INF FAKULTÄT FÜR INFORMATIK

Tagungsband der

1. Doktorandentagung Magdeburger-Informatik-Tage 2012 (MIT 2012)

Herausgeber:

Georg Krempl
Claudia Krull
Frank Ortmeier
Eike Schallehn
Sebastian Zug

17. Juli 2012

Impressum:

Verlag: Otto-von-Guericke-Universität Magdeburg

Verlagsnummer: 85720

Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Postfach 41 20
39016 Magdeburg

Herausgeber:
Georg Krempl
Claudia Krull
Frank Ortmeier
Eike Schallehn
Sebastian Zug

Redaktionsschluß: 7. Juli 2012

Redaktion/Gestaltung:
Eike Schallehn

Herstellung:
Magdeburger DigitalDruckerei GmbH

ISBN 978-3-940961-73-0

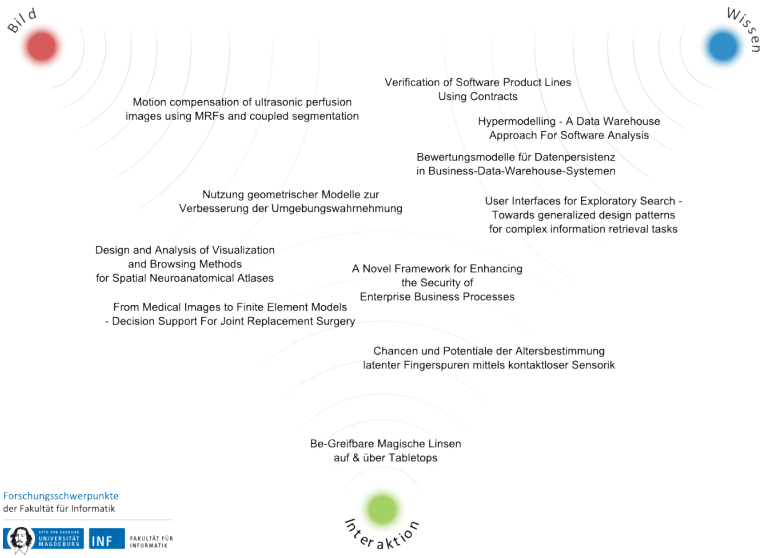
`http://www.cs.uni-magdeburg.de/`

Vorwort

Dieser Band ist eine Zusammenfassung der Beiträge zum 1. Magdeburger-Informatik-Tag (MIT). Hier präsentierten hervorragende junge Wissenschaftler der Fakultät für Informatik der Otto-von-Guericke-Universität ihre fortgeschrittenen Promotionsprojekte. Ziel der Tagung ist die Vorstellung anerkannter Forschungsergebnisse unserer Fakultät über Fachgebiets- und Universitätsgrenzen hinweg.

Das Forschungsprofil an der Fakultät Informatik richtet sich an den drei Schwerpunkten Bild, Wissen und Interaktion aus. Der Schwerpunkt „Bild“ beschäftigt sich mit der Repräsentation, Analyse und Vermittlung bildhafter Information. Dies beinhaltet speziell die Bereiche Bildverstehen, Modellierung, Bilderzeugung und Visualisierung. Forschungsarbeiten im Schwerpunkt „Wissen“ beschäftigen sich mit den methodischen und technologischen Grundlagen des Erwerbs, der Modellierung und Repräsentation, der Verwaltung und der Verarbeitung von Daten, Informationen und Wissen. Der Schwerpunkt „Interaktion“ adressiert mit Forschungsarbeiten zu Multimodalität, Zuverlässigkeit, Sicherheit und Technologie wichtige Herausforderungen moderner Mensch-Technik-Interaktion sowie der Interaktion technischer Geräte untereinander.

Die folgende Grafik ordnet die vorgestellten Arbeiten in Bezug zu diesen drei Leitmotiven ein.



Der vorliegende Tagungsband dokumentiert sowohl die Vielseitigkeit als auch die Konvergenz der Forschungsaktivitäten an der Fakultät für Informatik.

Magdeburg, den 17. Juli 2012

Georg Krempl
Claudia Krull
Frank Ortmeier
Eike Schallehn
Sebastian Zug

Inhaltsverzeichnis

Nutzung geometrischer Modelle zur Verbesserung der Umgebungswahrnehmung

André Dietrich
Otto-von-Guericke-Universität
Institut für Verteilte Systeme (IVS)
Gebäude 29, Universitätsplatz 2, 39106 Magdeburg
Email: dietrich@ivs.cs.uni-magdeburg.de

***Zusammenfassung*—Intelligente Systeme müssen sich in immer komplexeren und dynamisch verändernden Umgebungen zurechtfinden. Einerseits wächst die Zahl und Komplexität der zu erfüllenden Aufgaben, andererseits aber auch die in instrumentierten Umgebungen zur Verfügung stehenden Sensoren. Um seine Aufgaben lösen zu können, benötigt jeder Controller, jeder Problemlöser eine spezielle Sicht auf die jeweilige Umgebung. Eine intelligente Umgebungswahrnehmung bildet somit die Grundlage zur Bewältigung einer Vielzahl von Problemen in intelligenten Umgebungen. Diese Arbeit zeigt einen möglichen Weg auf, wie die Umgebungsabstraktion am Beispiel einer Wahrnehmungs-Middleware aus der Applikationsentwicklung herausgelöst werden kann.**

I. Einleitung und Motivation

Mit der voranschreitenden Entwicklung von Assistenzrobotern, die unmittelbar mit dem Menschen interagieren, vollzieht sich ein Wandel in diesem Forschungsfeld. Heutige autonome und teilautonome Robotersysteme finden sich im Unterschied zu den gekapselten Industrierobotern in Anwendungsszenarien wieder, die eine hohe Komplexität und Veränderlichkeit aufweisen. Mit dieser Anpassungsfähigkeit an reale Umgebungen können immer neue Anwendungsfelder erobert werden, wie in der Service-, Healthcare- oder Search-and-Rescue Robotik. Im Produktionsprozess geht der Trend von festen Fertigungszellen und Fließbändern hin zu flexiblen Produktionsstrecken, so werden vor allem in der Automobilindustrie zunehmend Automobile auf speziellen Kundenwunsch (Verlangen nach mehr Individualität und Differenzierung) angefertigt. Die neuen Möglichkeiten und Ideen können durch Begriffe wie „Internet of Things“, „Ambient“ und „Urban Intelligence“, „Ubiquitous/Pervasive Computing“, „Industrial/Building Automation“ oder „Smart Automobiles“ und viele mehr beschrieben werden.

Die breit gefächerten neuen Szenarien bergen jedoch viele grundlegende Probleme. Insbesondere die wachsende Zahl von verteilten Komponenten – Sensoren, Controller, Aktoren – die für die Umsetzung der komplexen Aufgaben nötig sind, erschwert die Umsetzung klassischer Konzepte, wie den etablierten Regelkreis zur Steuerung mechatronischer Systeme. Neue Arten von Steuerungsalgorithmen erweitern diese Ansätze um Aufgaben, wie die Überwachung (Supervision), Koordination und Planung, Situation Awareness, Diagnostik und Optimierung. Beispiele hierfür wären hierarchisch strukturierte Controller wie sie im Automobil zum Einsatz kommen [1], dynamisch rekonfigurierbare Controller bei Unmanned Aerial Vehicles (UAVs) [2] oder ein weit reichendes Spektrum von reaktiven bis hin zu deliberativen Controllern im Bereich der Robotik [3]. Ein bisher ungelöstes Problem betrifft die sensorische Wahrnehmung in veränderlichen intelligenten Umgebungen. Hier kann durch die Nutzung der zur Verfügung stehenden externen und heterogenen Sensorik (Gebäudeautomatisierung, Sensornetze, andere Roboter, mobile Geräte der Menschen) die Sicht auf die Umgebung erweitert werden. Neben der reinen Erfassung stellt sich aber die Frage, wie diese Daten zu ordnen, zu bewerten, zu selektieren sind.

So ist die flexible Integration zusätzlicher sensorischer und aktorischer Komponenten heute noch ein Problem und erfordert eine manuelle Integration und Anpassung des Applikationscodes. Als ein einfaches Beispiel kann hierfür die automatische Einparkhilfe dienen. Diese Steuerung benötigt eine wohl definierte Sicht auf die Umgebung (externes Modell), gewonnen durch die lokal zur Verfügung stehende Sensorik und ein internes Modell zur Steuerung des Automobils. Wird die Konfiguration des Automobils zum Beispiel über einen zusätzlichen Anhänger verändert, so ergeben sich mehrere Herausforderungen:

1) Die Umgebungswahrnehmung ist auf die zusätzlichen Sensoren des Anhängers auszuweiten.
2) Die Interpretation der vorhandenen Sensorik muss, wegen der möglichen Verdeckungen durch den Anhänger, angepasst werden.
3) Die neuen Abmessungen verändern die Entscheidungsfunktion über die Größe der nötigen Parklücke.
4) Auch das Fahrverhalten des Automobils während des automatischen Einparkens muss je nach Art und Ladung des Anhängers modifiziert werden.

Mit Nutzung heutiger Verfahren und Methoden ist dies bisher nur durch eine umfangreiche Rekonfiguration des Systems möglich, obwohl sich die Aufgabe „Einparken“ nicht geändert hat!

Abhilfe kann hier nur eine generelle Trennung von Applikationslogik, Wahrnehmung und Umgebungsmodellierung leisten. Wie zuvor erwähnt, benötigt jede Art der Applikation/Steuerung eine spezielle Sicht auf ihre Umgebung, dabei

kann es sich um einfache Parameter (Positionen, Abstände, Größen, Geschwindigkeiten, Vorhandensein von Sensorik, Zustände, etc.), zwei- und dreidimensionale Darstellungen, Graphen, semantische Modelle, vieles mehr und Kombinationen daraus handeln. Eine generelle Abstraktion der Umgebung über eine geeignete Middlewareschicht könnte immer wiederkehrende Aufgaben der sensorischen Interpretation, Validierung und Fusion übernehmen. Eine Applikation könnte darauf aufbauend über geeignete Schnittstellen die Art der Darstellung sowie daran geknüpfte Anforderungen definieren. Die Middleware übernimmt die Aufbereitung und Vorverarbeitung der Sensor- und Systeminformationen im Sinne eines generellen Umgebungsmodells und stellt diese in abstrahierte Form über die Anwendungsschnittstellen bereit. Der Ansatz ist mit einer Kommunikations-Middleware vergleichbar, die ebenfalls eine Abstraktion der darunter liegenden Netzwerke und Nachrichtenformate bildet.

A. Vorhaben

Das in der Arbeitsgruppe EOS der Otto-von-Guericke-Universität entwickelte zweischichtige Abstraktionsmodell [4] soll mit der vorliegenden Dissertation um eine zusätzliche Abstraktionsebene erweitert werden, siehe Abbildung 1.

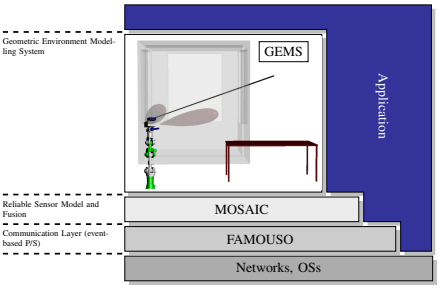


Abbildung 1. Mehrschichtiger Systementwurf mit FAMOUSO (Abstraktion der Kommunikation), MOSAIC (Abstraktion der Sensorik und Aktorik) und GEMS als Abstraktion der Umgebung

Auf der Kommunikationsebene wurde mit FAMOUSO eine Kommunikations-Middleware basierend auf dem Publish-Subscribe-Paradigma entwickelt, welches Echtzeitkommunikation auch zwischen hochintegrierten Systemen wie 8-bit-Mikrocontroller über verschiedenste Netze (wie CAN, Ethernet, 802.15.4, etc.) ermöglicht [5].

