(19) Deutsches Patent- und Markenamt

(10) **DE 10 2013 207 402 B4** 2025.02.20

(12) # Patentschrift

(21) Aktenzeichen: **10 2013 207 402.6**
(22) Anmeldetag: **24.04.2013**
(43) Offenlegungstag: **30.10.2014**
(45) Veröffentlichungstag der Patenterteilung: **20.02.2025**

(51) Int Cl.: ***G01N 27/62*** (2021.01)
G06F 17/16 (2006.01)
G16B 45/00 (2019.01)

Innerhalb von neun Monaten nach Veröffentlichung der Patenterteilung kann nach § 59 Patentgesetz gegen das Patent Einspruch erhoben werden. Der Einspruch ist schriftlich zu erklären und zu begründen. Innerhalb der Einspruchsfrist ist eine Einspruchsgebühr in Höhe von 200 Euro zu entrichten (§ 6 Patentkostengesetz in Verbindung mit der Anlage zu § 2 Abs. 1 Patentkostengesetz).

(73) Patentinhaber:
Bruker Daltonics GmbH & Co. KG, 28359 Bremen, DE

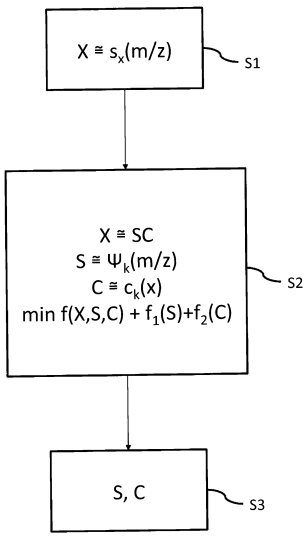
(74) Vertreter:
Fink Numrich Patentanwälte PartmbB, 81245 München, DE

(72) Erfinder:
Goldabaee, Mohammad, Dr., Paris, FR; Vandergheynst, Pierre, Prof. Dr., Ecublens, CH; Maaß, Peter, Prof. Dr., 28211 Bremen, DE; Kobarg, Jan Hendrik, 28213 Bremen, DE; Aleksandrov, Fedor, Dr., 28209 Bremen, DE

(56) Ermittelter Stand der Technik:
siehe Folgeseiten

(54) Bezeichnung: **Verfahren zum rechnergestützten Verarbeiten von räumlich aufgelösten Massenspektrometriedaten**

(57) Hauptanspruch: Verfahren zum rechnergestützten Verarbeiten von räumlich aufgelösten Massenspektrometriedaten mit einer Vielzahl von jeweiligen Messpositionen (x) zugeordneten Spektren, wobei ein Spektrum ($s_x(m/z)$) für mehrere Bildkanäle (m/z) jeweilige Intensitäten (I) umfasst, welche eine für den entsprechenden Bildkanal (m/z) detektierte physikalische Größe repräsentieren, wobei die Bildkanäle (m/z) Masse-zu-Ladungs-Verhältnisse darstellen und die für den entsprechenden Bildkanal (m/z) detektierte physikalische Größe die Anzahl der für das entsprechende Masse-zu-Ladungs-Verhältnis detektierten Moleküle repräsentiert, bei dem:
- die Massenspektrometriedaten mittels einer nicht-negativen Matrix-Faktorisierung zerlegt werden, gemäß der eine Datenmatrix (X), welche für jeweilige Messpositionen (x) die Intensitäten (I) für die Bildkanäle (m/z) des der jeweiligen Messposition (x) zugeordneten Spektrums ($s_x(m/z)$) umfasst, als ein Matrixprodukt einer Komponenten-Matrix (S) und einer Koeffizienten-Matrix (C) dargestellt wird, wobei die Komponenten-Matrix (S) für eine vorbestimmte Anzahl von Komponenten (k) jeweils Intensitäten eines der jeweiligen Komponente (k) zugeordneten Spektrums (Ψ_k (m / z)) für die Bildkanäle (m/z) umfasst und wobei die Koeffizienten-Matrix (C) für die vorbestimmte Anzahl von Komponenten (k) jeweils Koeffizienten für die Messpositionen (x) umfasst, wobei die nicht-negative Matrix-Faktorisierung derart ausgestaltet ist, dass die zerlegten Massenspektrometriedaten im Vergleich zu den Massenspektrometriedaten vor der Zerlegung eine geringere Datenmenge aufweisen;
- basierend auf der nicht-negativen Matrix-Faktorisierung eine Optimierung einer Kostenfunktion in der Form einer Maximierung oder Minimierung der Kostenfunktion in Abhängigkeit von der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) als Variablen durchgeführt wird, wobei die Kostenfunktion ein Funktional (f(X, S, C)) umfasst, das von der Datenmatrix (X), der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) abhängt;
- bei der Optimierung der Kostenfunktion eine erste und eine zweite Nebenbedingung einfließt, wobei gemäß der ersten Nebenbedingung ...



(56) Ermittelter Stand der Technik:

US	**2005 / 0 047 670**	**A1**
US	**2010 / 0 299 379**	**A1**
US	**2011 / 0 007 819**	**A1**

CN101487892A: Auszug aus den Datenbanken WPI/Thomson und EPODOC/EPO

GOLBABAEE, Mohammad ; ARBERET, Simon ; VANDERGHEYNST, Pierre: Distributed compressed sensing of hyperspectral images via blind source separation. In: Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the forty-fourth asilomar conference on, 7-10 Nov. 2010, S. 196-198. - ISBN 978-1-4244-9722-5

LEE, Daniel D. ; SEUNG, H. Sebastian: Learning the parts of objects by non-negative matrix factorization. In: Nature, Vol. 401, 1999, S. 788-791

RUDIN, Leonid I. ; OSHER, Stanley ; FATEMI, Emad: Nonlinear total variation based noise removal algorithms. In: Physica D, Vol. 60, 1992, S. 259-268. - ISSN 0167-2789

TREDE, Dennis [u.a.]: On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data. In: Journal of Integrative Bioinformatics, Vol. 9, 2012, No. 1, S.189-1-189-10

Beschreibung

[0001] Die Erfindung betrifft ein Verfahren zum rechnergestützten Verarbeiten von Massenspektrometriedaten, sowie ein entsprechendes Computerprogrammprodukt und ein entsprechendes Computerprogramm.

[0002] Die Erfindung betrifft das Gebiet der Bildverarbeitung von räumlich aufgelösten Hyperspektraldaten, welche sich dadurch auszeichnen, dass sie für mehr als einen Bildkanal Bildinformationen enthalten. Der Anwendungsfall der Erfindung sind räumlich aufgelöste Massenspektrometriedaten, bei denen die jeweiligen Bildkanäle Masse-zu-Ladungs-Verhältnisse repräsentieren. Bei der räumlich aufgelösten Massenspektrometrie (engl.: imaging mass spectrometry) werden für eine Vielzahl von Messpositionen der untersuchten Probe die darin enthaltenen Bestandteile in ionisierter Form herausgelöst und anschließend in einem Massenspektrometer in Abhängigkeit von deren Masse-zu-Ladungs-Verhältnis detektiert. Ein bekanntes Verfahren der räumlich aufgelösten Massenspektrometrie ist die MALDI-Massenspektrometrie, bei der auf die Probe eine Matrix aufgebracht wird und die Moleküle der Probe über Laserpulse abgelöst werden (MALDI = Matrix-Assisted Laser Desorption/Ionization). Um aus den Massenspektrometriedaten aussagekräftige Informationen zu erhalten, sind aus dem Stand der Technik verschiedene Analyseverfahren bekannt, wie z.B. Peak Picking, Clustering, Klassifikation und dergleichen (siehe Dokument [1]).

[0003] Bei der Verarbeitung von räumlich aufgelösten Hyperspektraldaten besteht das Problem, dass eine große Menge an Daten für jede Messposition und eine Vielzahl von Bildkanälen erzeugt werden. Die Speicherung und die Verarbeitung dieser Daten sind aufwändig.

[0004] Die Druckschrift US 2010/0299379 A1 offenbart die Zerlegung von multivariaten Daten mittels einer nicht-negativen Matrix-Faktorisierung basierend auf der Minimierung einer Kostenfunktion.

[0005] In dem Dokument CN101487892A wird die Zerlegung von Hyperspektraldaten mittels eines Faktor-Analyse-Modells beschrieben, um die Dimension der Hyperspektraldaten zu verringern.

[0006] Die Druckschriften US 2005/0047670 A1 sowie US 2011/0007819 A1 beschreiben weitere Verfahren zur Kompression von Hyperspektraldaten, wobei die Hyperspektraldaten räumlich aufgelöste Daten, beispielsweise aufgenommen von einem Satelliten, sind.

