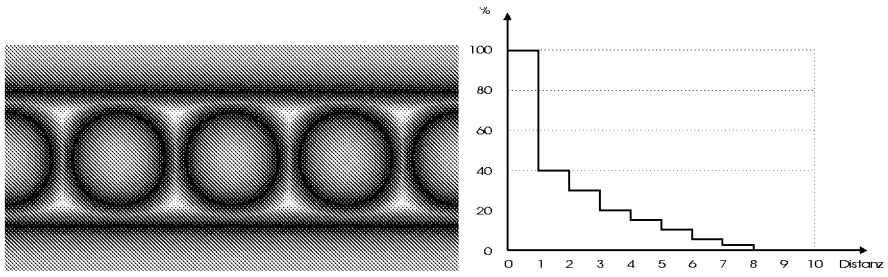
$$\kappa = \frac{P_0 - P_e}{1 - P_e} \tag{4}$$

Anstatt die Ähnlichkeit zweier Segmentierungen anhand der Schnittmenge zu bewerten, kann man auch die Distanz zwischen den beiden Oberflächen zum Vergleich heranziehen. So ist die Hausdorff-Distanz definiert als [5]:

$$H(A,B) = \max\{\boldsymbol{h}(A,B), \boldsymbol{h}(B,A)\} \text{ mit } \boldsymbol{h}(A,B) = \max_{a \in A} \min_{b \in B} d(a,b) \tag{5}$$

wobei A und B hier die Punktmengen der beiden Oberflächen darstellen und $d(a,b)$ eine Distanz zwischen zwei Punkten. Die Hausdorff-Distanz entspricht

Abb. 1. Grafische Darstellungen der Distanzverteilung: (a) Lokale Distanzunterschiede zwischen Würfel und Kugel; helle Punkte entsprechen hohen Distanzen. (b) Für jede Distanz ist der Anteil aller Punkte abzulesen, die diesen Abstand überschreiten.



(a) (b)

damit der maximalen Abweichung zwischen zwei Segmentierungen. Die durchschnittliche Oberflächendistanz lässt sich wie folgt berechnen:

$$\frac{\sum_{a\in A} \min_{b\in B} d(a,b) + \sum_{b\in B} \min_{a\in A} d(b,a)}{|A| + |B|} \tag{6}$$

Bei beiden Distanzmaßen fällt auf, dass die Entfernungen jeweils in beide Richtungen gemessen werden. Dieser Schritt ist notwendig, um das Symmetrie-Kriterium einer Metrik zu erfüllen, da sich von A nach B u.U. andere Punktkorrespondenzen mit anderen Distanzen entwickeln als von B nach A.

3 Material und Methoden

Die in der Literatur beschriebenen Vergleichsmaße wurden in vier Kategorien unterteilt: Überlappungsmaße, Oberflächendistanzen, Formdeskriptoren und Deformationsenergien. Jede Kategorie wurde auf ihre Eignung für die medizinische Bildverarbeitung untersucht, wobei diese i.A. unabhängig von bestimmten Modalitäten und Szenarien bewertet werden kann.

Um einen Überblick zu gewinnen, welche Methoden in der wissenschaftlichen Gemeinschaft populär sind, wurden 40 ausgewählte Publikationen der BVM-Tagungen aus den vergangenen drei Jahren analysiert. Alle untersuchten Veröffentlichungen beschäftigen sich mit neuen Segmentierungsverfahren in der medizinischen Bildverarbeitung und liefern potentiell auswertbare Ergebnisse.

Eine Auswahl von Gütemaßen wurde schließlich in einer objektorientierten Evaluationssoftware implementiert. Im Einzelnen handelt es sich dabei um die in Abschnitt 2 vorgestellten Maße: Tanimoto- und Dice-Koeffizient, Kappa-Wert, durchschnittliche Oberflächendistanz und Hausdorff-Distanz. Die Software kann sowohl mit grafischer Oberfläche als auch als Skript gestartet werden, um unterschiedlichen Anforderungen (interessante Einzelfälle oder hohes Datenaufkommen) gerecht zu werden. Zur Visualisierung der Oberflächendistanzen in der grafischen Variante wurden zwei neue Konzepte realisiert: Die Projektion der

Tabelle 1. Pearson-Korrelationskoeffizienten zwischen unterschiedlichen Vergleichsmaßen, berechnet über 5280 Einzelvergleiche.

Vergleichsmaße	Korrelation
Tanimoto – Dice	1,00
Tanimoto – Kappa	0,98
Dice – Kappa	0,98
Tanimoto – Mittlere Distanz	-0,40
Tanimoto – Hausdorff	-0,14
Mittlere Distanz – Hausdorff	0,68

Abstände auf eine 2D-Karte zur lokalen Auswertung (Abb. 2(a)) und die Darstellung in einem akkumulierten Histogramm zur globalen Auswertung (Abb. 2(b)).