Darauf aufbauend wurde das MOSAIC-Framework für die Sensor- und Aktor-Abstraktion in instrumentierten Umgebungen konzipiert [6]. Es bietet Methoden für die fehlertolerante Datenakquisition sowie eine Processing-Architektur mit Fokus auf dynamische Szenarien. Durch den Austausch von XML-Datenblättern wird eine nahtlose Datenintegration und -interpretation ermöglicht, dies beinhaltet unter anderem die Konfiguration des Knotens sowie aller übertragbarer Daten wie physikalische Einheiten, mögliche Störungen, deren Auftretenswahrscheinlichkeit sowie deren Auswirkungen auf den Sensormesswert.

Der ursprünglich zweischichtige Ansatz scheitert jedoch bei komplexen Sensoren (z.B. Kameras und 3D-Sensoren), komplexen Umgebungen (Straßenverkehr) oder komplexen Tasks (Überholvorgang). Um die diversitären Informationen einer sich dynamisch verändernden Umgebung richtig interpretieren, validieren und fusionieren zu können, müssen diese Informationen im Kontext betrachtet und eingeordnet werden. Das Ziel dieser Arbeit ist die schrittweise Entwicklung eines Konzeptes, mit dem mehr und mehr Funktionalität, die für eine allgemeine Umgebungsabstraktion und Sensordateninterpretation notwendig ist, aus der Applikationsentwicklung herausgelöst werden kann. Diese Funktionalität soll dann gekapselt verschiedensten Controllern zur Verfügung gestellt werden. Um diese Arbeit in geeignetem Maße einzuschränken, soll hier der Schwerpunkt auf die räumliche Umgebungserfassung gelegt werden. Chemische, thermische oder mechanische Eigenschaften könnten abgebildet werden, bleiben aber zunächst unbeachtet. Hierbei sollen Grundlagen für eine von Applikationen/Controllern unabhängige Umgebungswahrnehmung geschaffen werden und belegt werden, dass Wahrnehmung auch mithilfe einer separaten Middleware möglich ist.

B. Gliederung

Im nächsten Abschnitt werden zunächst die notwendigen Teilschritte zur Entwicklung einer Wahrnehmungs-Middleware aufgezeigt. Dabei werden zu jedem Teilschritt publizierte Arbeiten sowie die bisherigen und noch ausstehende Entwicklungen aufgeführt. Da bestehende Arbeiten immer nur Teilaspekte der Umgebungsabstraktion abdecken oder stark spezialisierte Lösungen für einen besonderen Anwendungsfall aufzeigen, werden die verwandten Arbeiten erst im dritten Abschnitt vorgestellt. Abschließend erfolgt eine Diskussion und Einordnung im Hinblick auf die verwandten Arbeiten.

II. Roadmap

In den folgenden Abschnitten sollen die jeweiligen Teilschritte skizziert und erläutert werden, die aus meiner Sicht für die Entwicklung einer Middleware zur Umgebungsabstraktion notwendig sind. Dabei wird Stück für Stück immer mehr Funktionalität aus der Kontrollapplikation extrahiert werden. Wie in Abschnitt I-A erläutert, sollen mit dieser Arbeit Grundlagen geschaffen und evaluiert werden, das heißt, dass der Hauptteil der praktischen Arbeit die Teilschritte „Modell zur geometrischen Einordnung“ und „Veränderliche Modelle für dynamische Umgebungen“ ausmachen. Nichtsdestotrotz werden auch der darauf folgende Teilschritt sowie die damit verbundenen Möglichkeiten erläutert. Einige daraus resultierende Implementierungen dienen als Proof-Of-Concept.

Zunächst soll von fixen Umgebungen und statischen Konfigurationen ausgegangen werden. Diese werden im weiteren Verlauf auf immer komplexer werdende verteilte Systeme und Anwendungen ausgeweitet, wobei auch der Aspekt der

Dynamik derartiger Systeme und der Umgebung aufgegriffen wird.

A. Ausgangspunkt

Wie in Abschnitt I bereits erörtert und in Abbildung 2 dargestellt, ist in den meisten der heutzutage verwendeten Sensor-Aktor-Systeme die eigentliche Abstraktion der sensorischen Umgebungswahrnehmung ein inhärenter Bestandteil der Kontrollapplikation. Das heißt, dass sämtliche Parameter, Eingangs-, Ausgangs- und Störgrößen bereits in der Implementierungsphase bekannt sein müssen, um angemessen auf Veränderungen in der Umgebung reagieren zu können.

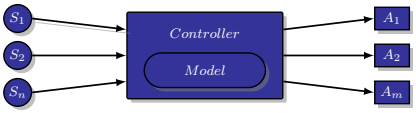


Abbildung 2. Allgemeines Konzept einer Sensor-Aktor-Kette

B. Modell zur geometrischen Einordnung

Das explizite Wissen über die verwendete Sensorik, Aktorik, die Umgebung und deren Konfiguration (Position und Orientierung), welches ein fest kodierter Teil in den meisten Kontrollapplikationen ist, soll im ersten Schritt aus dem Controller extrahiert werden (vgl. Abbildung 3). Hierbei wird von fixen und sich nicht dynamisch verändernden Systemen ausgegangen, wobei das Wissen über die Umgebung und die enthaltenen Komponenten mithilfe eines einheitlichen Beschreibungsformates definiert wird. Die jeweilige Interpretation dieser Informationen und der Umgang mit Sensordaten geschehen noch im Controller selbst (durch MOSAIC). Durch die Nutzung unterschiedlicher XML-Notationen können bereits einfache geometrische Modelle der Umgebung und der verwendeten Aktoren definiert werden, wie in Abbildung 1 dargestellt. Der Tisch, Manipulator mit Sensorik und das Modell des Raumes, werden dazu in ihrer Geometrie erfasst und abstrahiert. Diese Beschreibung der Umgebung und der darin enthaltenen statischen Objekte erfolgt durch ODE-XML (OpenDynamicsEngine), einer freien Bibliothek zur Simulation von Starrkörperdynamiken in der virtuellen Realität, vgl. [7]. Komplexere Roboter und Aktoren werden mit URDF (Unified Robot Description Format) beschrieben, dies beinhaltet eine dynamische und visuelle Beschreibung sowie ein vereinfachtes Kollisionsmodell und die Kinematik, vgl. [8]. Für eine realistische Beschreibung von Sensoren wird OpenRAVE (Open Robotics Automation Virtual Environment [9]) in Kombination mit MOSAIC verwendet.

Diese virtuelle Modellierung der Umgebung, Sensorik und Aktorik kann wie im Folgenden aufgeführt bereits zu Problemlösungen beitragen:

- Die Umrechnung von Translation und Rotationen zur Einordnung der Sensoren und der Sensormessungen kann automatisiert und anhand der Beschreibung des Systems geschehen.

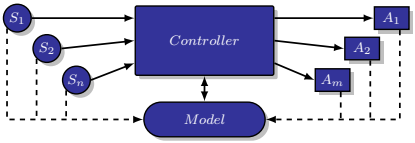


Abbildung 3. Extraktion geometrischer Eigenschaften

- Neben signalbasierten Methoden der Fehlerdetektion für Sensoren, w. z. B. in [10] vorgestellt oder durch Anwendung kombinierter Testverfahren [11], können Plausibilitätsprüfungen durch den Vergleich mit virtuellen Messungen erfolgen, vgl. [12].
- Die Nutzung von Filtermöglichkeiten zur taskorientierten Sensorselektion, aufgrund von Sensorpositionen, Abdeckungs- und Überwachungsbereichen, wie bereits teilweise in [13] vollzogen wurde, stellt eine wichtige Ergänzung zur reinen signalbasierten Sensorselektion dar (vgl. [14]).
- Ausfälle realer Sensormessungen können mithilfe der virtuellen Messergebnisse kurzfristig überbrückt werden.

Für das Beispiel der automatischen Einparkhilfe bedeutet dies, dass die Größenparameter des Automobils und des Anhängers aus der Beschreibung gewonnen werden können sowie die Abmessungen der Parklücke und die Positionen und Orientierungen der nutzbaren Sensoren. Es ist eine Sensorselektion möglich, wobei die durch den Anhänger verdeckten Sensoren, trotz der Erzeugung richtiger Messwerte, als unbrauchbar herausgefiltert werden. Beispielhaft ist dies dargestellt in Abbildung 6(a), wobei das einparkende Automobil eine Beschreibung seiner selbst, sowie der lokal zur Verfügung stehenden Sensorik (rot) nutzt, auf deren Basis bereits eine taskorientierte Sensorselektion erfolgen kann.

C. Veränderliche Modelle für dynamische Umgebungen

Um in instrumentierten Umgebungen operieren zu können und die eigene Sicht kontinuierlich auf die Änderungen in der Umgebung anzupassen, sei es durch die Interpretation externer Sensorik oder durch Integration externer Aktorik in das Umgebungsmodell, werden erweiterte Mechanismen benötigt. Das Wissen über die Umgebung und die enthaltenen Komponenten kann nun nicht mehr in einem XML-Dokument fest vorgegeben werden, sondern verlangt eine dynamische Integration. Hierzu soll auf den im vorigen Schritt entwickelten Mechanismen aufgebaut werden. Jede Entität besitzt eine XML-Beschreibung seiner selbst mit allen zur Verfügung stehenden Services und publizierten Topics [13]. Diese Beschreibungen können zwischen interessierten Komponenten ausgetauscht werden, sodass zusätzliche lokale (Plug&Play) oder externe Sensoren und Aktoren in die lokale Sicht des Controllers auf seine Umgebung eingebunden werden können, siehe Abbildung 4. Dies hat den zusätzlichen Vorteil, dass Steuersignale anderer Komponenten (w. z. B. Verfahrbewegungen oder Trajektorien fremder Roboter) innerhalb des Modells

interpretiert und nachvollzogen werden können.

Die hierfür nötigen Transformationen können jedoch nicht mehr aus einem XML-Dokument gewonnen werden, sondern müssen ebenfalls dynamisch erfolgen. Gegebenenfalls ist die Bestimmung einer Transformation auch nur über eine Folge von Einzeltransformationen bekannt, welche zusätzlich mit Unsicherheiten belegt sind. Hierfür soll das für Transformationen in `ROS` verwendete System `tf` verwendet und erweitert werden (vgl. Abschnitt III-A1). Die Einzeltransformationen werden zurzeit in `tf` noch in einer Baumstruktur organisiert und ein Umgang mit Unsicherheiten ist nicht gegeben. Diese Informationen in Verbindung mit der Sensorbeschreibung und aktuellen Sensorvaliditätswerten, wie sie MOSAIC auslieferte, können als Grundlage für eine taskorientierte Sensorselektion dienen, wie sie in [15] beschrieben wurde. Damit werden im Vergleich zu [16], [17], [18] auch erweiterte Fusions- und Validierungsansätze von heterogenen Sensordaten oder Lokalisationen, wie in [19] vorgestellt, ermöglicht.

Ein weiteres Problem dieser Ebene liegt in der verteilten Datenhaltung. Da jede Entität eine eigene lokale Sicht auf die direkte Umgebung besitzt und diese ständig mithilfe von Sensormessungen aktualisiert, müssen auch diese heterogenen Messdaten sowie deren Interpretationen (Hindernisse, komplexe Objekte und Geometrien, Personen, etc.) unter den verschiedenen Komponenten austauschbar sein. So können auch die Speicherung und der Austausch historischer Messdaten für einige Anwendungen notwendig sein (wie die Lokalisation oder die Erstellung von Statistiken und Bewegungsprofilen). Das bedeutet, dass neben einer geeigneten Organisation der Daten (Modellparameter, Sensormessungen, Systemzustände, etc.) hier auch der Schwerpunkt in der Schnittstellenentwicklung liegt, die einen Zugriff auf diese verteilten Daten ermöglicht. Für die Lösung dieses Problems wird aktuell an der Integration einer verteilten Datenbank gearbeitet. Dabei wird auf dem verteilten Datenbankverwaltungssystem `Cassandra` [20] aufgebaut, das speziell für umfangreiche Datenbanken und Skalierbarkeit in verteilten Systemen ausgelegt wurde. Die Knoten werden dezentral in Clustern organisiert, wobei Knoten hinzukommen und wegfallen können, was dem dynamischen Konzept der vorliegenden Arbeit entspricht.