[0007] Aufgabe der Erfindung ist es, ein Verfahren zum rechnergestützten Verarbeiten von räumlich aufgelösten Massenspektrometriedaten zu schaffen, welches eine Zerlegung der Daten in charakteristische Spektren und deren räumliche Verteilung bei gleichzeitiger Datenkompression ermöglicht.

[0008] Diese Aufgabe wird durch das Verfahren gemäß Patentanspruch 1 bzw. den maschinenlesbaren Träger gemäß Patentanspruch 13 gelöst. Weiterbildungen der Erfindung sind in den abhängigen Ansprüchen definiert.

[0009] Das erfindungsgemäße Verfahren dient zum rechnergestützten Verarbeiten von räumlich aufgelösten Massenspektrometriedaten, welche eine Vielzahl von jeweiligen Messpositionen zugeordneten Spektren umfassen, wobei ein Spektrum für mehrere Bildkanäle jeweilige Intensitäten umfasst, welche eine für den entsprechenden Bildkanal detektierte physikalische Größe repräsentieren. Dabei stellen die Bildkanäle Masse-zu-Ladungs-Verhältnisse dar, und die für den entsprechenden Bildkanal detektierte physikalische Größe repräsentiert die Anzahl der für das entsprechende Masse-zu-Ladungs-Verhältnis detektierten Moleküle. In einer bevorzugten Variante sind die Massenspektrometriedaten die bereits oben erwähnten MALDI-Massenspektrometriedaten. Nichtsdestotrotz können die räumlich aufgelösten Massenspektrometriedaten auch mit einem anderen Verfahren, z.B. mittels DESI-Massenspektrometrie (DESI = Desorption Electrospray Ionization) oder SIMS (= Secondary Ion Mass Spectrometry), gewonnen sein.

[0010] Im Rahmen des erfindungsgemäßen Verfahrens werden die Massenspektrometriedaten mittels einer nicht-negativen Matrix-Faktorisierung zerlegt. Dazu wird eine Datenmatrix, welche für jeweilige Messpositionen die Intensitäten für die Bildkanäle des der jeweiligen Messposition zugeordneten Spektrums umfasst, als ein Matrixprodukt einer Komponenten-Matrix und einer Koeffizienten-Matrix dargestellt. In der Datenmatrix werden die jeweiligen Messpositionen z.B. durch Spalten oder gegebenenfalls auch Zeilen repräsentiert. Die Parameter der nicht-negativen Matrix-Faktorisierung sind derart gewählt, dass die hierdurch erzeugte Datenmenge geringer ist als die der ursprünglichen Massenspektrometriedaten.

[0011] Die Komponenten-Matrix umfasst für eine vorbestimmte Anzahl von Komponenten jeweils Intensitäten eines der jeweiligen Komponente zugeordneten bzw. die jeweilige Komponente charakterisierenden Spektrums aus Intensitäten für die Bildkanäle. Die einzelnen Komponenten werden durch die Spalten der Komponenten-Matrix bzw. gegebenenfalls auch durch deren Zeilen repräsentiert. Die Koeffizienten-Matrix umfasst für die vorbestimmte Anzahl von Komponenten jeweils Koeffizienten für die Vielzahl von Messpositionen, wobei die Komponenten

vorzugsweise durch die Zeilen bzw. gegebenenfalls auch durch die Spalten der Matrix repräsentiert werden. Die Komponenten-Matrix enthält somit die charakterisierenden Spektren der Massenspektrometriedaten, und die Koeffizienten-Matrix repräsentiert die räumliche Verteilung dieser Spektren. Aufgrund der nicht-negativen Matrix-Faktorisierung sind alle Einträge der Komponenten-Matrix und der Koeffizienten-Matrix größer gleich Null.

[0012] Im erfindungsgemäßen Verfahren wird basierend auf der nicht-negativen Matrix-Faktorisierung eine Optimierung einer Kostenfunktion in Abhängigkeit von der Komponenten-Matrix und der Koeffizienten-Matrix als Variablen durchgeführt, d.h. bei der Optimierung werden die Einträge der Komponenten-Matrix bzw. der Koeffizienten-Matrix variiert. Die Optimierung kann dabei eine Minimierung und ggf. auch eine Maximierung der Kostenfunktion sein.

[0013] Die Kostenfunktion umfasst ein Funktional, das von der Datenmatrix, der Komponenten-Matrix und der Koeffizienten-Matrix abhängt. Bei der Optimierung der Kostenfunktion fließt eine erste und eine zweite Nebenbedingung ein, wobei gemäß der ersten Nebenbedingung eine Optimierung der Kostenfunktion hin zu einer möglichst geringen Anzahl von Einträgen der Komponenten-Matrix ungleich Null (d.h. hin zu einer möglichst großen Anzahl von Einträgen gleich Null) erfolgt und wobei gemäß der zweiten Nebenbedingung eine Optimierung der Kostenfunktion hin zu einer möglichst geringen Anzahl von Einträgen der Koeffizienten-Matrix ungleich Null (d.h. hin zu einer großen Anzahl von Einträgen gleich Null) erfolgt. Die durch die nicht-negative Matrix-Faktorisierung erzeugten Einträge der Komponenten-Matrix und der Koeffizienten-Matrix werden schließlich gespeichert.

[0014] Erfindungsgemäß werden in der an sich bekannten Methodik der nicht-negativen Matrix-Faktorisierung neue Nebenbedingungen eingeführt, welche eine spärliche Besetzung der Komponenten-Matrix und der Koeffizienten-Matrix definieren. Mit der nicht-negativen Matrix-Faktorisierung kann insbesondere eine hohe Kompression der ursprünglichen Massenspektrometriedaten erreicht werden, wobei gleichzeitig aussagekräftige Informationen in den Massenspektrometriedatensätzen erhalten bleiben, die z.B. basierend auf der weiter unten beschriebenen Visualisierung von den Komponenten zugeordneten Spektren bzw. Koeffizienten dargestellt werden können.

[0015] Die Optimierung der Kostenfunktion im Rahmen der im erfindungsgemäßen Verfahren verwendeten nicht-negativen Matrix-Faktorisierung kann mit an sich bekannten mathematischen Verfahren durchgeführt werden, wobei Beispiele solcher Verfahren in der detaillierten Beschreibung genannt sind. Da die Verfahren zum Stand der Technik gehören und dem Fachmann geläufig sind, werden diese nicht weiter im Detail beschrieben.

[0016] In einer Variante des erfindungsgemäßen Verfahrens, welche nicht in den Ansprüchen definiert ist, ist die nicht-negative Matrix-Faktorisierung derart ausgestaltet, dass eine vorgegebene Genauigkeit der Rekonstruktion der Massenspektrometriedaten aus den zerlegten Massenspektrometriedaten oder eine vorgegebene Datengröße der zerlegten Massenspektrometriedaten erreicht wird. Ferner kann das oben beschriebene Funktional ggf. durch eine Zufallsmatrix erweitert werden, so dass die Massenspektrometriedaten über eine randomisierte Linearkombination reduziert werden.

[0017] In einer besonders bevorzugten Ausführungsform umfasst das in der Kostenfunktion enthaltene Funktional die L2-Norm der Differenz zwischen der Datenmatrix und dem Matrixprodukt aus der Komponenten-Matrix und der Koeffizienten-Matrix, wobei die Optimierung der Kostenfunktion hin zu möglichst kleinen L2-Normen erfolgt.

[0018] In einer weiteren, besonders bevorzugten Ausführungsform wird die erste Nebenbedingung durch eine erste Funktion repräsentiert, welche Bestandteil der Kostenfunktion ist und zu dem Funktional addiert wird, so dass die Optimierung der Kostenfunktion zur Bevorzugung einer Komponenten-Matrix mit einer geringen Anzahl von Einträgen ungleich Null führt. Vorzugsweise umfasst die erste Funktion die L1-Norm der Komponenten-Matrix. Ebenso kann die erste Funktion eine Summe von Normen der Komponenten-Matrix sein, wobei die Summe vorzugsweise die L1-Norm und die L2-Norm der Komponenten-Matrix umfasst. Die Verwendung einer Summe von Normen in einem Funktional ist an sich bekannt und wird deshalb nicht weiter im Detail beschrieben.

