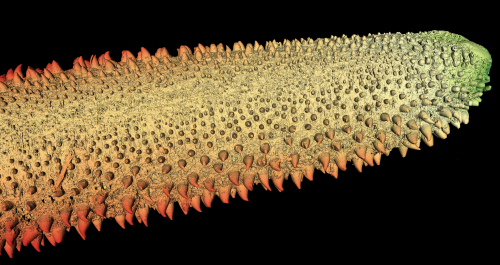
Identifizieren biologischer Strukturen

Ausgangssituation und Zielsetzung

Bei dieser Aufgabe geht es darum, eine Tätigkeit weitgehend zu automatisieren, die sehr zeitaufwändig ist, wenn sie von Hand durchgeführt wird: In einem großen Datensatz, der anhand computertomografischer Bilder von einem bestimmten biologischen Objekt erstellt wurde, sollen bestimmte Strukturen ausfindig gemacht werden. Bei dem untersuchten Objekt handelt es sich um die Spitze des schnabelartigen Fortsatzes am Kopf eines Segelfisches, das sogenannte Rostrum, und bei den ausfindig zu machenden Strukturen um sehr kleine Zähne, sogenannte Mikrozähne, die sich in großer Zahl auf dem Rostrum befinden. Die folgende Abbildung visualisiert einen Ausschnitt aus einem solchen Datensatz:



Das Problem dabei ist, dass keine genaue Definition der gesuchten Mikrozähne vorliegt. Für einen menschlichen Experten ist es dennoch prinzipiell sehr einfach, die gewünschten Strukturen zu identifizieren. Als Basis für die Entwicklung einer automatisierten Lösung stehen zwei Dateien zur Verfügung:

Ein Datensatz (data0.csv), der als zweidimensionale Matrix mit Höhenwerten dargestellt ist, wobei jede Zeile für einen Punkt auf der Längsachse des Rostrums steht und jede Spalte für einen Punkt auf dem Ring um die Längsachse herum. An die letzte Spalte schließt sich also unmittelbar wieder die erste Spalte an.

Eine Liste (labels0.csv) mit den Koordinaten der von einem Experten von Hand markierten Mikrozähne. Die erste Koordinate gibt die Zeile und die zweite die Spalte (jeweils beginnend mit 0) im Datensatz data0.csv an. Wir gehen davon aus, dass diese Liste korrekt und vollständig ist.

Alle Daten liegen im CSV-Format vor, wobei ein Komma als Trennzeichen verwendet wird.

Zur endgültigen Bestimmung der Qualität (Evaluation) wird zu einem späteren Zeitpunkt im Praktikum ein neuer frischer Datensatz (data1.csv und labels1.csv) zur Verfügung gestellt.

Bitte beachten: Alle Datensätze enthalten echte Daten, die nur für diese Lehrveranstaltung verwendet und nicht weitergegeben werden dürfen!

Konkrete Aufgabenstellung

1. Untersuchen Sie den gegebenen Datensatz 0 (also data0.csv) und versuchen Sie, Kriterien zu ermitteln, mit denen sich die zugehörigen Markierungen (labels0.csv) erklären lassen.
2. Entwickeln Sie einen Algorithmus, der die Markierung der Datensätze, also das Auffinden der Mikrozähne automatisiert.
3. Evaluieren Sie Ihren Algorithmus anhand der manuellen Label (labels0.csv) und bestimmen Sie Precision, Recall und F-Score für Ihren Algorithmus.
4. Verbessern Sie die Leistung ihres Algorithmus. Verwenden Sie für die Optimierung nur eine Teilmenge des Datensatzes 0, um auf den restlichen Daten eine möglichst unabhängige Evaluation durchführen zu können. Im Endergebnis sollte auf dem später ausgegebenen Datensatz 1 (also data1.csv) ein F-Score von mindestens 0,84 erreicht werden.

Der Recall ist der Quotient von der Anzahl der vom Algorithmus korrekt gefundenen Zähne und der Anzahl der tatsächlich vorhandenen Zähne.

Die Precision ist der Quotient von der Anzahl der vom Algorithmus korrekt gefundenen Zähne und der Gesamtanzahl der vom Algorithmus gefundenen Zähne.

Precision und Recall besitzen den Wertebereich 0...1.

Der F-Score ist der harmonische Mittelwert von Precision und Recall.

Ein Zahn gilt als "korrekt gefunden", wenn seine Koordinaten mit denen eines Zahns aus der Datei labels0.csv (bzw. labels1.csv) übereinstimmen oder einen Punkt in dessen Nachbarschaft markieren, der dieselbe Höhe besitzt. Bei einer Evaluation darf jeder Zahn aus der Datei labels0.csv (bzw. labels1.csv) jedoch höchstens einmal gezählt werden, d.h. weitere unmittelbare Nachbarpunkte dürfen nicht als weitere korrekt gefundene Zähne gezählt werden.

Bitte beachten Sie: Lösen Sie das Problem allgemein. Ihr Algorithmus soll die manuelle Markierung ersetzen und später auf unbekannte Datensätze zuverlässig angewendet werden können. Aus diesem Grund wird der Datensatz 1 auch erst später zur Verfügung gestellt.

Die Software ist in Java zu implementieren. Abzugeben sind der Quellcode der Java-Klassen sowie eine kurze schriftliche Darstellung der Ergebnisse als PDF-Dokument.

Weitere Hinweise

Denken Sie an eine software- und programmiertechnisch korrekte Form der Implementierung. Denken Sie auch an den angemessenen Einsatz von Datenstrukturen und Algorithmen. Zu einer tadellosen Lösung gehört auch eine aussagekräftige und vollständige Kommentierung des Quelltextes. Die Kommentierung sollte im Idealfall javadoc-fähig sein.

Die Bearbeitung der Aufgabe soll in Zweierteams erfolgen. Es wird vorausgesetzt, dass sich beide Teampartner mit dem Lösungsweg und auch mit der konkreten Implementierung der Lösung bestens auskennen.

Packen Sie alle zu Ihrer Lösung gehörenden Klassen sowie das PDF-Dokument in ein Zip-Archiv, benennen Sie es nach dem Schema vorname1.nachname1-vorname2.nachname2-ISys2.zip (Beispiel: stefan.krause-michael.breuker-ISys2.zip) und laden Sie das Archiv rechtzeitig im Lernraum-Kurs hoch. Verspätete Abgaben werden nicht berücksichtigt.