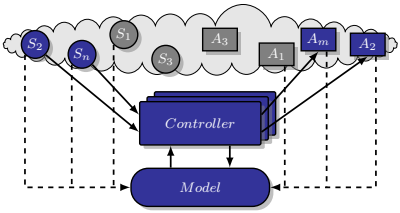


Abbildung 4. Erweiterung des Modellansatzes auf verteilte Systeme

Wie in Abbildung 6(a) dargestellt, wäre die automatische Einparkhilfe hiermit auch in der Lage zusätzlich die externe Sensorik (grau) der anderen Fahrzeuge in das eigene Umgebungsmodell zu integrieren, um damit den eigenen Erfassungs- und Wahrnehmungsbereich zu erweitern. Durch Einbezug der anderen Geometrien können auch Zustände (Beifahrertür ist offen) direkt im Modell interpretiert und dargestellt werden.

D. Middlewareintegration

Die zuvor entwickelten Methoden können für die Selektion und Integration von Umgebungsdaten, Aktoren und Sensorinformationen in die eigene Sicht auf die Umgebung sowie deren Validierung genutzt werden, während die Aufbereitung und Fusionierung von Daten noch immer in der Kontrollapplikation (mithilfe von MOSAIC) geschieht. In diesem letzten Schritt könnte, aufbauend auf den vorhergehenden Implementierungen, eine verteilte Middleware zur sensorischen Umgebungswahrnehmung entwickelt werden, die die sensorische Wahrnehmung und deren Interpretation gänzlich von der Kontrollapplikation abtrennt, vgl. Abbildung 5. Auch andere Aufgaben, die ursprünglich innerhalb des Controllers ausgeführt wurden, wie Lokalisation, Tracking, Kartenerzeugung, Hindernis- oder Objekterkennung, Anchoring, etc. werden in die Middleware ausgelagert. Ein Controller könnte über eine geeignete Schnittstelle seine Anforderungen an die sensorische Wahrnehmung definieren, das heißt, welcher Bereich der Umgebung, mit welcher Validität, Aktualität und in welchem Format. Diese Sicht auf die Umgebung wird dann über einen View zur Verfügung gestellt.

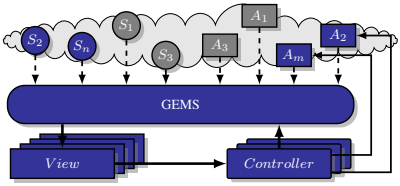


Abbildung 5. Vollständige Trennung von Umgebungswahrnehmung und Controller

1) Views: stellen eine abstrahierte und taskspezifische Sicht auf die Umgebung dar, wobei jede beliebige Art der Abstraktion denkbar ist. Benötigt die automatische Einparkhilfe z. B. einen View in Form einer Occupancy Grid Map mit einer bestimmten Auflösung und Güte (vgl. Abbildung 6(b)), so müsste die Middleware entscheiden, welche der zur Verfügung stehenden Sensorsysteme sich am besten zur Positionsbestimmung eignen (ggf. der Laserscanner des Jeeps) und diesen View ständig mit den verfügbaren sensorischen Informationen aktualisieren. Des Weiteren sind differenzierte Darstellungen der Umgebung auch für die Mensch-Maschine-Interaktion wichtig. In [21] wurden Visualisierungsmöglichkeiten der „Erweiterten Realität"für unterschiedliche Nutzerrollen in kooperativen Tasks mit Robotern vorgestellt. Somit benötigt ein Entwickler eine detailreiche Sicht auf das jeweilige System. Der damit interagierende Nutzer hingegen benötigt Wissen

über Vorhaben und Intentionen, wie geplante Arbeitsschritte, Arbeitsbereiche, Trajektorien, etc. Dass solche Methoden auch in anderen Szenarien nutzbar sind, wie bei der Patientenüberwachung, konnte in [22] bereits gezeigt werden.

2) Gesamtvalidität: Da jede Darstellung auch mit Unsicherheiten behaftet ist, wird eine Metrik benötigt, mit deren Hilfe die Validität solcher Views ausgedrückt werden kann. Hierzu könnte das bereits in MOSAIC genutzte Validitätsmaß [6] herangezogen werden. Aus den Einzelvaliditätswerten der zur Überwachung eines Szenarios eingesetzten Sensoren, den Abdeckungsbereichen und Aktualisierungsraten könnte eine Gesamtvalidität bestimmt werden. Damit wäre eine Anwendung in der Lage, ein Mindestmaß an Validität und Aktualität für einen View zu definieren. Es wäre dann Aufgabe der Middleware zu prüfen, ob genügend Sensorik in der jeweiligen Umgebung zur Verfügung steht und eine geeignete Sensorkombination auszuwählen. Sollten Anforderungen nicht erfüllbar sein, so muss die Middleware Auskunft darüber bzw. über die maximale Validität geben können, sodass sich der Controller ggf. darauf einstellen kann (z. B. durch veränderte Verfahrgeschwindigkeiten).

III. Verwandte Arbeiten

Im folgenden Teil soll ein Überblick gegeben werden, wie heutige Systeme ihre Umgebung wahrnehmen bzw. wie, mit welchen Hilfsmitteln und auf welchen Ebenen Wahrnehmung passiert.

A. Konventionelle Systeme

In den nächsten beiden Abschnitten soll das konventionelle Herangehen bei der Entwicklung von Robotik-Applikationen kurz erläutert werden sowie verschiedene Abstraktionsmöglichkeiten der Umgebung aufgelistet werden.

1) Robot Operating System (`ROS`): entwickelt sich dank einer offenen und rasant wachsenden Cummunity stetig weiter und ist mittlerweile zu einem Quasi-Standard bei der Implementierung und Integration von Robotikanwendungen geworden. Eine große Bibliothek sowie standardisierten Daten und Interfaces für eine Vielzahl unterschiedlicher Sensoren und Aktoren sind darin bereits enthalten (vgl. [23]). Für die räumliche Einordnung dieser Informationen (relative Position und Orientierung) wird das Paket `transformation` (`tf`) genutzt (vgl. [24]), das speziell für verteilte Systeme konzipiert wurde. Jeder Knoten publiziert seine lokalen Transformationen unter einem einheitlichen Topic. Dieses beinhaltet Translation, Rotation, einen Zeitstempel, eine Frame-ID und eine Child-Frame-ID (Elter- und Kind-Koordinatensystem). Mithilfe dieser IDs kann dann von jedem beteiligten Knoten ein Baum über den Einzeltransformationen aufgespannt werden. `tf` behält die Übersicht über die Einzeltransformationen kombiniert mit den Zeitstempeln.

Neben dem Einsatz zum visuellen Debugging `rviz` (vgl. [25]) bildet `tf` auch die Grundlage für viele weitere Module, wie den NavigationStack oder verschiedene Lokalisierungspakete. Es gibt zwar keine zentrale Einheit zur Abstimmung der Koordinatensysteme, deren IDs müssen aber schon zur Designzeit von Hand eindeutig festgelegt werden, Mehrdeutigkeiten sind nicht erlaubt. Der Vorteil der dezentralen Organisation geht mit hohem Bandbreitenverbrauch einher, da alle Nachrichten per Broadcast übertragen werden.

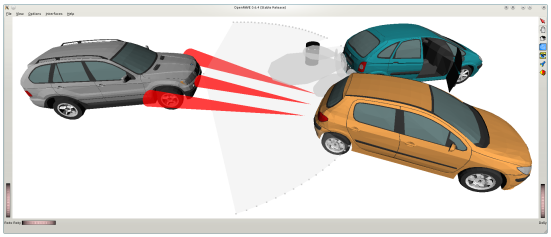
Die eigentliche Schwäche ergibt sich aus dem fehlenden Umgang mit Unsicherheiten. Nimmt man an, dass sämtliche Transformationen fehlerbehaftet sind, wie die Position einer mobilen Roboterplattform, so wächst die Ungenauigkeit bei der Koordinatentransformation mit zunehmender Pfadlänge, wobei ein einheitlicher Umgang mit fehlerbehafteten Koordinatentransformationen bereits in [26] vorgestellt wurde. Durch die Nutzung mehrerer unsicherer Transformationen kann eine Lösung gesucht werden, die dem Grad der durch die Applikation geforderten Gesamtunsicherheit entspricht.

`ROS` ist zwar speziell für Robotikapplikationen in verteilten Systemen konzipiert, die Modellbildung, das heißt, die Abstraktion der Umgebung durch die sensorischen Eingaben, bzw. die geeignete Fusionierung, Validierung und Interpretation muss immer noch vom Entwickler vorgegeben werden, wobei ein enorme Anzahl an verschiedensten Paketen und Werkzeugen zur Verfügung steht.

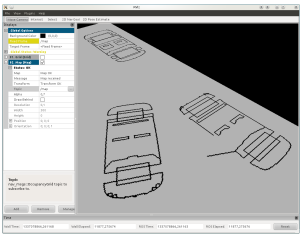
2) Räumliche Umgebungsmodelle: Mithilfe von `ROS` und ähnlichen Entwicklungsumgebungen können leicht anwendungsspezifische Darstellungen der Umgebung erzeugt werden. Da sie unserem Denken am ehesten entsprechen, sind räumliche Umgebungsmodelle für Menschen sehr leicht zu interpretieren. Jedoch sind sie für ein maschinelles Situationsbewusstsein aufgrund des sehr hohen Abstraktionsniveaus nur schwer erfassbar. So können Abstandsmessungen in 2D bzw. 3D-Strukturen überführt werden, die es einem autonomen System erlauben sich in Indoor- oder Outdoor-Umgebung zu orientieren und Hindernisse zu umgehen (vgl. [27], [28], [29]). Solche Darstellungen oder Karten können bereits einfach erzeugt und zwischen verschiedenen Entitäten ausgetauscht werden.

Mithilfe der Objekterkennung können aus 3D-Punktwolken komplexere geometrische Objekte abgeleitet werden. Solche Strukturen und Objekte lassen sich daraufhin mit zusätzlichen Attributen verknüpfen, wodurch sie eine zusätzliche Semantik erhalten. Dies wurde zum Beispiel in den Arbeiten von Rusu (siehe [30], [31]), die sich vor allem auf den Servicerobotereinsatz im Küchenumfeld konzentrieren, gezeigt. In [32], [33] wurden Objekte der Umgebung zusätzlich mit physikalischen Eigenschaften innerhalb der Modellrepräsentation versehen. Dies erlaubte durch zusätzliche Simulationen in ODE [7] Auskünfte über zukünftige Umgebungssituationen zu erzeugen. Des Weiteren konnte gezeigt werden, dass geometrische Modelle und die Attribuierung der darin enthaltenen Objekte (w. z. B. durch Masse, Farbe, etc.) essenziell für die Kooperation mit Menschen sind. So können Befehle der Form "Gib mir den blauen Schraubendreher links von mir!" nur erfüllt werden, wenn der Standpunkt und somit die Sicht des Befehlsgebers eingenommen werden können.

Die Vorwegnahme zukünftiger Situationen basierend auf Simulationen und zusätzlichen Modellen der Bewegung sind vor allem bei der Trajektorienplanung [34], [35] oder sogar bei



(a) Implementierung der automatischen Einparkhilfe mit OpenRAVE, mit heterogener Sensorik (Jeep: Laserscanner, Citroen: Ultraschall, Peugeot: Infrarot)



(b) 2D-View als Occupancy Grid Map, Darstellung in `rviz`

Abbildung 6. Repräsentationen verschiedener Ebenen, durch Anwendung von Filtern auf die allgemeine 3D-Darstellung (a) wurde der View (b) erzeugt

der Kollisionsvermeidung im Automobil [36] äußerst hilfreich.

Diese Beispiele hoch spezialisierter Applikationen nutzen feste Modelle der Umgebung, um die zuvor definierten Aufgaben, für die sie konzipiert wurden, erfüllen zu können. Momentan existieren keine einheitlichen Schnittstellen, die es diesen Systemen ermöglichen, ihre Wahrnehmung und ihre Modelle der Umgebung miteinander zu teilen. Ein Staubsaugerroboter kann sein Wissen über die Umgebung nicht mit einem Wakamru teilen, obwohl beide in sich überlagernden Arbeitsbereichen davon profitieren würden.