[0019] In einer weiteren Ausgestaltung des erfindungsgemäßen Verfahrens wird die zweite Nebenbedingung durch eine zweite Funktion repräsentiert, welche Bestandteil der Kostenfunktion ist und zu dem Funktional addiert wird, so dass die Optimierung zu einer Bevorzugung einer Koeffizienten-Matrix mit einer möglichst geringen Anzahl von Einträgen ungleich Null führt. Insbesondere kann als zweite Funktion die TV-Norm (TV = Total Variation) der Koeffizienten-Matrix und ggf. auch die L1-Norm der Koeffizienten-Matrix verwendet werden. Die TV-Norm ist an sich aus dem Stand der Technik bekannt und insbesondere auch für zweidimensionale Bilder definiert, wie durch das Dokument [4] belegt ist. Die Koeffizienten-Matrix kann dabei als ein entsprechendes Bild aufgefasst werden, für das die TV-Norm ermittelt wird. Bei der TV-Norm werden in einem Bild die Sprunghöhe und Kantenlänge bewertet. Die

Verwendung dieser Norm in der Kostenfunktion gewährleistet somit glatte Kanten in dem entsprechenden Bild der Koeffizienten-Matrix.

[0020] In einer weiteren Ausgestaltung des erfindungsgemäßen Verfahrens wird die Koeffizienten-Matrix als 2D-Wavelets repräsentiert, wobei als zweite Nebenbedingung eine begrenzte Anzahl von 2D-Wavelet-Koeffizienten festgelegt wird. Die Verwendung einer solchen zweiten Nebenbedingung ist an sich aus dem Stand der Technik bekannt und insbesondere in der Druckschrift [5] beschrieben. Die entsprechende Wavelet-Repräsentation mit den Nebenbedingungen findet sich in Gleichung (4) dieses Dokuments. Erfindungsgemäß wird zusätzlich die entsprechende erste Nebenbedingung in der Optimierung mit berücksichtigt. Ggf. kann die zweite Nebenbedingung auch derart ausgestaltet sein, dass die L1-Norm der 2D-Wavelet-Koeffizienten möglichst klein wird.

[0021] In einer weiteren, besonders bevorzugten Ausführungsform wird die Anzahl von Komponenten der nicht-negativen Matrix-Faktorisierung derart festgelegt, dass die Gesamtanzahl der gespeicherten Einträge der Komponenten-Matrix und der Koeffizienten-Matrix einem vorgegebenen Prozentsatz der Anzahl der Einträge der Datenmatrix entspricht, wobei der vorgegebene Prozentsatz insbesondere zwischen 1% und 10% der Anzahl der Einträge der Datenmatrix liegt. Durch geeignete Wahl der Anzahl der Komponenten kann ggf. auch eine Datenreduktion mit dem Faktor 10^3 oder mehr ermöglicht werden. Beispielsweise kann die Anzahl von Komponenten zwischen 3 und 20 bzw. zwischen 5 und 10 liegen. Die Anzahl der Komponenten hängt dabei von der Datengrundlage ab und kann auch andere Werte annehmen.

[0022] In einer weiteren, besonders bevorzugten Ausführungsform wird aus den gespeicherten Einträgen der Komponenten-Matrix und der Koeffizienten-Matrix eine Visualisierung auf einem Display derart generiert, dass für zumindest einen Teil der Komponenten jeweils ein Bild aus die Messpositionen repräsentierenden Bildpunkten erzeugt wird, wobei ein jeweiliger Bildpunkt den Koeffizienten für die jeweilige Komponente der Koeffizienten-Matrix und die durch den jeweiligen Bildpunkt repräsentierte Messposition z.B. farblich codiert. Der Begriff des Bildpunkts ist dabei weit zu verstehen und kann ein Pixel bzw. auch mehrere benachbarte Pixel auf dem Display repräsentieren. Wie sich aus der speziellen Beschreibung ergibt, werden mit dieser Ausführungsform des erfindungsgemäßen Verfahrens aussagekräftige Bilder zu der räumlichen Verteilung von verschiedenen Stoffgruppen in der durch die Massenspektrometrie analysierten Probe erhalten. Mit anderen Worten wird eine Softsegmentierung erreicht, bei der den einzelnen Gewebebereichen nicht fest eine Gewebestruktur zugeordnet wird, sondern für die verschiedenen Gewebebereiche der relative Anteil der entsprechenden Komponente angegeben ist.

[0023] In einer weiteren Ausführungsform des erfindungsgemäßen Verfahrens wird aus den gespeicherten Einträgen der Komponenten-Matrix und der Koeffizienten-Matrix eine Visualisierung auf einem Display derart generiert, dass für zumindest einen Teil der den Komponenten zugeordneten Spektren jeweils eine graphische Darstellung erzeugt wird, welche die Intensitäten des der jeweiligen Komponente zugeordneten Spektrums in Abhängigkeit von den Bildkanälen, d.h. den Masse-zu-Ladungs-Verhältnissen, anzeigt. Auf diese Weise wird die Information erhalten, wie hoch der Anteil bestimmter Zusammensetzungen in den entsprechenden Komponenten ist. Dies bildet z.B. die Grundlage für eine nachfolgende Biomarker-Identifikation.

[0024] In einer weiteren Variante des erfindungsgemäßen Verfahrens werden aus den gespeicherten Einträgen der Komponenten-Matrix und der Koeffizienten-Matrix durch Matrixmultiplikation dieser Matrizen die Massenspektrometriedaten rekonstruiert. Das erfindungsgemäße Verfahren gewährleistet dabei eine nahezu verlustlose Rekonstruktion der ursprünglichen Daten. Dabei kann eine Toleranz δ vorgegeben werden und abhängig davon die Parameter der nicht-negativen Matrix-Faktorisierung derart gewählt werden, dass die Abweichung der Rekonstruktion von den ursprünglichen Daten geringer als die Toleranz δ ist.

[0025] Die Erfindung betrifft darüber hinaus einen maschinenlesbaren Träger mit darauf gespeichertem Programmcode zur Durchführung des erfindungsgemäßen Verfahrens bzw. einer oder mehrerer bevorzugter Varianten des erfindungsgemäßen Verfahrens, wenn der Programmcode auf einem Computer ausgeführt wird.

[0026] Die Erfindung betrifft in einer Ausgestaltung, welche nicht in den Ansprüchen definiert ist, ferner ein Computerprogramm mit einem Programmcode zur Durchführung des erfindungsgemäßen Verfahrens bzw. einer oder mehrerer bevorzugter Varianten des erfindungsgemäßen Verfahrens, wenn der Programmcode auf einem Computer ausgeführt wird.

[0027] Ausführungsbeispiele der Erfindung werden nachfolgend anhand der beigefügten Figuren detailliert beschrieben.

[0028] Es zeigen:

Fig. 1 ein Diagramm, welches beispielhaft mittels MALDI-Spektrometrie erfasste Daten wiedergibt, die mit dem erfindungsgemäßen Verfahren verarbeitet werden;

Fig. 2 ein Diagramm, welches die wesentlichen Schritte einer Ausführungsform des erfindungsgemäßen Verfahrens verdeutlicht;

Fig. 3 und **Fig.** 4 Schnittdarstellungen eines Rattengehirns, welche basierend auf einer Ausführungsform des erfindungsgemäßen Verfahrens generiert wurden; und

Fig. 5 Diagramme von charakteristischen Spektren eines Rattengehirns, welche mit einer Ausführungsform des erfindungsgemäßen Verfahrens generiert wurden.

[0029] Im Folgenden wird eine Ausführungsform des erfindungsgemäßen Verfahrens basierend auf sog. MALDI-Massenspektrometriedaten beschrieben. Im Rahmen dieser Spektrometrie werden die Moleküle der zu analysierenden Gewebeprobe mit einer Matrix-Lösung überzogen, und ein Laserstrahl (z.B. eines Stickstoff-Lasers) wird auf verschiedene Messpositionen des entsprechenden Probenträgers gerichtet. Über die Laserpulse des Laserstrahls erfolgt eine Teilchenablösung von ionisierten Molekülen des Gewebes, deren Masse-zu-Ladungs-Verhältnis in an sich bekannter Weise mit einem Massenspektrometer erfasst wird. Hierzu kann z.B. ein TOF-Massenspektrometer (TOF = Time of Flight) oder FTICR-Massenspektrometer (FTICR = Fourier Transform Ion Cyclotron Resonance) genutzt werden. Das Verfahren der Erfindung ist jedoch auch auf andere Massenspektrometriedaten anwendbar, sofern die Daten ortsaufgelöst an verschiedenen Messpositionen Spektren enthalten, die in Abhängigkeit von dem Masse-zu-Ladungs-Verhältnis die Anzahl von mit dem Massenspektrometer detektierten Ionen wiedergeben. Ferner kann das Verfahren der Erfindung auch allgemein auf Hyperspektraldaten angewendet werden (z.B. Raman-Spektroskopie, Remote Sensing), sofern die Daten ortsaufgelöst an verschiedenen Messpositionen Intensitäts-Vektoren mit Einträgen für verschiedene Bildkanäle enthalten. Eine Intensität kann dabei z.B. den Anteil reflektierten Lichtes einer bestimmten Wellenlänge umfassen, wobei die Wellenlänge den Bildkanal darstellt.