In einer Studie mit 12 Medizinstudenten wurden von jedem Probanden in vier Durchläufen fünf unterschiedliche Volumen-Datensätze segmentiert. Jedes Volumen bestand dabei aus fünf aufeinanderfolgen Schichten einer Leber-CT-Aufnahme mit 3mm Schichtabstand. Ein Inter-Observer-Vergleich zwischen allen Probanden setzte sich damit aus 5280 unterschiedlichen Einzelvergleichen zusammen, in denen für jedes implementierte Gütemaß ein Ergebnis berechnet wurde. Alle Ergebnisse wurden mit dem Kolmogorov-Smirnov-Test auf Normalverteilung untersucht und der Pearson-Korrelationskoeffizient zwischen den unterschiedlichen Maßen berechnet.

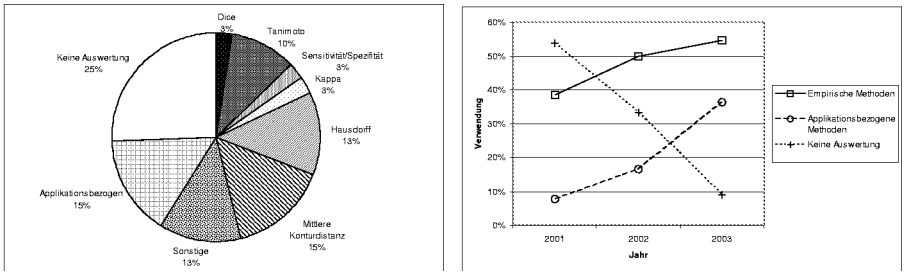
4 Ergebnisse

Die Literaturrecherche hat ergeben, dass 60% der BVM-Autoren empirische Vergleichsmaße zur Auswertung verwenden, 15% benutzen applikationsbezogene Methoden und ein Viertel verzichtet auf eine Evaluation (Abb. 3(a)). Die populärsten empirischen Vergleichsmaße sind – zumindest im deutschen Sprachraum – die mittlere und die maximale Oberflächendistanz, wobei letztere der Hausdorff-Distanz entspricht. Unter den Überlappungsmaßen sticht der Tanimoto-Koeffizient hervor, alle anderen Methoden tauchen nur selten auf.

Wenn man die Ergebnisse zeitlich sortiert, wird eine interessante Entwicklung sichtbar (Abb. 3(b)): Offensichtlich ist die Evaluation von Segmentierungsverfahren erst in den letzten beiden Jahren ins Bewusstsein der wissenschaftlichen Gemeinschaft vorgedrungen. Während vor drei Jahren noch nicht einmal die Hälfte der Veröffentlichungen quantitativ ausgewertet wurde, waren es zwei Jahre später bereits über 90%. Der Anteil applikationsbezogener Methoden ist dabei am stärksten gestiegen, aber auch für empirische Methoden zeichnet sich ein deutlicher Trend nach oben ab.

Die Ergebnisse der Studie zeigen, dass Tanimoto, Dice und Kappa wie erwartet stark korrelieren (Tab. 1). Aus diesen drei wurde der Tanimoto-Koeffizient stellvertetend für alle Überlappungsmaße mit der durchschnittlichen und maximalen Distanz verglichen, wobei nur eine schwache Korrelation sichtbar wurde. Auch die beiden Distanzmaße sind untereinander nur mittelmäßig korreliert.

Abb. 2. Verwendete Evaluationsmethoden aus 40 Publikationen (a) und deren zeitliche Entwicklung (b).



(a) (b)

5 Diskussion

Von den in der Literatur beschriebenen Evaluationsverfahren sind für die medizinische Bildverarbeitung nicht alle gleichermaßen geeignet. Zu den aussagekräftigsten und gleichzeitig am vielseitigsten zu verwendenden zählen die empirischen Diskrepanzmethoden. Um eine hohe Akzeptanz dieser Verfahren bei Medizinern zu erreichen, sollten jedoch nur Vergleichsmaße mit direktem Bezug zur Wirklichkeit verwendet werden. Für die in diesem Beitrag vorgestellten Kategorien trifft diese Anforderung auf Überlappungsmaße und Oberflächendistanzen zu. Wie die Ergebnisse der Studie zeigen, sind zur umfassenden Evaluation von Segmentierungsergebnissen mindestens drei Vergleichsmaße nötig: Eines aus der Gruppe der Überlappungsmaße (z.B. Tanimoto-Koeffizient), die durchschnittliche Oberflächendistanz und die Hausdorff-Distanz. Alle drei Maße sind untereinander nur schwach bis mittelmäßig korreliert und berechnen damit jeweils unterschiedliche Aspekte der Übereinstimmung. Sie können somit als orthogonale Achsen eines einzigen Bewertungssystems aufgefasst werden, welches wir für einen umfassenden und nachvollziehbaren Vergleich von Segmentierungsergebnissen zur Benutzung empfehlen.

Literaturverzeichnis

1. Zhang YJ: A Survey on Evaluation Methods for Image Segmentation. Pattern Recognition 29(8):1335–1346, 1996.
2. van Rijsbergen CJ: Information Retrieval. Butterworths, London, 1979.
3. Dice LR: Measures of the Amount of Ecologic Associations between Species. Journal of Ecology 26, 1945.
4. Tanimoto TT: An Elementary Mathematical Theory of Classification and Prediction. IBM Research, 1958.
5. Veltkamp RC: Shape Matching: Similarity Measures and Algorithms. Procs. Shape Modelling International:188–197, 2001.