B. Trennung von Wahrnehmung und Applikation

In [37] wird versucht, Prinzipien der menschlichen Wahrnehmung auf mechatronische Systeme anzuwenden. Das vorgestellte System trennt strikt Konzepte der Wahrnehmung vom Bewusstsein. Wahrnehmung ist immer künstlich, es ist nur ein Modell der Umgebung, konstruiert auf der Basis der zur Verfügung stehenden sensorischen Informationen, Vorwissen und Annahmen. Bewusstsein kann auf mehreren Ebenen existieren, definiert durch Ziele, Bewertung, Planung und Vorhersagen sowie maschinelles Lernen.

Die Grundlage bildet auch hier der noch in Abschnitt III-C erläuterte Kreislauf in Intelligenten Umgebungen. Die Entscheidungs- und Kontrollalgorithmen werden hier nur durch eine Menge von fixen Regeln abgebildet, welche die sensorischen Eingaben direkt auf aktorische Ausgaben mappen, ein persistenter Datenspeicher ist ebenfalls vorgesehen. Bemerkenswert ist jedoch, dass Instanzen der Wahrnehmung und des Bewusstseins aus dem beschriebenen Kreislauf herausgelöst sind. Die Wahrnehmung wird durch zwei Instanzen repräsentiert, durch ein internes Modell des Systems und ein externes Modell der Umgebung (Weltwissen), die beide miteinander verzahnt sind. Eingaben für die beiden Modelle bilden nur die sensorischen Informationen. Der Wahrnehmung übergeordnet, liegen die eigentlichen Planungs- und Vorhersageinstanzen, die gänzlich von der sensorischen Wahrnehmung entkoppelt sind und als Eingabegrößen nur die beiden Systemmodelle nutzen.

Mithilfe dieser Art der Trennung können Kontrollalgorithmen und Problemlöser gänzlich von der Wahrnehmung entkoppelt implementiert werden und sind damit universell einsetzbar.

1) Reconfigurable Control Systems: Das Problem im Umgang mit Dynamik und Unsicherheit in der Umgebungswahrnehmung wird in [2] auf zwei Ebenen gelöst. Der im vorhergehenden Abschnitt aufgezeigte Ansatz der strikten Trennung von Perzeption und Bewusstsein wird für UAVs fortgeführt. Ein Sensor-Management-Modul verarbeitet, validiert und fusioniert die Roh-Sensordaten für die interne und externe Wahrnehmung. Die Bewertung abstrahierter Daten geschieht daraufhin in anderen High-Level-Modulen, die für die Situation Awareness, Fehlerdetektion und -isolation zuständig sind.

In Abhängigkeit von der aktuellen Situation, der Aufgabe und möglichen Fehlern geschieht dann die Auswahl von verschiedenen Betriebsmodi (auf Grundlage einer Regelbasis). Eine Situation wird hier durch einen Eigenschaftsvektor beschrieben (mit Elementen w. z. B. „Ist Lokalisation mit GPS möglich?“, Tageszeit, etc.). Diese Schicht kann auch als das höhere Bewusstsein des UAVs aufgefasst werden. Ausgehend von den ausgewählten Betriebsmodi wird das reaktive Verhalten (Trajectory Controller, Attitude Controller, etc.) bestimmt.

Die Abstraktion der Umgebung, die höheren Instanzen zur Verfügung gestellt wird, ist stark vereinfacht und auf die Bedürfnisse des UAVs zugeschnitten. Änderungen am System und der Sensorik sind durch die Nutzung von Modulen zwar möglich, haben aber immer Auswirkungen auf die Komponenten der höheren Ebenen.

2) Verteiltes Umgebungsmodell VerUM: Im VerUM-Projekt wurde ein System zur verteilten Datenhaltung und Analyse in Fahrzeugen [38] entwickelt. Die möglichen Situationen sind wie beim vorherigen Beispiel fest vorgegeben und stark abstrahiert. Eine Situation wird durch ein Tuple über den Mengen der vorgegebenen Aktionen, Interaktionen und Verkehrsregeln beschrieben. Eine Aktion kann ein Spurwechsel oder das Einparken sein. Einem Fahrzeug folgen oder Anfahren definieren eine Interaktion. Durch diese einfache Semantik können Situationen aufgrund von einfachen Sensordaten und Kontextwissen (Fahrsituation innerstädtisch, Autobahn oder Baustelle, etc.) einfach identifiziert werden. In [39] konnte

das Wissen um die aktuelle Situation genutzt werden, um das aufkommende Datenvolumen auf dem Bus signifikant zu reduzieren, da nicht in jeder Situation alle Sensordaten genutzt werden müssen. Es handelt sich im Automobil zwar um ein verteiltes System, die Anwendungen sind es jedoch nicht, da die Sensorkonfiguration für jede Anwendung fest vorgegeben ist und nicht geändert werden kann.

3) Middlewareeinsatz: In [40] wurde bereits vorgeschlagen, taskorientierte Wahrnehmung mithilfe einer Middleware zu realisieren. Auch hier ist Wahrnehmung von der Kontrollapplikation ausgeschlossen, es existiert jedoch ein Rückkanal, der es der Applikation ermöglicht, explizite Anforderungen an die Wahrnehmungsebene zu übermitteln, sodass sich irrelevante Sensorinformationen herausfiltern lassen.

C. Dynamische & Intelligente Umgebungen

Bis jetzt wurden nur einzelne und stark spezialisierte Systeme betrachtet. Im Folgenden soll diese Betrachtung auf intelligente und komplexe Umgebungen ausgeweitet werden, da diese per definitionem die dynamische Integration neuer Komponenten und den intelligenten Austausch von Informationen mithilfe von Adaption und Selbstkonfiguration unterstützen sollten. Eine Übersicht verschiedener Intelligenter Umgebungen wird in [41] gegeben. Das allgemeine Konzept einer solchen Umgebung ist ein erweiterter Perception-Control-Kreislauf. Basierend auf dem wahrgenommenen Umgebungszustand und verfolgten Zielen werden Schlüsse bezüglich möglicher Aktionen und Zustandsänderungen gezogen. Wahrnehmung ist hiernach ein Bottom-Up Prozess, wobei Sensoren die Umgebung erfassen und diese Informationen über eine Kommunikationsschicht verbreiten. Innerhalb einer Datenbank können die rohen Informationen gespeichert werden, um daraus durch andere Komponenten höherwertiges Wissen zu extrahieren. Dieses Wissen ist wiederum die Basis verschiedener Entscheidungs- und Kontrollalgorithmen, die die Ausführung diverser Aktionen delegieren (Top-Down). Eine Aktion wird mit Hilfe von Aktoren ausgeführt, die wiederum den Zustand der Umgebung und damit die Wahrnehmung verändern.

Unter allen Intelligenten Umgebungen zeigt das PEIS-Ecology Projekt [42] die größten Ähnlichkeiten mit unserem Vorhaben und soll daher etwas genauer beleuchtet werden. Es wird von intelligenten PEIS-Objekten (Physically Embedded Intelligent System) ausgegangen, seien es Sensoren, Aktoren oder einfach nur Objekte des alltäglichen Lebens (RFID-Tagged), die ihre Funktionalität anderen Objekten in einer sich dynamisch ändernden Umgebung zur Verfügung stellen können. Hierbei wird die Philosophie eines verteilten Roboters verfolgt, der zur Erfüllung einer Aufgabe dynamisch zusammengesetzt werden kann. Es ist zu beachten, dass die PEIS-Objekte innerhalb der Umgebung nicht a priori bekannt sein müssen und dynamisch hinzukommen und wieder verschwinden können. Die Heterogenität wird durch eine einheitliche und eventbasierte tuple-space Kommunikation gekapselt [43]. Selbstkonfiguration soll durch Selbstreflexion (eigenständiges Bewerten einer Situation) durchgeführt werden können. PEIS ist im Grunde eine konsequente Fortführung des Situationskalküls, siehe Abschnitt III-D. Die Bewertung einer Situation und die Erzeugung einer Konfiguration geschehen mithilfe von logischem Schließen. Hierzu liegt eine deklarative/logische Beschreibung des Systems, aller Komponenten und der möglichen Funktionalität vor. PEIS ist ein Top-Down-Approach. Die eigentliche Grundlage für eine intelligente Umgebungserfassung, die Wahrnehmung und die damit verbundene korrekte Einordnung und Prüfung von Sensorwerten wird hier nicht behandelt, sondern als Servicefunktionalität vorausgesetzt.

D. Situationskalkül

Einen gänzlich anderen Weg wird mit dem Situationskalkül beschritten, wobei eine Modellwelt mithilfe der Prädikatenlogik beschrieben wird. Dieses wurde in [44] eingeführt und in [45] weiterführend konzipiert. Dabei wird davon ausgegangen, dass das gesamte Umgebungswissen (Objekte der Umgebung, durchführbare Aktionen sowie die Auswirkungen von Handlungen, etc.) bereits vollständig in einer Wissensbasis vorliegt. Es wird vor allem für das Finden von Handlungssequenzen, die zum Erreichen eines bestimmten Zieles führen, genutzt. Zeitliche Aspekte, wie die Dauer einer Handlung, lassen sich hiermit nicht behandeln. Die erste sprachliche Implementierung erfolgte mit GoLOG (alGOL in LOGic) in [46].

IV. Diskussion

Intelligente Handlungsweisen von autonomen Systemen in komplexen Umgebungen erfordern eine intelligente Wahrnehmung. Diese benutzt eine allgemeine Repräsentation der Umwelt zur geometrischen Einordnung der Messdaten und zu deren Validierung. Entsprechend dem verteilten Ansatz der geschilderten Szenarien hat dies in einem dezentralen und dynamisch adaptiven Kontext zu geschehen.

Wie in Abschnitt III anhand der verwandten Arbeiten gezeigt, existiert dafür kein generelles Konzept, das eine vollständige Trennung von Umgebungswahrnehmung und Applikation unterstützt. Die aufgeführten Lösungen sind auf einen konkreten Einsatzbereich fokussiert und nicht allgemein gültig anwendbar. Vor diesem Hintergrund wurde in Abschnitt I und II ein mehrschichtiger Ansatz beschrieben, der als Kernelement eine generische Umgebungsrepäsentation ermöglicht. Die dafür nötigen und bereits umgesetzten Teilschritte wurden in den zuvor benannten Publikationen beschrieben. Die Integration der verteilten Datenhaltung sowie die vollständig automatische Transformation dieser Daten bedarf einer Weiterentwicklung der bestehenden Konzepte und Implementierungen.

Acknowledgment

This work is (partly) funded by the German Ministry of Education and Research within the project ViERforES-II (grant no. 01IM10002B).

LITERATUR

[1] S. Li, K. Li, R. Rajamani, and J. Wang, "Multi-objective coordinated control for advanced adaptive cruise control system," in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*. IEEE, 2009, pp. 3539–3544.

[2] L. Wills, S. Kannan, S. Sander, M. Guler, B. Heck, J. Prasad, D. Schrage, and G. Vachtsevanos, "An open platform for reconfigurable control," *Control Systems Magazine, IEEE*, vol. 21, no. 3, pp. 49–64, 2001.

[3] D. Nakhaeinia, S. Tang, S. Mohd Noor, and O. Motlagh, "A review of control architectures for autonomous navigation of mobile robots," *International Journal of the Physical Sciences*, vol. 6, no. 2, pp. 169–174, 2011.

[4] S. Zug, M. Schulze, A. Dietrich, and J. Kaiser, "Programming abstractions and middleware for building control systems as networks of smart sensors and actuators," in *Proceedings of Emerging Technologies in Factory Automation (ETFA '10)*, Bilbao, Spain, 9 2010.

[5] M. Schulze, "Adaptierbare ereignisbasierte Middleware für ressourcenbeschränkte Systeme," Doktorarbeit, Fakultät für Informatik, Otto-von-Guericke Universität Magdeburg, 2011.

[6] S. Zug, "Architektur für verteilte, fehlertolerante Sensor-Aktor-Systeme," Doktorarbeit, Fakultät für Informatik, Otto-von-Guericke Universität Magdeburg, 2011.