[0030] In den nachfolgend beschriebenen Ausführungsformen wurden mit der räumlich aufgelösten MALDI-Massenspektrometrie Schichten eines Rattengehirns untersucht. **Fig.** 1 zeigt typische Spektren, die mittels MALDI-Spektrometrie für einen Gewebeschnitt aus einem Rattengehirn aufgenommen wurden. Entlang der Achse m/z ist das Masse-zu-Ladungs-Verhältnis in Dalton angegeben. Ferner codiert die Achse x entsprechende zweidimensionale Positionen des Schnitts des Rattengehirns. Dabei sind beispielhaft für drei Positionen entsprechende Spektren s_1, s_2 und s_3 wiedergegeben, welche jeweils die Anzahl der detektierten ionisierten Moleküle in der Form eines Intensitätswerts I für entsprechende Masse-zu-Ladungs-Verhältnisse m/z angeben. In Abhängigkeit von der Auflösung der MALDI-Spektrometrie wird durch einen m/z-Wert ein bestimmtes Intervall von m/z-Werten um diesen Wert zusammengefasst. Die mit der MALDI-Spektrometrie erfassten Daten können anschließend in geeigneter Weise visualisiert werden, z.B. können Bilder des Rattengehirns für jeweilige m/z-Werte generiert werden, wobei in einem Bild für die unterschiedlichen Messpositionen die Intensität für den entsprechenden m/z-Wert farbcodiert ist. In **Fig.** 1 sind beispielhaft zwei m/z-Werte bei 4966 Da und 6717 Da angedeutet, für welche durch Codierung der entsprechenden Intensitäten für die unterschiedlichen Messpositionen solche Bilder generiert werden können.

[0031] Bei der räumlich aufgelösten Massenspektrometrie erweist es sich als nachteilhaft, dass eine sehr große Menge an Daten generiert wird. In dem dargestellten Beispiel besteht ein Datensatz für einen Schnitt durch ein Rattengehirn aus 20000 Spektren mit jeweils 8000 spektralen m/z-Werten, d.h. insgesamt aus 160 Millionen Werten. Für gewöhnlich erreichen räumlich aufgelöste Massenspektrometrie-Daten deutlich größere Datenmengen, sowohl durch eine wesentlich höhere Zahl gemessener Spektren (z.B. Faktor 100 bis 1000) als auch durch besser aufgelöste m/z-Werte (z.B. Faktor 2 bis 2000). Ein Ziel der nachfolgend beschriebenen Ausführungsform des erfindungsgemäßen Verfahrens ist die Komprimierung dieser Datenmenge mit geringem Informationsverlust.

[0032] Die wesentlichen Schritte des hier beschriebenen erfindungsgemäßen Verfahrens sind in **Fig.** 2 dargestellt. Das Verfahren geht aus von einem Schritt S1, in dem die oben beschriebenen MALDI-Massenspektrometriedaten in der Form einer Matrix X angeordnet werden. Die Matrix codiert für jeweilige Messpositionen x das entsprechende Spektrum s_x (m/z) der Intensitätswerte für die unterschiedlichen Masse-zu-Ladungs-Verhältnisse m/z. Die Matrix X umfasst N Zeilen, welche mit i = 1, ..., N indiziert werden, wobei jede Zeile einen m/z-Wert repräsentiert. Demgegenüber sind in der Matrix X M Spalten vorgesehen, welche mit j = 1, ..., M indiziert werden, wobei jede Spalte eine Messposition bzw. Bildposition repräsentiert. Die Bildpositionen beziehen sich dabei auf zweidimensionale Positionen x eines Schnitts durch ein Rattengehirn bzw. gegebenenfalls auch auf unterschiedliche Schnitte eines Rattengehirns, wobei in diesem Fall die Position einen dreidimensionalen Wert darstellt, der neben der Position innerhalb eines bestimmten Schnitts auch angibt, um welchen Schnitt es sich handelt.

[0033] In einem Schritt S2 der **Fig.** 2 wird die Datenmatrix X der Massenspektrometriedaten einer nichtnegativen Matrix-Faktorisierung unterzogen. Diese

nicht-negative Matrix-Faktorisierung zeichnet sich dadurch aus, dass Nebenbedingungen einfließen, gemäß denen die mit der Faktorisierung bestimmten Matrizen spärlich besetzt sind, d.h. wenige Einträge ungleich Null aufweisen, wie im Folgenden noch näher erläutert wird. Im Rahmen der nicht-negativen Matrix-Faktorisierung werden die Spektren $s_x(m/z)$ der Matrix X wie folgt durch Komponenten bzw. Basisspektren ψ_k (m / z), k = 1,...., K , approximiert:

$$s_x(m/z) = \sum_{k=1}^{K} c_k(x)\psi_k(m/z). \qquad (1)$$

[0034] Mit anderen Worten wird gemäß der nicht-negativen Matrix-Faktorisierung die Matrix X dargestellt aus einem Produkt der Matrizen S und C, welche nur positive Einträge aufweisen dürfen. Die Matrix S weist dabei die Dimension N×K auf, wobei die Zeilen die m/z-Werte und die Spalten die Komponenten k darstellen und jeder Eintrag einen Intensitätswert des entsprechenden Basisspektrums ψ_k(m / z) repräsentiert. Demgegenüber ist die Matrix C eine Koeffizienten-Matrix, welche die Dimension K×M aufweist. Die Zeilen dieser Matrix repräsentieren dabei die Komponenten k = 1, ..., K und die Spalten die Messpositionen. Die Werte $c_k(x)$ stellen die Einträge und damit die Koeffizienten dieser Matrix dar.

[0035] Gemäß der hier beschriebenen Ausführungsform wird eine nicht-negative Matrix-Faktorisierung verwendet, welche auf der Minimierung einer Kostenfunktion von drei Funktionen f, f_1 und f_2 basiert. Das heißt, die nicht-negative Matrix-Faktorisierung minimiert folgenden Ausdruck:

$$\min f(X,S,C) + f_1(S) + f_2(C). \qquad (2)$$

[0036] Dabei wird als Funktional f(X, S, C) die euklidische bzw. L2-Norm der Differenz zwischen der Datenmatrix X und dem Matrixprodukt SC verwendet, d.h. es gilt:

$$f(X,S,C) = \|X - SC\|^2. \qquad (3)$$

[0037] Die Funktionen f_1 und f_2 stellen Strafterme der, welche Lösungen mit spärlich besetzten Matrizen S bzw. C bevorzugen. Das heißt, der Term f_1 ist derart festgelegt, dass im Rahmen der Minimierung der Kostenfunktion Matrizen S bevorzugt werden, welche eine geringe Anzahl von Einträgen ungleich Null und damit eine große Anzahl von Einträgen gleich Null aufweisen. Analog ist die Funktion f_2 derart definiert, dass im Rahmen der Minimierung der Kostenfunktion Matrizen C bevorzugt werden, welche eine geringe Anzahl von Einträgen ungleich Null und damit eine große Anzahl von Einträgen gleich Null aufweisen. Mit diesen Straftermen wird erreicht, dass die komprimierte Datenmenge, welche durch die Matrizen S und C repräsentiert wird, wichtige Informationen aus den ursprünglichen Massenspektrometriedaten enthält, wie weiter unten anhand der mit der komprimierten Datenmenge erzeugten Visualisierungen erläutert wird.

[0038] In der hier beschriebenen Ausführungsform wird als Strafterm f_1 die L1-Norm verwendet, d.h. es gilt:

$$f_1(S) = \|S\|_{\ell 1}. \qquad (4)$$

[0039] Demgegenüber beschreibt der Strafterm f_2 (C) die sog. TV-Norm (TV = Total Variation), d.h. es gilt:

$$f_2(C) = |C|_{TV}. \qquad (5)$$

[0040] Die TV-Norm ist an sich aus dem Stand der Technik bekannt, wie durch die bereits oben erwähnten Druckschrift [4] belegt ist.

[0041] Das oben beschriebene Minimierungsproblem gemäß Gleichung (2) kann mit an sich bekannten Verfahren gelöst werden, wie z.B. einer sog. Blind-Source-Separation mit L1-Nebenbedingungen bzw. mit einem OMP-Verfahren (OMP = Orthogonal Matching Pursuit) oder ähnlichen Verfahren des sog. Dictionary Learnings. Insbesondere kann z.B. das IHT-Verfahren (IHT = Iterative Hard Thresholding, siehe Druckschrift [2]) bzw. ein multiplikativer Algorithmus gemäß der Druckschrift [3] eingesetzt werden.