[7] R. Smith. (2007) The open dynamics engine. [Online]. Available: http://ode.org

[8] W. Meeussen, J. Hsu, and R. Diankov. (2012, 4) URDF - Unified Robot Description Format. [Online]. Available: http://www.ros.org/wiki/urdf

[9] R. Diankov, "Automated construction of robotic manipulation programs," Ph.D. dissertation, Carnegie Mellon University, Robotics Institute, 8 2010.

[10] A. Dietrich, S. Zug, and J. Kaiser, "Detecting External Measurement Disturbances Based on Statistical Analysis for Smart Sensors," in *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE)*, 7 2010, pp. 2067–2072.

[11] S. Zug, A. Dietrich, and J. Kaiser, *Fault-Handling in Networked Sensor Systems*. St. Franklin, AUS: Concept Press Ltd., 2012.

[12] A. Dietrich, S. Zug, and J. Kaiser, "Modelbasierte Fehlerdetektion in verteilten Sensor-Aktor-Systemen," in *11./12. Forschungskolloquium am Fraunhofer IFF*. Fraunhofer Institut für Fabrikbetrieb und Automatisierung (IFF), 2011.

[13] ——, "Model based Decoupling of Perception and Processing," in *ERCIM/EWICS/Cyberphysical Systems Workshop, Resilient Systems, Robotics, Systems-of-Systems Challenges in Design, Validation & Verification and Certification*, Naples, Italy, 9 2011.

[14] S. Zug, M. Schulze, A. Dietrich, and J. Kaiser, "Reliable Fault-Tolerant Sensors for Distributed Systems," in *Proceedings of the Fourth ACM International Conference on Distributed Event-Based Systems (DEBS '10)*. Cambridge, United Kingdom: ACM Press New York, NY, USA, 7 2010, pp. 105–106.

[15] S. Zug, A. Dietrich, M. Schappeit, C. Steup, and J. Kaiser, "Flexible Daten-Akquisition & Interpretation für verteilte Sensor-Aktor-Systeme im Produktionsumfeld," in *10. Magdeburger Maschinenbauertage*, 9 2011.

[16] S. Zug and A. Dietrich, "Examination of Fusion Result Feedback for Fault-Tolerant and Distributed Sensor Systems," in *IEEE International Workshop on Robotic and Sensors Environments (ROSE 2010)*, Phoenix, AZ, USA, 2010.

[17] S. Zug, A. Dietrich, and J. Kaiser, "An Architecture for a Dependable Distributed Sensor System," *IEEE Transactions on Instrumentation and Measurement*, vol. 60 Issue 2, pp. 408 – 419, 2 2011.

[18] V. Köppen, S. Zug, A. Dietrich, and M. Mory, "Business-Management-Inspired Sensor Data Fusion," in *International Conference on Wireless Technologies for Humanitarian Relief (ACWR2011)*, Kerala, India, 12 2011.

[19] S. Zug, C. Steup, A. Dietrich, and K. Brezhnyev, "Design and implementation of a small size robot localization system," in *IEEE International Symposium on Robotic and Sensors Environments (ROSE 2011)*, Montreal, Quebec, Canada, Sep. 2011.

[20] (2012) The apache cassandra project. [Online]. Available: http://cassandra.apache.org/

[21] A. Dietrich, M. Schulze, S. Zug, and J. Kaiser, "Visualization of Robot's Awareness and Perception," in *First International Workshop on Digital Engineering (IWDE)*. Magdeburg, Germany: ACM Press New York, NY, USA, 6 2010.

[22] T. Kiebel, A. Dietrich, M. Schulze, S. Zug, and J. Kaiser, "Identifying patients and visualize their vitality data through Augmented Reality," in *The 7th IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS '10)*, San Francisco, CA, USA, 11 2010.

[23] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, vol. 3, no. 3.2, 2009.

[24] T. Foote, E. Marder-Eppstein, and W. Meeussen. (2012, 4) tf - ros. [Online]. Available: http://www.ros.org/wiki/tf

[25] D. Hershberger and J. Faust. (2012, 5) rviz - ros. [Online]. Available: http://www.ros.org/wiki/rviz

[26] R. C. Smith and P. Cheeseman, "On the Representation and Estimation of Spatial Uncertainty," *The International Journal of Robotics Research*, vol. 5, no. 4, pp. 56–68, 1986.

[27] D. Hähnel, W. Burgard, and S. Thrun, "Learning compact 3d models of indoor and outdoor environments with a mobile robot," *Robotics and Autonomous Systems*, vol. 44, no. 1, pp. 15–27, 2003.

[28] H. Surmann, A. Nüchter, and J. Hertzberg, "An autonomous mobile robot with a 3d laser range finder for 3d exploration and digitalization of indoor environments," *Robotics and Autonomous Systems*, vol. 45, no. 3, pp. 181–198, 2003.

[29] S. Thrun, W. Burgard, and D. Fox, "A real-time algorithm for mobile robot mapping with applications to multi-robot and 3d mapping," in *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, vol. 1. IEEE, 2000, pp. 321–328.

[30] R. Rusu, Z. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927–941, 2008.

[31] R. Rusu, Z. Marton, N. Blodow, A. Holzbach, and M. Beetz, "Model-based and learned semantic object labeling in 3d point cloud maps of kitchen environments," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 3601–3608.

[32] K. Hsiao, N. Mavridis, and D. Roy, "Coupling perception and simulation: Steps towards conversational robotics," in *International Conference on Intelligent Robots and Systems*, vol. 1. IEEE, Oct 2003, pp. 928–933.

[33] D. Roy, K. Hsiao, and N. Mavridis, "Mental imagery for a conversational robot," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 3, pp. 1374–1383, 2004.

[34] E. Yoshida, C. Esteves, I. Belousov, J. Laumond, T. Sakaguchi, and K. Yokoi, "Planning 3-d collision-free dynamic robotic motion through iterative reshaping," *Robotics, IEEE Transactions on*, vol. 24, no. 5, pp. 1186–1198, 2008.

[35] H. Yu, R. Beard, and J. Byrne, "Vision-based navigation frame mapping and planning for collision avoidance for miniature air vehicles," *Control Engineering Practice*, vol. 18, no. 7, pp. 824–836, 2010.

[36] N. Kaempchen, B. Schiele, and K. Dietmayer, "Situation assessment of an autonomous emergency brake for arbitrary vehicle-to-vehicle collision scenarios," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 10, no. 4, pp. 678–687, 2009.

[37] H. Caulfield and J. Johnsonb, "Artificial perception and consciousness," in *Sixth International Conference on Education and Training in Optics and Photonics: 28-30 July 1999, Cancún, Mexico*. Society of Photo Optical, 2000, p. 112.

[38] A. Hermann and J. Desel, "Driving situation analysis in automotive environment," in *Vehicular Electronics and Safety, 2008. ICVES 2008. IEEE International Conference on*. IEEE, 2008, pp. 216–221.

[39] A. Hermann, "Fahrsituationsspezifische Datenverteilung im Verteilten Umgebungsmodell für Fahrzeugsoftware," in *Lecture Notes in Informatics. GI-Jahrestagung*, 2007, pp. 541–545.

[40] A. Rotenstein, A. Rothenstein, M. Robinson, and J. Tsotsos, "Robot middleware must support task-directed perception," in *Proc. ICRA 2nd Int. Workshop on Software Development and Integration into Robotics, Rome, Italy*, 2007.

[41] D. Cook and S. Das, "How smart are our environments? an updated look at the state of the art," *Pervasive and Mobile Computing*, vol. 3, no. 2, pp. 53–73, 2007.

[42] A. Saffiotti, M. Broxvall, B. Seo, and Y. Cho, "The peis-ecology project: a progress report," in *Proc. of the ICRA-07 Workshop on Network Robot Systems, Rome, Italy*. Citeseer, 2007, pp. 16–22.

[43] M. Broxvall, M. Gritti, A. Saffiotti, B. Seo, and Y. Cho, "Peis ecology: Integrating robots into smart environments," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 212–218.

[44] J. McCarthy, "Situations, actions, and causal laws," Stanford University Artificial Intelligence Project, Stanford, California, Tech. Rep., 1963.

[45] J. McCarthy and P. J. Hayes, *Machine Intelligence*, 4th ed. Edinburgh University Press, 1969, ch. Some philosophical problems from the standpoint of artificial intelligence, pp. 463–502.

[46] H. Levesque, R. Reiter, Y. Lespérance, F. Lin, and R. Scherl, "GOLOG: A logic programming language for dynamic domains," *The Journal of Logic Programming*, vol. 31, no. 1-3, pp. 59–83, 1997.

Hypermodelling

A Data Warehouse Approach For Software Analysis

Tim Frey
Otto-von-Guericke-University Magdeburg, Germany
tim.frey@tim-frey.com
Supervisor: Gunter Saake

***Abstract*— Separation of concerns is a widespread principle for good software design. Research showed limitations in object oriented programming and refers to a multi-dimensional concern space. In this work, we present the Hypermodelling approach that utilizes Data Warehouse technology to explore and analyze the concern space of a program. We give an overview about the associated work, applications and the potential benefit of Hypermodelling. Now, managers can control software structure with Data Warehouse technology. Developers can use Hypermodelling for code search and software analysts can investigate a software project with multi-dimensional operations at various abstraction levels. Future research paths indicate potential synergies with other approaches and the possibility to uncover additional application scenarios.**

Separation of Concerns, Hypermodelling, Code Reporting, Data Warehouse, Orthographic Software Modeling

I. Introduction

Software engineering focuses in the creation of software systems, i.e. programs [1]. Software engineers, often called developers or programmers, encode functionality to create such a program [2]. However, the creation of a program is a complex scenario that contains many obstacles to achieve the desired quality and functionality of the program [3,4]. Literature defines a software system sometimes in broad terms as the outcome of a development effort, containing software and hardware [2]. Anyway, in this work, we focus on the challenges that are faced within the software development process. In special, we concentrate on the problems, directly associated with the manufactured software system that is created by writing source code.

Nowadays, the complexity in programming is faced with modules [5]. Literature refers to a module as a collection of algorithms and data structures to do a closed operation within itself. Furthermore, the correctness is also achievable without knowledge about a certain programming system [6]. Often literature refers to concrete mechanisms in a programming language, like packages or Java classes as modules [7,8]. In this work, we focus mainly on the Java language as proxy for other object-oriented languages and its modularization techniques. Therefore, we define the term module as a fixed set of functionality, like a Java class or function. Such a module encapsulates a certain functionality that can be summoned or composed with other modules by language means. The purpose of a module is to enable reusability, comprehensibility, interchange and testing. Hence, a module follows the purpose of encapsulation of a set of functionality. This enables a reuse of this functionality and the composition with other modules to derive more complex functionality. Through ability to compose different modules together also the interchange of modules gets easier. Furthermore, modules can be tested separately to ensure that their desired functionality works in the module itself. In this work we focus mainly on the general idea how modules are composed together and functionality is split into modules.

Modern code bases consist out of millions lines of source code that are belonging to plenty of modules[1]. One main challenge is the immense size of that code bases. Those are hard to maintain, to overview and to navigate. Developers face the challenge to investigate and extend such code bases. The main issue with this complex code bases is that developers spend most time reading and searching source code. Out of this reason, we concentrate on a new code base investigation and analysis technique to advance software engineering. Thus, our contribution is a new code base investigation technique called Hypermodelling that allows investigating code bases with Data Warehouse technology. Now, developers can use Data Warehouse technology that was built of big data to analyze huge amounts of source code. Furthermore, we show advanced application scenarios where Hypermodelling can be applied.

In the following, we give a brief overview about the inspiring, motivating prior work that led to our called Hypermodelling approach. Thereby, we reveal two main questions that our research answers. Afterwards, we present insights into the Hypermodelling approach and point out different application scenarios. Lastly, we do conclusions and give an outlook to future work.

II. State of the Art

We embed our research into the related work by presenting the motivating and related research that led to our technique. We start by describing separation of concerns and the multi-dimensional viewpoints there. Then, we describe Orthographic Software modeling that also follows a multi-dimensional idea and describe its addition to separation of

[1] http://de.wikipedia.org/wiki/Lines_of_Code

concerns. Afterwards, we depict the role of frameworks within software development and formulate the demand that those need to be considered in software investigation scenarios. As related multi-dimensional technique, we describe Data Warehousing and its application to built dashboards for issue tracking systems. Lastly, we derive our research questions out of the demands that are given through the prior work.

A. *Separation of Concerns*

Separation of concerns is the approach to regard a software system out of separate vantage points. The goal of this principle is to study every aspect of a system in its own [9]. Hence, developers try to separate concerns into their own module to enable to study them on their own. Through the composition of these different aspects a software system is created. Different programming paradigms offer various mechanisms to decompose functionality into modules. Hence, programmers face the challenge to identify the concerns of a system and to create modules that cover a certain concern and then compose these concerns together.

The possibility to apply separation of concerns within a program is depending on the modularization techniques of a programming language and the skills of a developer. Either way, both are often not perfect. The concerns that programmers want to encode in a program evolve during a project. There, new and unplanned concerns arise and change and programmers have to adept the program to the new desires. This leads often to imperfect separated programs. Otherwise also not all program paradigms enable a perfect separation of concerns through limited modularization techniques [10]. Therefore, concerns are traditionally interwoven within object oriented modules. This way, modules are responsible at the same time for multiple concerns or a concern is distributed over various modules [11, 12].

In order to overcome this limitation of imperfect separation, multiple approaches introduce multi-dimensional modularization techniques, like aspect-oriented [13] programming or the Hyperspace [14] approach. The multi-dimensional approaches share altogether that they see the concern space of a program as a multi-dimensional space where the concerns influence each other [15,16]. The composition of the concerns then results in the final software system.

Therefore, we note the following conclusion. Concerns are interwoven within modules of software. Concerns represent a multidimensional space that can be expressed by multi-dimensional modularization techniques. In case of object oriented programming a module often is associated with multiple concerns at the same time and it is not possible to regard every concern on its own through the lack of modularization. This violates the main idea of separation of concerns where each concern should be studied on its own.

B. *Orthographic Software Modelling*

In current development environments developers use a collection of graphical models, which are loosely related to each other. The different models (UML- Diagrams) represent different perspectives of a software system to a developer [17]. Every view of a model provides different aspects of the system and when one model is altered this may influence what is shown in another view. Therefore, the problem is that the maintenance of these views is creating consistency, management and navigation problems.

The reason for this is that commonly the models are only partly related on the same data model. So, when the data in one model is changed it may not influence another related representation model what would be needed to ensure consistency. Furthermore, the views, used for presenting models to the developer, have no interaction in their navigation with each other. Currently, the linkage of the views to the implementation and the interaction of those have to be done by the developers in their heads. Developers know which aspect is represented in different contexts in the diverse views, but their navigation on those is not supported by tools. Also, developers know how the models relate to each other and in which model which property is defined. Thus, each view has to be navigated on its own without the ability to combine the navigation and interaction for the various views. We can see the lack of consistent contextual model tooling support for a clearer design of interaction for models and views [18].

The Orthographic Software Modeling (OSM) principle is trying to overcome this limitation by introducing a multi-dimensional viewpoint towards a software system. This viewpoint is based on an analogy to the Computer Aided Manufacturing (CAD) where different projections of the same facts are presented to engineers. Engineers model each perspective on its own and through their relations a holistic model is built. In case of software the diverse views of CAD can be the various UML models. Therefore, the main idea is that the different projections (models) of the software system can be used to manage the complexity [17, 18]. In order to enable this idea, a consistent model out of which these views are created is needed to ensure the consistency between the various views. To create the views out of this model, different mechanisms are needed to provide the ability to generate the views. Moreover, OSM proposes additionally a coherent navigation and management mechanism to navigate between the various views. There, different abstraction levels and hierarchies of models that relate to each others are part of the navigation. This relation of projections serves as foundation to realize an integration of the various perspectives and models at the same time. In general, the goal is to define a holistic model that is used as source for the diverse model projections.

All together, we see OSM as an addition to the pure multi-dimensional viewpoint of (multi-dimensional) separation of concerns to the explicit claim to raise abstraction, whereby SOC mainly deals with the separation.

C. Frameworks

Another mean that is used to advance and speed up programming are Frameworks. Frameworks offer (half) ready application frames that developers can use to embed their own application logic in [19]. Commonly, such ready to use functionality is offered in Java as jar files that can be added to an application and the functionality within them can be used. The usage of this functionality is done through modularization techniques. For example, objects of the framework can be instantiated, methods called or parent classes of the framework get inherited. This way, custom code of a specific application can be easily weaved together with framework code [20].

Developers know the functionality of the frameworks and how they associate their own application logic with it. When developers read code they understand with which components of a framework the custom code is associated. An application integrates into the diverse layers of the framework. Through the known framework architecture inferences to which layers application logic belongs can be built. When a developer reads code of a class that extends a data access class of a framework he knows that the extended class is belonging to the persistency layer of an application. Therefore, when we think about code analysis, there are structures already embedded in frameworks that are well known by developers. Hence, this already existing knowledge should be used within code analysis.

When we consider a relation to OSM we can see frameworks as one mean of abstraction of the traditional modularization means to apply SOC. However, currently Frameworks are just a pure implementation mean and no explicit viewpoint towards a software system. Frameworks offer ready to use programmatic models of common functionality within programs and we lack to utilize this model as a viewpoint when we think about them. Thus, we see the need that to credit separation of concerns that a software system should be capable to be studied from every concern of its own. Frameworks are such kind of viewpoint and their logic is clearly a concern that should be explicitly considered by a future concern investigation technique.

D. Data Warehousing

Data Warehouses are mostly used within business scenarios within enterprises. They integrate the data from various sources into a central repository. In order to load data into the Data Warehouse and prepare it for analysis a so called Extract transform load process is used [21]. Within this process the data from the various data sources is extracted and transformed into the right schema of the Warehouse and loaded into it. In there, the data gets staged and loaded into so called data cubes. Data cubes are multi-dimensional structures that can be analyzed from different perspectives [22]. Every part of the data and their relation may be treated as a dimension within these cubes. The data in these cubes is analyzed with queries that allow multi-dimensional operations to reveal relations of the data. Selections, filtering of dimensions as well as navigation between different dimensional hierarchies (called Rollup, Drill-down or Drill-across), are possible. Measures between the data indicate the relations of the dimensions. Through the central approach of measures within such data cubes the hierarchies of the data can be used to apply abstractions. Every dimension of the data can be a center of analysis and source of an aggregation towards the other dimensions. Through this approach the navigation path how the dimensions relate to each other is easily and efficiently accessible. This enables us at the same time to generate reports on actual and aggregated data. Often, there is also tool support to create reporting dashboards, containing graphical elements that can be used by people that do not know how to build queries manually. Out of these reports decisions are made and the business plans are adapted.

In order to achieve specific future business goals out of the reports, Data Warehouses are used for advanced planning scenarios. There, future business figures are persisted within the Data Warehouse and used to indicate how the data dimensions should relate to each other, based on figures for their future. Through the business realization of the future plan new data is generated within the operative systems (for instance, the amount of cars that are sold within a region). This data is then loaded in the Warehouse, again, as addition to the already existing one. This way, the old and the new data can be used in reports and queries. In such queries the actual reality can be compared with the planned future to determine the progress of the future plan.

All together, Data Warehousing is about a multi-dimensional perspective to data. In general, we see the multi-dimensional viewpoint of Data Warehousing as similar to the multi-dimensional viewpoint of concerns with the addition of an explicit abstraction method in Data Warehousing. However, the two viewpoints have not been merged, yet. Also, we see the viewpoint of OSM as another multi-dimensional viewpoint representative of software. OSM uses models as abstraction whereby Data Warehousing uses measures and dimensional data hierarchies as mean of abstraction.

E. Software Project Control Centers

In software development project managers, quality assurance and deciders need to overview a big amount of data. Currently, software development projects lack a holistic approach to provide relevant and filtered information for specific roles at a central point. In order to advance project steering, an approach is needed that allows reproducible abstractions of the relevant information to estimate the progress of an entire project at once. Thus, different work follows the goal to introduce so called management dashboards, i.e. software cockpits, to enable a better overview over software projects [23, 24, 25]. This dashboards show custom indicators about the status of a software development project and support the controlling of a projects business goals. The general vision of such a

cockpit is similar to one in a plane that offers the relevant information for the captains and his co-pilots and allows them to steer the plane and get all relevant information for decisions at once. In order to provide analysis support about the project Data Warehouses are used to store extracted data that provides facts about a software project.

In the case of current cockpits [23, 24], a small excerpt of the code structure is stored in the Warehouse to allow drill downs to the positions where the different metrics occur (a single hierarchy: projects down to classes and methods). Managers and developers can use those dashboards to gain an overview of the development progress. However, first evaluations showed that a cockpit is useful in projects. One main advantage in contrast to other analysis tools is the integration of data out of various sources together in the cockpit. Furthermore, structured interviews of involved employees led to confident viewpoints about the positive effects of the introduced cockpit. More detailed industrial applications and empirical evaluations of a small panel indicate further benefits of a software cockpit [24].

All together, code cockpits advance software project management. The developed code cockpits applied Data Warehousing in traditional means. Only project related data was used to analyze the software projects and the multi-dimensionality of the concern space was kept aside. Anyway, the usefulness of code cockpits indicates that cockpits about software contain a general usefulness in practice and we see the need to consider cockpits for source code and its multi-dimensional structure.

F. Conclusion - Demands on a new technique

We described the multi-dimensional concern structure of source code and the difficulty and the demand to be capable to study every concern on its own. OSM added the general idea of abstraction and multi-dimensional navigation of a software system within development. Data Warehousing technology offers high performance infrastructures that offer support for scale-ups and scale-outs. A first use case of software cockpits built with Data Warehousing indicated a usefulness of that technology within software engineering.

We see the need to add the idea of abstraction that OSM introduced to the general thought of separation of concerns. We propose to reformulate the original statement, to study every concern on its own, with the demand that every concern should also be capable to be studied in abstraction. Furthermore, we see the necessity to consider a multi-dimensional navigation as method concern exploration. OSM described modeling and navigation problems. As additional fact we introduced that frameworks are already certain sets of functionality embedded in source code itself. Lastly, we saw that there is a potential benefit in using reports in software development projects.

With Data Warehousing we have a successful method at hand that is already applied successfully in other areas in enterprises. This technology allows multi-dimensional navigation and also abstraction. Additionally, successful approaches have been made to apply the technology in software engineering. Therefore, we see the need to explore this technology in combination with the principle of separation of concerns. In order to evaluate if the technology can be applied for abstraction and navigation and has an advanced usefulness within software development, aside the obvious, we postulate two main questions:

1.) Can a program be loaded into a Data Warehouse and how can we express therein concern relations?
2.) Are there further benefits aside abstraction and navigation for software engineering? Can we depict further application scenarios that open up further research?

Like we can see on the research questions, the demand on the contribution of the thesis is the technical realization and further applications of the technique. Hence, this thesis embeds into the research of separation of concerns, but also grazes other work in the area of software engineering to demonstrate the usefulness of the application of Data Warehouse technology.

III. HYPERMODELLING

In this section we describe briefly the research about our Hypermodelling technique that utilizes Data Warehouse technology for software investigations. All already published results [26, 27, 28, 29] can be found at the Hypermodelling project homepage[2]. Furthermore, several videos[3] show the technology in action. Lastly, an access to the demos can be requested by the author. Hence, this section is a short sum up, what was already achieved and will be described in the final thesis.

First, we give a brief insight how we utilize Data Warehouse technology to express concern associations in software. This technical viewpoint corresponds to the first research question. Afterwards, we describe a general framework how Hypermodelling embeds into Data Warehouse technology. Thereby, we reveal the diverse areas how it may be applied within software engineering. In the following, we give peak insights of the diverse applications that we built and the affected areas. Through this, we show that there is an integrative benefit of the application of Data Warehouse technology in diverse areas of software engineering, what corresponds to our second research question.

At the end of our thoughts about separation of concerns, we ask ourselves how it relates to the human mind. The arising question is, how the mind structures information and if there is research in psychology available that goes in the same direction. Therefore, we give lastly an excursus about program comprehension out of research in psychology to look a bit beyond our own nose.