[0042] Die oben beschriebenen Nebenbedingungen, gemäß denen Lösungen mit spärlich besetzten Matrizen S bzw. C bevorzugt werden, können gegebenenfalls auch auf andere Weise bzw. mit anderen Funktionen f_1 und f_2 realisiert werden. Die Nebenbedingung in Bezug auf die spärliche Besetzung der Matrix C kann z.B. auch über eine Darstellung der Matrix C in einer Wavelet-Basis mit entsprechenden Bedingungen an die Spärlichkeit erreicht werden, wie in der Druckschrift [5] (siehe dort insbesondere Gleichung (4)) beschrieben ist.

[0043] Als Ergebnis der Matrix-Faktorisierung gemäß Schritt S2 der **Fig.** 2 erhält man schließlich die beiden Matrizen S und C, welche in einem Schritt S3 in einem entsprechenden Speicher abgelegt werden. Durch geeignete Wahl der Anzahl K an Komponenten k kann hierdurch eine beträchtliche Datenreduktion im Vergleich zur ursprünglichen Datenmatrix X erreicht werden. In einer Ausführungsform des Verfahrens, bei dem K = 6 Komponenten verwendet werden, ergeben sich für das oben beschriebene Beispiel von 20000 Spektren mit 8000 m/z-Werten für die Matrix S 8000×6 und für die Matrix C 20000×6 Werte, d.h. insgesamt 168000 Werte, im Vergleich zu 160 Millionen Werten in der ursprünglichen Datenmatrix X. Somit kann eine Datenkompression mit einem Faktor von ungefähr 950 erreicht werden. Dabei bleiben jedoch die wesentlichen Informationen

aus den Spektren in den Matrizen S und X erhalten, was durch die oben beschriebenen Strafterme sichergestellt wird.

[0044] In einer Variante des erfindungsgemäßen Verfahrens können bei Bedarf die ursprünglichen Massenspektrometriedaten durch Multiplikation der Matrizen S und C mit geringen Verlusten wiederhergestellt werden. In einer weiteren, besonders bevorzugten Ausführungsform werden jeweilige Zeilen der Koeffizienten-Matrix C und damit entsprechende Komponenten k visualisiert. Dabei wird das Gewebe auf einem Display über eine Vielzahl von Bildpunkten repräsentiert, wobei jeder Bildpunkt einer Messposition x entspricht und den Wert des Koeffizienten der Komponente k aus der Matrix C codiert. Dies ist anhand von **Fig.** 3 und **Fig.** 4 für MALDI-Massenspektrometriedaten eines Schnitts durch ein Rattengehirn angedeutet.

[0045] **Fig.** 3 zeigt den Schnitt eines Rattengehirns für die Komponente k = 1 und **Fig.** 4 den Schnitt des Rattengehirns für die Komponente k = 2. Dabei wurde die nicht-negative Matrix-Faktorisierung für sechs Komponenten (K = 6) auf MALDI-Massenspektrometriedaten des entsprechenden Schnitts des Rattengehirns angewendet. In **Fig.** 3 und **Fig.** 4 werden dabei durch dunklere Farben höhere Werte der Koeffizienten codiert. Wie man aus **Fig.** 3 erkennt, wird durch die Komponente k = 1 der Bereich CO des Cortex des Rattengehirns bzw. die Bereiche H1 und H2 des Hippocampus beschrieben. Demgegenüber werden in **Fig.** 4 durch die Komponente k = 2 Gewebebereiche T1 und T2 beschrieben, welche dem Thalamus des Rattengehirns entsprechen. Aus **Fig.** 3 und **Fig.** 4 wird somit ersichtlich, dass durch die Darstellung entsprechender Zeilen der Koeffizienten-Matrix C biologisch sinnvolle Bereiche des analysierten Gewebes repräsentiert werden. Dabei wird eine Softsegmentierung des Gewebes erreicht, indem die Relevanz der entsprechenden Basisspektren bzw. Komponenten an den unterschiedlichen Messpositionen über die Darstellung der Koeffizienten codiert wird. Dies folgt daraus, dass die Koeffizienten angeben, wie stark eine entsprechende Komponente k zu dem an der Position x gemessenen Spektrum $s_x(m/z)$ der ursprünglichen Massenspektrometriedaten beiträgt. Somit wird durch die Repräsentation der Koeffizienten mittels Bildpunkten gemäß **Fig.** 3 und 4 die charakteristische metabolische Struktur von Regionen angezeigt, welche eine molekulare Zusammensetzung aufweisen, die der Komponente k = 1 bzw. k = 2 entspricht.

[0046] In einer weiteren Ausführungsform des erfindungsgemäßen Verfahrens werden für zumindest einen Teil der Komponenten k die Werte der entsprechenden Spalte der Matrix S in einer graphischen Darstellung wiedergegeben. Dies ist in **Fig.** 5 für die Komponente k = 1 und k = 2 dargestellt. Man erhält somit ein einzelnes charakteristisches Spektrum für eine Komponente k, das für die verschiedenen m/z-Werte die entsprechenden Intensitäten I wiedergibt. Auf diese Weise wird die Information dargestellt, welche Molekülzusammensetzungen in den verschiedenen Komponenten dominierend sind. Diese Information kann z.B. im Rahmen der Identifikation von Proteinen bzw. Biomarkern genutzt werden. Insbesondere können die metabolischen Strukturen mit der größten Dominanz in den Daten ermittelt werden. Wie man aus **Fig.** 5 erkennt, ergeben sich für die Komponenten k = 1 und k = 2 charakteristische Spektren, welche für bestimmte m/z-Werte Peaks aufweisen, woraus Rückschlüsse auf die molekulare Zusammensetzung der entsprechenden Komponenten gezogen werden können.

[0047] Die im Vorangegangenen beschriebenen Ausführungsformen des erfindungsgemäßen Verfahrens weisen eine Reihe von Vorteilen auf. Insbesondere wird durch eine nicht-negative Matrix-Faktorisierung mit entsprechenden Straftermen eine hohe Kompression von Massenspektrometriedaten erreicht, wobei wichtige Informationen aus den Massenspektrometriedaten erhalten bleiben. Diese Informationen können im Rahmen einer Visualisierung wiedergegeben werden.

Literaturverzeichnis:

[1] Dennis Trede, Jan Hendrik Kobarg, Janina Oetjen, Herbert Thiele, Peter Maass, and Theodore Alexandrov. On the importance of mathematical methods for analysis of MALDI-Imaging mass spectrometry data. Journal of Integrative Bioinformatics, 9(1):189, 2012.

[2] T. Blumensath, M.E. Davies (2008): Iterative Thresholding for Sparse Approximations. Journal of Fourier Analysis and Applications, Volume 14, Issue 5-6, pp 629-654

[3] D.D. Lee, H.S. Seung (1999): Learning the parts of objects by non-negative matrix factorization. Nature, Volume 401, pp 788-791

[4] L.I. Rudin, S. Osher, E. Fatemi (1992): Nonlinear total variation based noise removal algorithms. Physica D, Volume 60, pp 259-268

[5] M. Golbabaee, S. Arberet, P. Vandergheynst (2010): Distributed compressed sensing of hyperspectral images via blind source separation. The Asilomar Conference on Signals, Systems, and Computers, November 2010

Patentansprüche

1. Verfahren zum rechnergestützten Verarbeiten von räumlich aufgelösten Massenspektrometriedaten mit einer Vielzahl von jeweiligen Messpositionen (x) zugeordneten Spektren, wobei ein Spektrum (s_x

(m/z)) für mehrere Bildkanäle (m/z) jeweilige Intensitäten (I) umfasst, welche eine für den entsprechenden Bildkanal (m/z) detektierte physikalische Größe repräsentieren, wobei die Bildkanäle (m/z) Masse-zu-Ladungs-Verhältnisse darstellen und die für den entsprechenden Bildkanal (m/z) detektierte physikalische Größe die Anzahl der für das entsprechende Masse-zu-Ladungs-Verhältnis detektierten Moleküle repräsentiert, bei dem:
- die Massenspektrometriedaten mittels einer nicht-negativen Matrix-Faktorisierung zerlegt werden, gemäß der eine Datenmatrix (X), welche für jeweilige Messpositionen (x) die Intensitäten (I) für die Bildkanäle (m/z) des der jeweiligen Messposition (x) zugeordneten Spektrums (s_x(m/z)) umfasst, als ein Matrixprodukt einer Komponenten-Matrix (S) und einer Koeffizienten-Matrix (C) dargestellt wird, wobei die Komponenten-Matrix (S) für eine vorbestimmte Anzahl von Komponenten (k) jeweils Intensitäten eines der jeweiligen Komponente (k) zugeordneten Spektrums (ψ_k(m / z)) für die Bildkanäle (m/z) umfasst und wobei die Koeffizienten-Matrix (C) für die vorbestimmte Anzahl von Komponenten (k) jeweils Koeffizienten für die Messpositionen (x) umfasst, wobei die nicht-negative Matrix-Faktorisierung derart ausgestaltet ist, dass die zerlegten Massenspektrometriedaten im Vergleich zu den Massenspektrometriedaten vor der Zerlegung eine geringere Datenmenge aufweisen;
- basierend auf der nicht-negativen Matrix-Faktorisierung eine Optimierung einer Kostenfunktion in der Form einer Maximierung oder Minimierung der Kostenfunktion in Abhängigkeit von der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) als Variablen durchgeführt wird, wobei die Kostenfunktion ein Funktional (f(X, S, C)) umfasst, das von der Datenmatrix (X), der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) abhängt;
- bei der Optimierung der Kostenfunktion eine erste und eine zweite Nebenbedingung einfließt, wobei gemäß der ersten Nebenbedingung eine Optimierung der Kostenfunktion hin zu einer möglichst geringen Anzahl von Einträgen der Komponenten-Matrix (S) ungleich Null erfolgt und gemäß der zweiten Nebenbedingung eine Optimierung der Kostenfunktion hin zu einer möglichst geringen Anzahl von Einträgen der Koeffizienten-Matrix (C) ungleich Null erfolgt;
- die durch die nicht-negative Matrix-Faktorisierung erzeugten Einträge der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) gespeichert werden.

2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet**, dass die Massenspektrometriedaten MALDI-Massenspektrometriedaten (MALDI = Matrix-Assisted Laser Desorption/Ionization) oder DESI-Massenspektrometriedaten (DESI = Desorption Electrospray Ionization) oder SIMS-Massenspektrometriedaten (SIMS = Secondary Ion Mass Spectrometry) sind.

3. Verfahren nach Anspruch 1 oder 2, **dadurch gekennzeichnet**, dass das Funktional (f(X, S, C)) die L2-Norm der Differenz zwischen der Datenmatrix (X) und dem Matrixprodukt aus der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) umfasst, wobei die Optimierung der Kostenfunktion hin zu möglichst kleinen L2-Normen erfolgt.

4. Verfahren nach einem der vorhergehenden Ansprüche, **dadurch gekennzeichnet**, dass die erste Nebenbedingung durch eine erste Funktion (f_1(S)) repräsentiert wird, welche Bestandteil der Kostenfunktion ist und zu dem Funktional (f(X, S, C)) addiert wird, so dass die Optimierung der Kostenfunktion zur Bevorzugung einer Komponenten-Matrix (S) mit einer möglichst geringen Anzahl von Einträgen ungleich Null führt.

5. Verfahren nach Anspruch 4, **dadurch gekennzeichnet**, dass die erste Funktion (f_1(S)) die L1-Norm der Komponenten-Matrix (S) umfasst oder eine Summe von Normen der Komponenten-Matrix (S), wobei die Summe vorzugsweise die L1-Norm und die L2-Norm der Komponenten-Matrix (S) umfasst.

6. Verfahren nach einem der vorhergehenden Ansprüche, **dadurch gekennzeichnet**, dass die zweite Nebenbedingung durch eine zweite Funktion (f_2(C)) repräsentiert wird, welche Bestandteil der Kostenfunktion ist und zu dem Funktional (f(X, S, C)) addiert wird, so dass die Optimierung der Kostenfunktion zu einer Bevorzugung einer Koeffizienten-Matrix (C) mit einer möglichst geringen Anzahl von Einträgen ungleich Null führt.

7. Verfahren nach Anspruch 6, **dadurch gekennzeichnet**, dass die zweite Funktion (f_2(C)) die TV-Norm der Koeffizienten-Matrix (C) umfasst.

8. Verfahren nach einem der vorhergehenden Ansprüche, **dadurch gekennzeichnet**, dass die Koeffizienten-Matrix (C) als 2D-Wavelets repräsentiert wird, wobei als zweite Nebenbedingung eine begrenzte Anzahl von 2D-Wavelet-Koeffizienten festlegt wird oder gemäß der zweiten Nebenbedingung die L1-Norm der 2D-Wavelet-Koeffizienten möglichst klein wird.

9. Verfahren nach einem der vorhergehenden Ansprüche, **dadurch gekennzeichnet**, dass die vorbestimmte Anzahl von Komponenten (k) derart festgelegt wird, dass die Anzahl der gespeicherten Einträge der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) einem vorgegebenen Prozentsatz der Anzahl der Einträge der Datenmatrix (X) entspricht, wobei der vorgegebene Prozentsatz ins-

besondere zwischen 1% und 10% der Anzahl der Einträge der Datenmatrix liegt.

10. Verfahren nach einem der vorhergehenden Ansprüche, **dadurch gekennzeichnet**, dass aus den gespeicherten Einträgen der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) eine Visualisierung auf einem Display derart generiert wird, dass für zumindest einen Teil der Komponenten (k) jeweils ein Bild aus die Messpositionen (x) repräsentierenden Bildpunkten erzeugt wird, wobei ein jeweiliger Bildpunkt den Koeffizienten ($c_k(x)$) für die jeweilige Komponente (k) der Koeffizienten-Matrix (C) und die durch den jeweiligen Bildpunkt (BP) repräsentierte Messposition (x) codiert.

11. Verfahren nach einem der vorhergehenden Ansprüche, **dadurch gekennzeichnet**, dass aus den gespeicherten Einträgen der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) eine Visualisierung auf einem Display derart generiert wird, dass für zumindest einen Teil der den Komponenten (k) zugeordneten Spektren ($\psi_k(m / z)$) jeweils eine graphische Darstellung erzeugt wird, welche die Intensitäten des der jeweiligen Komponente (k) zugeordneten Spektrums ($\psi_k(m / z)$) in Abhängigkeit von den Bildkanälen (m/z) anzeigt.

12. Verfahren nach einem der vorhergehenden Ansprüche, **dadurch gekennzeichnet**, dass aus den gespeicherten Einträgen der Komponenten-Matrix (S) und der Koeffizienten-Matrix (C) durch Matrixmultiplikation dieser Matrizen die Massenspektrometriedaten rekonstruiert werden.

13. Maschinenlesbarer Träger mit darauf gespeichertem Programmcode zur Durchführung eines Verfahrens nach einem der vorhergehenden Ansprüche, wenn der Programmcode auf einem Computer ausgeführt wird.