[2] http://hypermodelling.com
[3] http://www.youtube.com/user/hypermodelling

A. *Technology – Brief insight*

Our main approach to apply Data Warehousing technology is based on the idea is that a source code fragment belongs to multiple concerns at the same time. Hence, in Hypermodelling a class or a method can be associated with multiple concerns at the same time. All fragments that are associated with a specific concern are members of a slice belonging to this concern. In Data Warehousing those slices are called dimensions. This leads to the technique to use Data Warehouse similar methods to query for fragments, belonging to one or more concern. The main problem is that the associations are expressed with measures in a Data Warehouse, whereby the associations in source code are done by means or modularization.

Our solution is to express any kind of association of code fragments as a count of "1". We use the dimensions as structural points in a program and the associations as a count to build up a relation matrix. For us, the association of a class with concerns, like its parent classes, is the same like the associations of measures with dimensions. Hence, the class association with its parents is a connection through the measure "1".

```
1: @Entity
2: @Deprecated
3: class Customer{
4:         @Deprecated
5:         Customer(){
6:         …}
7: …}
8: class CustomerDAO extends DaoSupport
9:
10:        @SuppressWarnings("deprecation")
11:        Customer createCustomer(){
12:                return new Customer();
13:        }      ...}
```

Listing 1. Example for annotated source code

Table 1. Concerns of Listing 1 in a table

Element / Dimensions	extends DaoSupport	@Entity	@Deprecated	@SuppressWarnings
Customer	-	1	1	-
Customer.Customer()	-	-	1	-
CustomerDAO	1	-	-	-
CustomerDAO. createCustomer()	-	-	-	1

We present an example in Listing 1 and Table 1. Listing 1 shows Java source code. There, we use the Java Metadata annotations [30] and inheritance to show associations. A Customer class and a Data-Access-Object (DAO) are shown. The CustomerDAO class extends a helper class (DaoSupport) for table access. This is commonly done when frameworks are used. Table 1 shows rows that are representing source code fragments and columns representing concerns. The "1" indicates that a fragment belongs to a concern. For example the constructor of the class Customer is marked @Deprecated. Likewise the rest of the table-listing associations can be done.

Furthermore, also hierarchies exist in source code. A class, for instance, is member within a hierarchical package structure. This way, source code shows huge similarities to Data Warehouse data structures. In a Data Warehouse the dimensions can be ordered into hierarchies and we can use such hierarchies to describe the hierarchies in source code.

All together, once the source code is brought into such a relation structure within a Data Warehouse all the means within a Data Warehouse can be used to investigate the source code from any perspective. This way, we credit the claim of to study every concern on its own through introducing the multi dimensional exploration of data within a Data Warehouse. Furthermore, we credit the demand of OSM to have abstraction through the possibility of Data Warehouses to aggregate dimensions and therefore the relation of concerns to each other. Lastly, Data Warehouses technology also allows using dimensional hierarchic structures that exist within source code as point of exploration and navigation. This way, we fulfill our demands to the new techniques. Our work provides details about the concrete technical application and the specific data cubes and how source code can be structured and associations can be computed within a Data Warehouse.

B. *The Hypermodelling Framework*

In order to use the whole toolbox of Data Warehouse technology, we abstracted our idea to a generic framework. We present this framework in Figure 1. There, we see that Data of various sources is extracted into a Data Warehouse.

We use a relational schema as base that contains most of the associations within source code in the center. The main reason to use a relational schema is to be compatible with most Data Warehouses that are often built on relational Databases. This relational schema makes it possible to integrate data from various other sources that are associated with source code into the Warehouse, too. Since the relational source code model is in the center, every kind of data that has a relation to it, can reference it. Our work is going to provide the relational schema for most of the Java source code structure.[4]

[4]We use mainly the relations of Methods, classes, method parameters and annotations and inheritance within our schema.

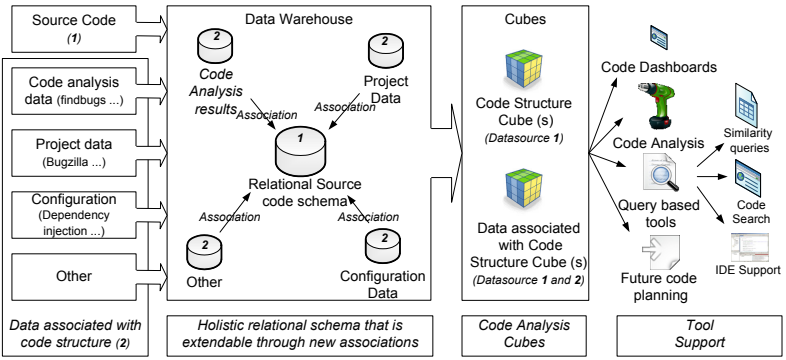


Figure 1. Hypermodelling Tool Framework

Out of the relational schema, we load the structure into the multi-dimensional cubes. One option is to create a cube that contains all the source code structure. Another option is to build cubes, containing source code structure as well as data from external tools. This way, we utilize the possibility to investigate "external" data and its relation to source code structure. In our work, we present a holistic cube for the source code structure that can be used as base to create smaller cubes for specific investigations or as reference to derive new cubes. Furthermore, we are going to show cubes that contain data of other tools. On top of the cubes, various applications can be built.

All together, the variety of applications shows that Hypermodelling enables future research. In order to give deeper insights and to support this point, we built exemplary research and prototypes for those. This way, we depict the diverse areas that Hypermodelling affects. This provides support to answer our second research question: Hypermodelling provides further benefits and enables further research. In the following, we provide an overview about our applications and research that we describe in our detailed work. There, we show exemplary application scenarios for each area.

1) Code Analysis

Framework manufacturers face the challenge to determine which parts of frameworks are used and varied. Application developers want to know on which framework elements their application is depending. Currently, programs need to be parsed to extract information about framework usage what consumes time and effort and makes information mining inflexible. Hypermodelling overcomes the current limitations. We demonstrate that Hypermodelling is suitable to explore software variance. We present reports based on real application data of one project example to reveal multiple facts about the software variance. We show visualizations at different granularity levels. This supports our theory that Hypermodelling can be used to explore software variance in an easy way. A detailed description about our work can be found in [29].

2) Code Cockpits: Hypermodelling Reporting

Do we have a vendor lock in? How many classes of a framework do we extend in our code? These questions may be asked by software development managers when talking about an application development. In order to reveal such facts, a lot of effort in application investigation is needed. We overcome this by presenting the usage of the Hypermodelling approach to built software cockpits. We present a cockpit for the variance of software that is based on our research about software variance. We reveal a schematic cockpit view and evaluate the effort to implement it. Project managers can now use our cockpit to investigate software variances more easily. Important indicators about variance can now be investigated at a central spot. This avoids costly, time-consuming and deep investigations in the first place. Further investigations can discover cockpits for other roles to cover and control the whole development cycle in a holistic approach. Furthermore, the reasonable effort to create such cockpits enables the possibility to create different cockpits and evaluate or compare the usage of those. A live demo is available on request and further information and videos of cockpits can be found at our homepage.

3) Code Search and Similarity

Code recommendation systems ease programming by proposing developers mined and extracted use cases of a code base [31]. Currently, recommendation systems are based on hardcoded datasets what makes it complicate to adapt them. Another challenge is the adaptable live detection of code clones. In order to overcome current limitations, we advance clone detection and code

recommendation systems by presenting the utilization of the Hypermodelling approach. We present the generic idea to advance recommendation and clone detection based on queries and evaluate our application with industry source code through demo queries [36]. Consequently, recommender systems and clone detection can be customized with flexible queries via Hypermodelling. This enables further research about more complex clone detection and context sensitive code recommendation.

Additionally, we show how Hypermodelling can be used as code search engine that is available online[5] and based on queries internally. Different parameters can be used to discriminate the search results to demonstrate the idea. Furthermore, we show style issues in the search results that show that such kind of data could also be used in code search scenarios.

4) IDE (Integrated Development Environment) Support

Imagine a developer, who wants to alter the service layer of an application. Even though the principle of separation of concerns is widespread not all elements belonging to the service layer are clearly separated. Thus, a programmer faces the challenge to manually collect all classes that belong to the service layer. We present how the Hypermodelling approach can be used to query the necessary code fragments that belong to the service layer. Thereby, we show how framework information can be used in queries. The service layer can now be uncovered just via a query.

Further information can be found in the associated publication [28] and the videos online[6] that show the plug-in for the IDE in action.

5) Future Code Planning: Hypermodelling the Future

Imagine a developer, who modularizes a traditional object oriented program with annotations. Can we plan the migration and the adaption to the new mechanisms? Can we measure progress of our migration plan? We present an outlook for a method that can be used to plan the realization of concerns with figures, like it is done in enterprise planning scenarios. We evaluate our preliminary approach at an excerpt of a demo application that is upgraded against a new framework version. Our contribution enables further research about the planning of future program versions and if Hypermodelling can be used in more complex scenarios. Furthermore, we propose investigations about the necessary management support and the needed indicators to measure the progress of a concern encapsulation association movement.

C. Excursus: Concerns as Categories in Psychology

Lastly, after studying a lot of literature about separation of concerns we have to ask why separation of concerns is such a desirable goal. Dijkstra formulated the goal out of the way how scientists think. Nevertheless, software is created by human developers and we have to consider human aspects in software engineering. Research about program comprehension gives indications of how developers study source code. However, in psychology plenty of research about cognition exists. Our postulations address the need for an advanced software engineering technique and the need for abstraction and for a multi-dimensional viewpoint. Therefore, we see the need to make an excursus to widen our viewpoint of information structuring in psychology. We reveal the main theories there and derive an exemplary program comprehension model based on it [32, 33, 34]. We show that psychologists use different theories, how information is structured and that content is not classified sharply. Furthermore, the context is influencing the cognition. We present our research as a program comprehension model.

All together, this research supports our Hypermodelling attempt. Hypermodelling offers to regard a program from different perspectives at different abstraction levels and to allow multi-concern associations at the same time. Furthermore, from what we depicted in psychology, the multiple viewpoints that are altered depending on the context and fuzzy categories support the usefulness of our technique. We see Hypermodelling as a view technology that meets the demand of what different theories depict as categories in psychology.

IV. Conclusion and Future Work

We described the background and the inspiration to derive the Hypermodelling technique. Thereby, we showed that different approaches target different solutions. We combined excerpts of those approaches to develop Hypermodelling. We got into details how concern associations in source code can be expressed in a multi-dimensional way. Now, developers can use Data Warehousing technology and its whole tool set to investigate source code. Different application scenarios depicted the need for future research and the usefulness of the approach. Diverse areas can benefit from it. Related research about categories in psychology shows a relation to separation of concerns and supports our approach that we need a technology that supports multiple viewpoints.

However, we see the future work threefold. First, talks with developers in the industry brought us interested feedback and we think that Hypermodelling can serve as integration technology there. More business and a consistent tool chain of code reports and code investigation should be derived as advertisement factor.

More and more companies are manufacturing software outside the original software business. For instance, the cars today are running on plenty of software artifacts. Hence, we see a market for new code investigation techniques. Thus, the main goal is to push it more into the business world.

Secondly, Hypermodelling itself opens a lot of further questions. For instance, we need to consider how we express

[5] http://codesearch.hypermodelling.com
[6] http://www.youtube.com/watch?v=qeu8aPFDlP0

time and the movement of software artifacts through versions. Also, our application examples in diverse areas can be studied in more detail. There, the usefulness of various software cockpits or the integrated development environment can be studied empirically. In other areas like code recommendation and code search a more detailed investigation about similar techniques and synergies can be done. Additionally, the idea to plan future software versions with Data Warehouse technology can be interesting to control software development projects.

Currently, our technique is based on a relation source code model to create the multi-dimensional cubes. Further research needs to identify if this is the right way or if it would be better to load the associations directly in a multi-dimensional model. Likewise, the new high performance architectures for data analysis need to be considered as technology for source code analysis. They offer fast technology that offers even statistical engines on top of those for analysis.[7] They could be ideal candidates for future software analysis.