Es folgen 5 Seiten Zeichnungen

Anhängende Zeichnungen

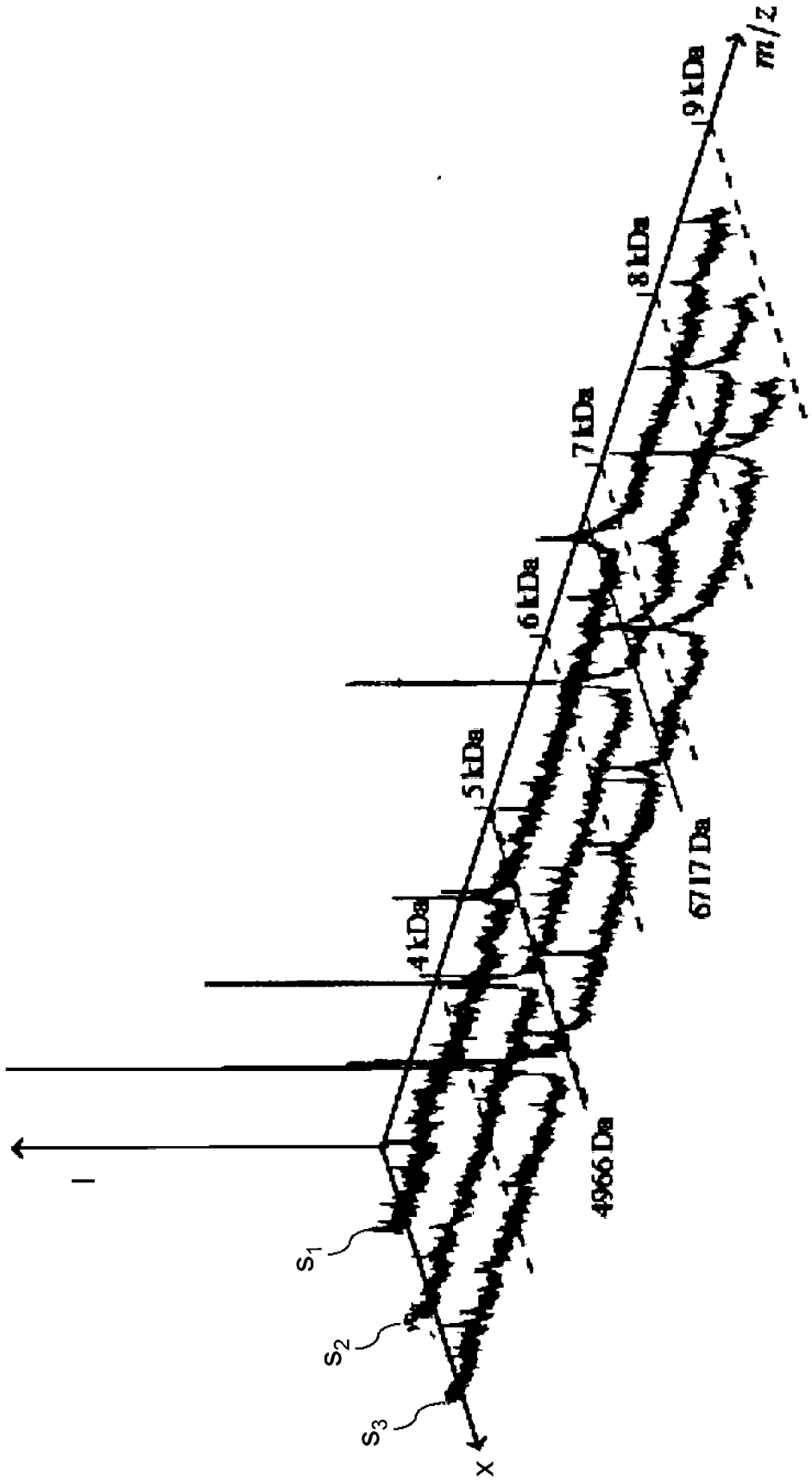


Fig. 1

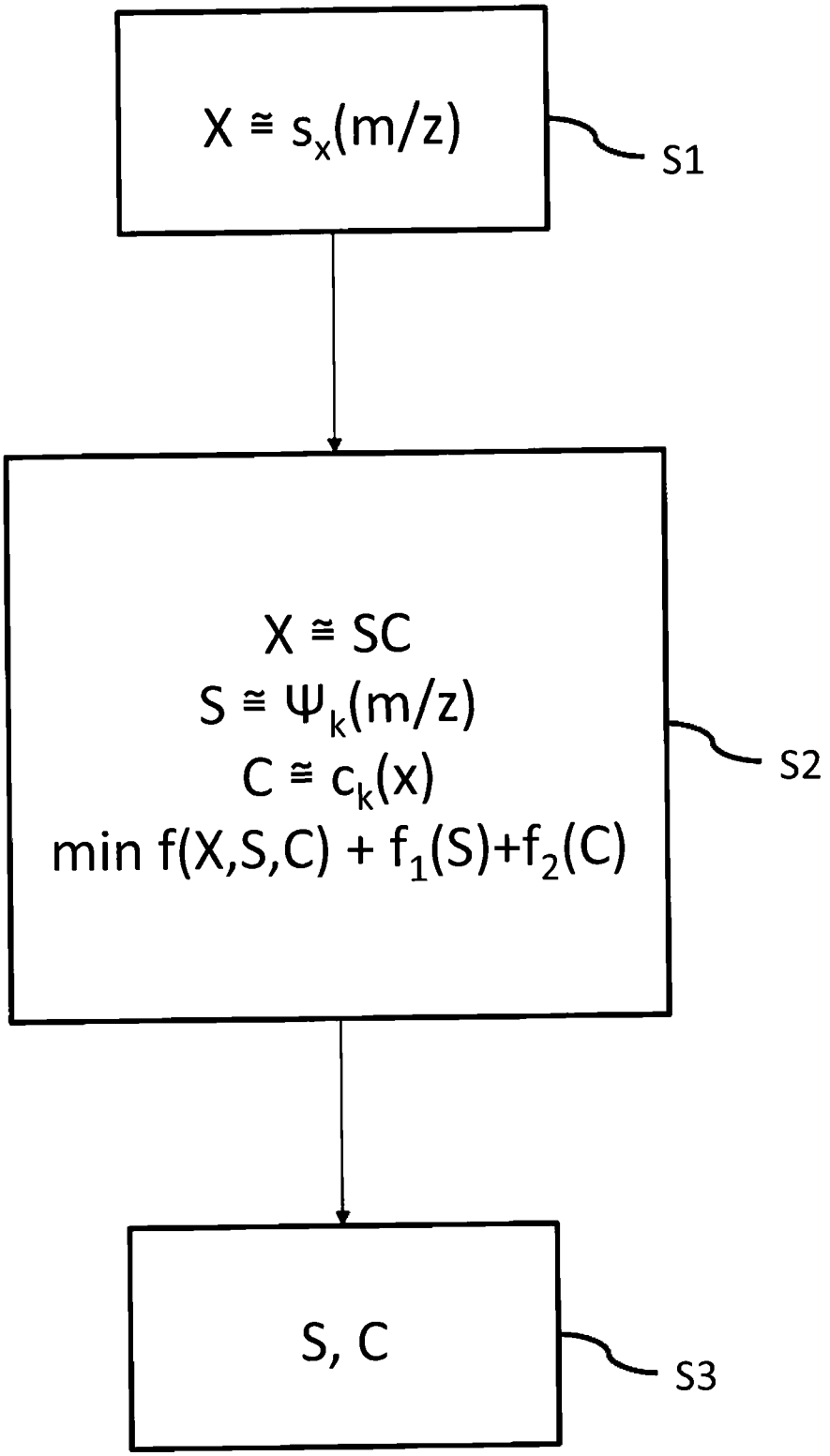

$X \cong s_x(m/z)$
S1
$X \cong SC$
$S \cong \Psi_k(m/z)$
$C \cong c_k(x)$
$\min f(X,S,C) + f_1(S)+f_2(C)$
S2
S, C
S3

Fig. 2

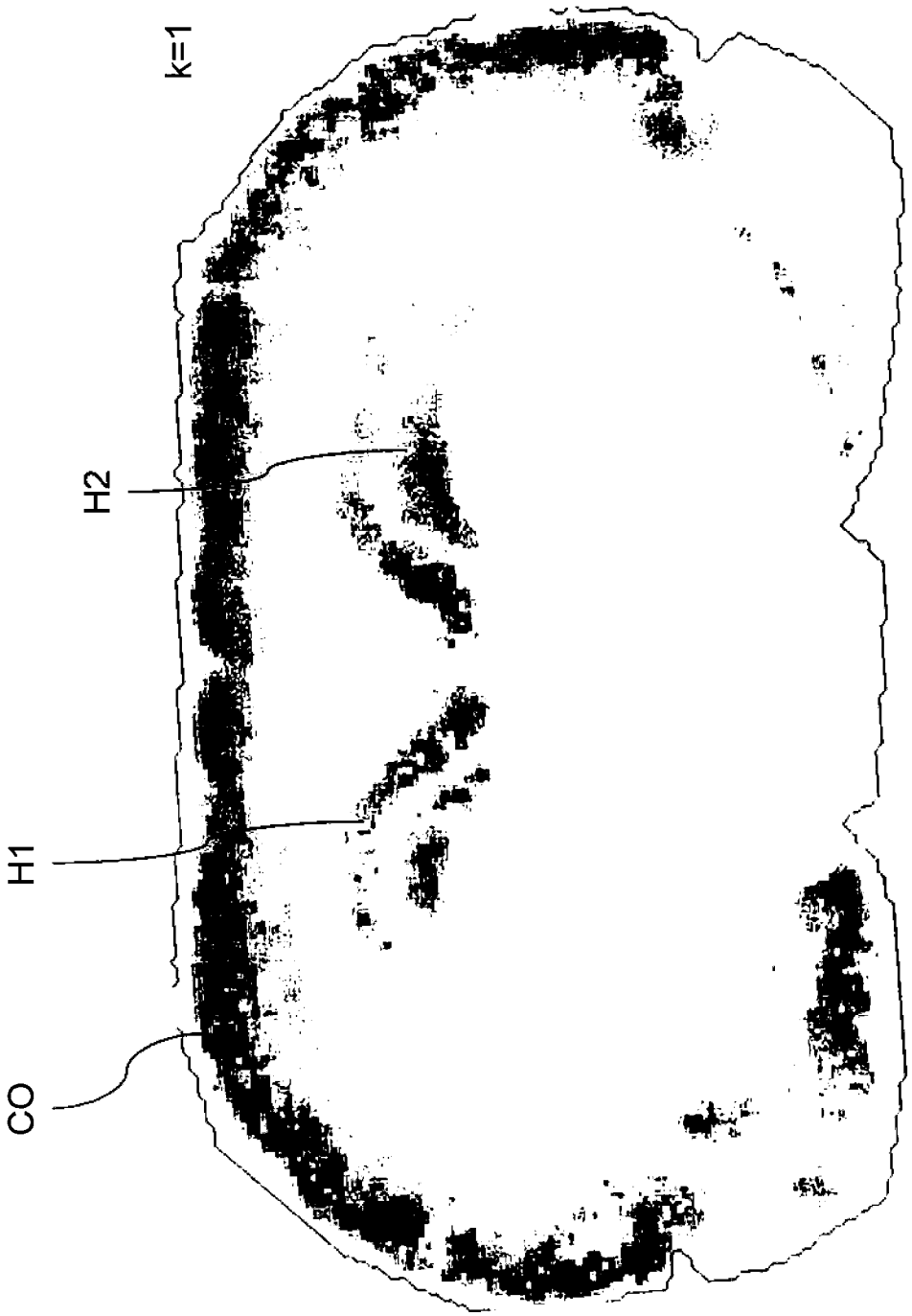


Fig. 3

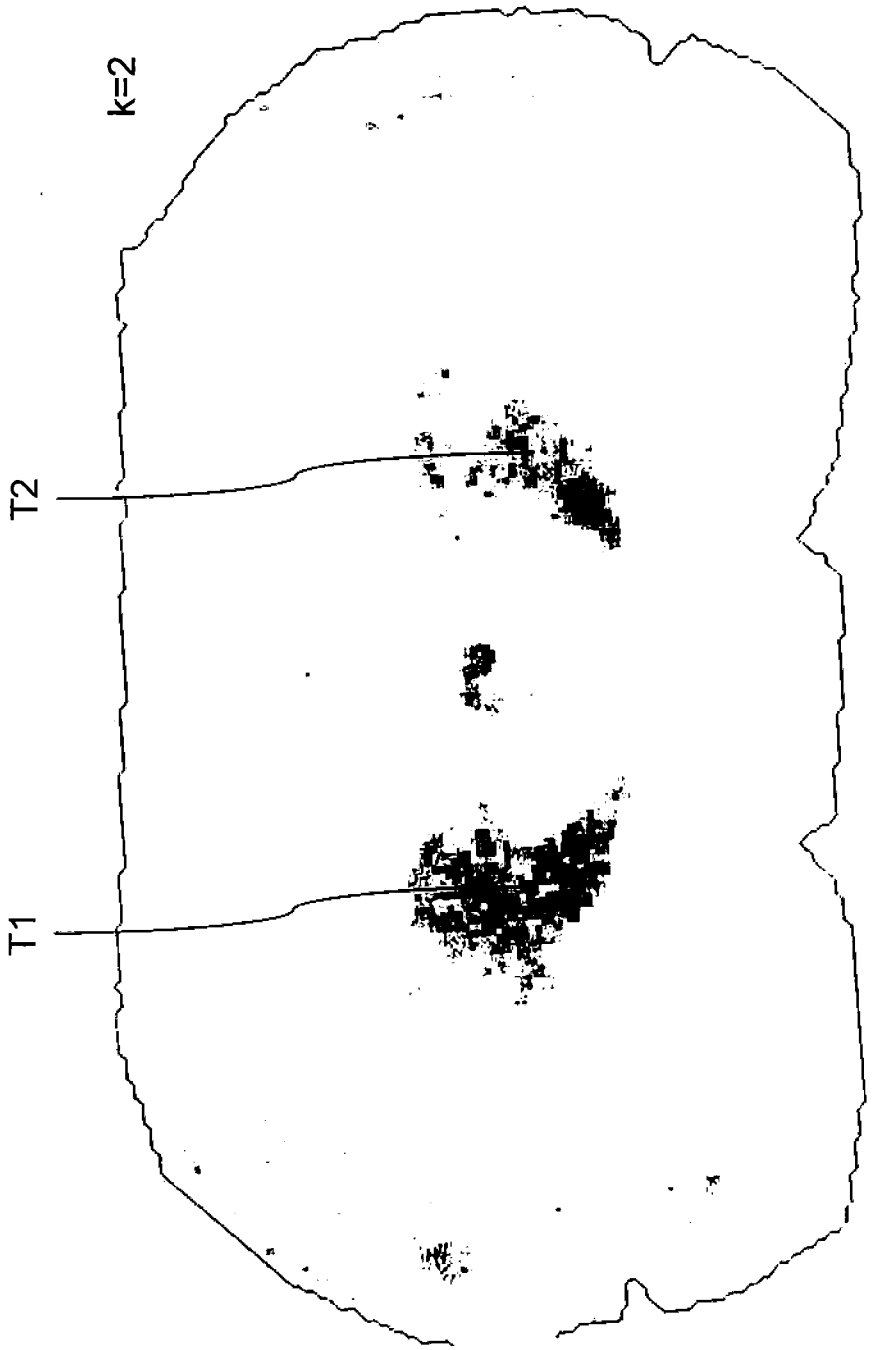


Fig. 4

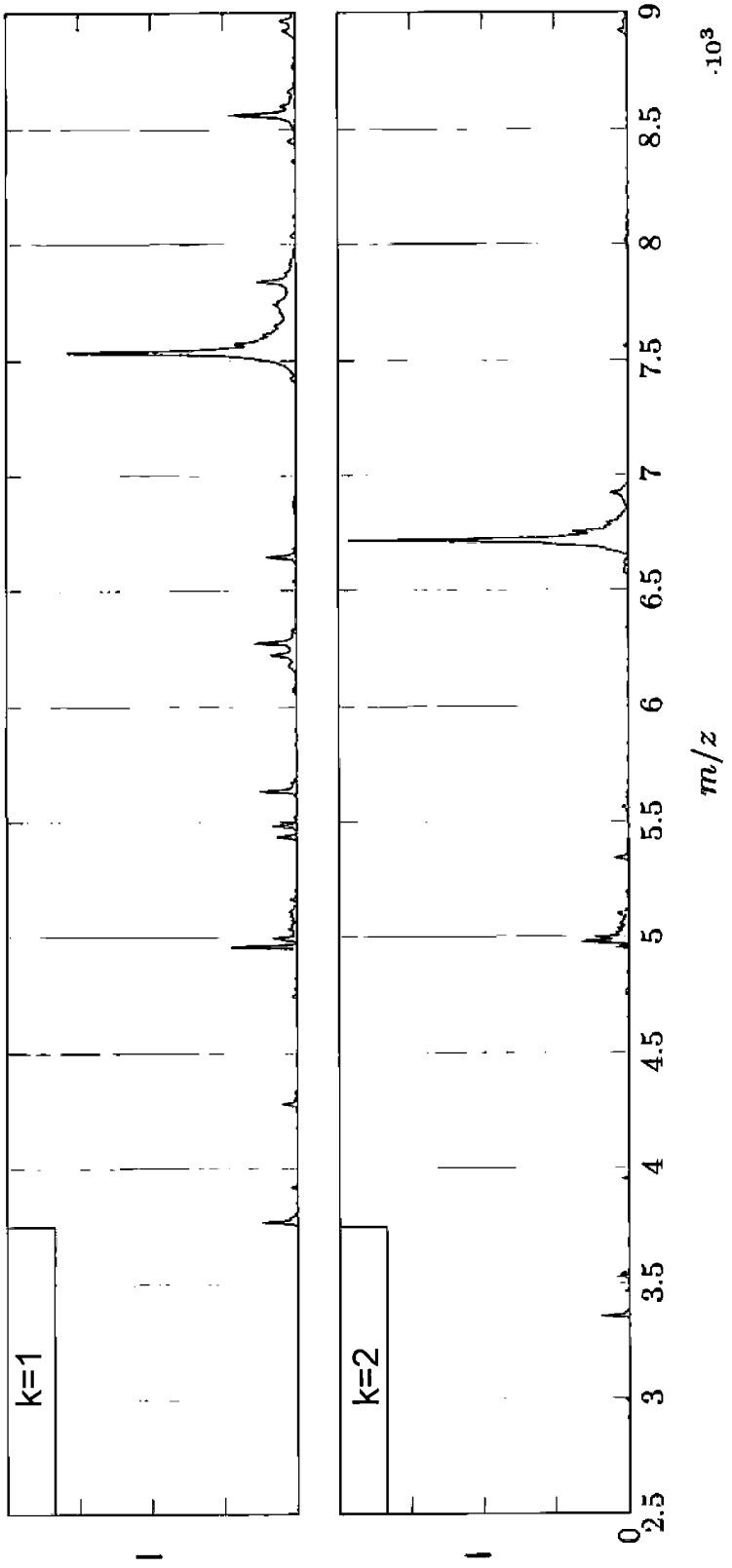


Fig. 5

 Das Dokument wurde durch die Firma Luminess hergestellt.