However, Hypermodelling targets currently explicit defined concerns. This means that we focused on concerns that are defined via specified modularization techniques or concerns that are associated with data of development tools. In addition to those, supplemental concerns in programs exist that can be revealed manually [11,12] or with mining algorithms [35]. We believe that our approach is suitable to support the concern retrieval tools as well as the automated concern mining. Therefore, we see the need for further investigations how these methods can be integrated into our approach to reveal potential synergies.

Third, we see a trail to investigate our excursus about psychology further. We depicted how categories are structured in psychological research. We see the Hypermodelling approach as possibility to investigate if concerns are structured in software like categories. In order to do so, we see the query possibility to reveal associations in source code as key factor to investigate if and how the structures of the mind manifest similarly in code. Furthermore, we also saw there that the perception is influenced by the context that we see. Therefore, we see Hypermodelling as a potential base to built new and contextual code investigation techniques.

All together, we see Hypermodelling as a beneficial step in software analysis. We presented a multi-dimensional technology that enables to slice and dice software from various viewpoints. Additionally, our approach is capable to integrate other viewpoints in the future. Software systems and their complexity are still growing. Therefore, we are see Hypermodelling as a contribution to advance today's software engineering research and practice.

[7]For instance SAP HANA http://www.sap.com/germany/plattform/in-memory-computing/in-memory-appliance/index.epx

REFERENCES

[1] A. Endres and D. Rombach. A Handbook of Software and Systems Engineering, Empirical Observations, Laws, and Theories, Person Education. Addison Wesley. 2003

[2] I. Sommerville, Software Engineering, Addison Wesley. 2010

[3] W. J. Brown, R. C. Malvea et al. Refactoring Software, Architectures, and Projects in Crisis. Addison Wesley. 1998

[4] M. Aksit and L. Bergmans, Obstacles in object-oriented software development . OOPSLA '92 conference proceedings on Object-oriented programming systems, languages, and applications. ACM. 1992

[5] D. L. Parnas. On the criteria to be used in decomposing systems into modules. Communications of the ACM, 15(12), Dec. 1972, pp. 1053–1058.

[6] P. Rechenberg, Informatik handbuch. Hanser verlag. 2006

[7] U. Grude. Java ist eine Sprache. Vieweg+Teubner Verlag.2005

[8] R. Merker. Programmieren lernen mit Java.2006

[9] E. W. Dijkstra. Selected Writings on Computing: A Personal Perspective. On the role of scientific thought. Springer. 1982

[10] W. L. Hursch and C. Lopes. Separation of Concerns. College of Computer Science, Northeastern University Boston, MA 02115, USA. 1995

[11] I. Majid and M. P. Robillard. NaCIN - An Eclipse Plug-In for Program Navigation-based Concern Inference. In Proceedings of the Eclipse Technology Exchange at OOPSLA.ACM. 2005

[12] M. P. Robillard, F. Weigand-Warr. ConcernMapper: simple view-based separation of scattered concerns. In Proceedings of the 2005 OOPSLA Workshop on Eclipse technology eXchange. ACM. 2005

[13] G. Kiczales, E. Hilsdale, et al..An Overview of AspectJ. European Conference on Object-Oriented Programming (ECOOP). pages 327-353.Springer. 2001

[14] H. Ossher, P. Tarr. Multi-dimensional separation of concerns and the hyperspace approach. In Proceedings of the Symposium on Software Architectures and Component Technology. Kluwer. 2001.

[15] W. Chung, W. Harrison, V. Kruskal et al.. Working with Implicit Concerns in the Concern Manipulation Environment,'05 Workshop on Linking Aspect Technology and Evolution (LATE).pages 1-5. ACM. 2005.

[16] R. E. Filman, T. Elrad, S. Clarke, M. Aksit Aspect-Oriented Software Development. Addison-Wesley Professional; 1st edition. 2004

[17] C. Atkinson, D. Stoll. Orthographic Modelling Environment. FASE'08/ETAPS'08 Proceedings of the Theory and practice of software,11th international conference on Fundamental approaches to software engineering. pages 93-96. Springer. 2008

[18] C. Atkinson, D. Brenner, P. Bostan et al. . Modeling Components and Component-Based Systems in KobrA. in A. Rausch, R. Reussner, R. Mirandola, F. Plasil (eds.): "The Common Component Modeling Example: Comparing Software Component Models Springer. 2008

[19] S.H. Kaisler. Software paradigms. John Wiley and Sons. 2005

[20] R. Johnson. Expert One-on-One J2EE Design and Development. Wrox. 2002

[21] W. H. Inmon. Building the Data Warehouse. 4th ed., J.Wiley & Sons, New York. USA. 2005.

[22] M. Golfarelli, D. Maio, S. Rizzi. The Dimensional Fact Model: A Conceptual Model For Data Warehouses. In International Journal of Cooperative Information Systems, 7, pages 215-247. 1998.

[23] S. Larndorfer, R. Ramler, C. Buchwiser. Experiences and Results from Establishing a Software Cockpit at BMD Systemhaus. 35th Euromicro Conference on Software Engineering and Advanced Applications, 2009 pp. 188-194

[24] M. Ciolkowski, J. Heidrich, F. Simon, M. Radicke. Empirical results from using custom-made software project control centers in industrial environments. ESEM '08 Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement. ACM 2008

[25] J. Heidrich, J. Münch, J. Wickenkamp. Practical Guidelines for

Introducing Software Cockpits in Industry. Proceedings of the 5th Software Measurement European Forum. 2009.pp49-64

[26] T. Frey. Vorschlag Hypermodelling: Data Warehousing für Quelltext. 23rd GI Workshop on Foundations of Databases. CEUR-WS. pages 55-60. Austria. 2011

[27] T. Frey, V. Köppen, G. Saake. Hypermodelling - Introducing Multi-dimensional Concern Reverse Engineering. In 2nd International ACM/GI Workshop on Digital Engineering (IWDE), Germany, 2011.

[28] T. Frey, Hypermodelling for Drag and Drop Concern Queries. Proceedings of Software Engineering 2010 (SE2012). Gesellschaft für Informatik (GI), Springer. Germany. 2012

[29] T. Frey, V. Köppen. Exploring Software Variance with Hypermodelling - An exemplary approach. In 5. Arbeitstagung Programmiersprachen (ATPS'12). im Rahmen der Software Engineering 2012. Gesellschaft für Informatik (GI) . Springer. Germany. 2012.

[30] J. A. Bloch. Metadata Facility for the Java Programming Language. 2004.

[31] M. Bruch, M. Mezini, M. Monperrus. Mining Subclassing Directives to Improve Framework Reuse, In Proceedings of the 7th IEEE Working Conference on Mining Software Repositories, IEEE, 2010.

[32] T. Frey, M. Gelhausen. Strawberries are nuts. CHASE '11 4th international workshop on Cooperative and human aspects of software engineering. ACM. 2011.

[33] T. Frey, M. Gelhausen, H. Sorgatz, V. Köppen. On the Role of Human Thought – Towards A Categorial Concern Comprehension. In Proceedings of the Workshop on Free Composition (FREECO) at the ACM Onward! and SPLASH Conferences. USA.2011.

[34] T. Frey, M. Gelhausen, G. Saake. Categorization of Concerns – A Categorical Program Comprehension Model. In Proceedings of the Workshop on Evaluation and Usability of Programming Languages and Tools (PLATEAU) at the ACM Onward! and SPLASH Conferences. USA. 2011

[35] P. F. Baldi, C. V. Lopes, E. J. Linstead, S. K. Bajracharya. A theory of aspects as latent topics. In OOPSLA '08 Proceedings of the 23rd conference on Object-oriented programming systems languages and applications. ACM. 2008

[36] T. Frey, V. Köppen. Hypermodelling Live - OLAP for Code Clone Recommendation. Baltic DB & IS 2012. Tenth International Baltic Conference on Databases and Information Systems July 8-11, Vilnius, Lithuania. CEUR-WS. 2012

A Novel Framework for Enhancing the Security of Enterprise Business Processes

Ahmed A. Hussein
University of Otto Magdeburg, Magdeburg, Germany
College of Computers and Information Sciences, King Saud University, Saudi Arabia
ahussein@ksu.edu.sa

Supervisor: Prof. Reiner R. Dumke
University of Otto Magdeburg, Magdeburg, Germany
dumke@ics.cs.uni-magdeburg.de

***Abstract*—Modern enterprises have their own business activities represented as business processes in an IT environment. The dynamic nature of these processes leads to a problem in providing their security and adaptation. To face this problem a security enhancement approach is proposed to guarantee obtaining secured BPEL based on defined enterprise security ontology criteria. The proposed approach introduces a framework that provides an enhanced securely adapted business processes at pre-production phase and checks for their security at production phase. The framework composes of two layers; the first layer is the validation and adaptation of the BPEL security which comes into place at the pre-production phase while the second layer is the Security Enhancement Engine (SEE) that is used at both phases. The behavior of the business processes security enhancing framework layers is modeled using Petri Net and the illustration of both layers' behavior by using the modular PNML is extended. Our framework is them applied to a real case study of a banking environment.**

Keywords: Busines processes; Security Ontology; Petri Net; BPEL; PMNL; QoS; WSDL

I. INTRODUCTION

Modern Software applications face major challenges to describe their behavior based on a mathematical foundation. Unified Markup Language (UML) as one of the choices for describing the behavior; through its diagrams such as state chart diagram; is considered a semi-formal technique, while Petri net provides graphical representation based on a mathematical foundation. The advantages of this involve verifying and validating system behavior.

The core representation of organizations is the business processes in which allows organizations to illustrate and implement their main services in an efficient secured and maintainable manner. The Business processes schema is represented by a BPEL (Business Process Execution Language) that contains detailed descriptions of services carrying out processes as web services. As a standard XML schema, BPEL allows for defining new tags such as a new security and internal relationships tags. Web services are referenced inside the BPEL and presented in a WSDL form; it is an xml standard representation that illustrates functions' attributes and behaviors.

Today's companies need to improve their organization activities by enhancing the security, efficiency and flexibility of their business process which are considered the core input to our BPSE framework [1]. The BPSE framework introduces new security implementation and verification criteria [2], and provides a security enhancement approach for business process.

Digital signature and the message authentication code are the major two methods to apply data integrity and authentication. In digital signature method, public key cryptography is used, while in the Message Authentication Code (MAC), a shared secret key is used instead of the private key [15].

Due to dynamic nature of business processes and to guarantee achieving both security and performance of them, Petri net techniques are used to illustrate the behavior of the whole framework, its Layer's components (Validation Engine, Security Ontology Engine, Generator Module, Hashing Module and Monitor Module), and to define the behavior structure between components using a modular Petri net Markup Language (PMNL) [10] which is an extension of the XML-based interchange format for Petri nets that consists of three parts, the PNML document, PNML types and features, and the PNML technology.

Recently, modern security technologies had been introduced to secure enterprise applications, however most of these current applications do not use such technologies in a fully utilized manner; this is clearly shown through the use of separate and not related standard security ontology criteria in different domains in the enterprise as shown in [13].

Securing business process depends mainly on securing the middle-ware existing in the organizational security. My contribution in this regard is not only to enhance the security of the middle-ware separately but also to integrate the security technologies used in other domains with it. New security tags, token and a developed hashing algorithm had been introduced that relate the infrastructure security technologies with the middle-ware in the enterprise.

The rest of this paper is organized as follows; the related work is given in section II. Section III describes the Business

Process Security Enhancement (BPSE) framework. The framework's layers are described in sections IV and V. Sections VI and VII illustrate the behavior of the framework and its model. In section VIII a case study is given. Finally the paper is concluded in Section IX.

II. WORK PLAN

Phase I:

Literature review on:

1- Service –oriented Architecture (SOA)
2- Service Oriented Security (SOS)
3- Model-Driven Architecture and UML
4- Security ontology criteria
5- Business Processes and Business Process Management Notation (BPMN)
6- Business Processes Security
7- Web services Security

Phase II: Sketching and building the proposed Business Process Security Enhancement (BPSE) framework based on a Defined Enterprise Security Ontology Criteria

Phase III: Proposing Security Enhancement Engine (SEE) as a separate layer for enhancing business process security

Phase IV: Creating the prototype that simulate the proposed framework model

Phase V: Verifying and validating the prototype through a real case study.

Phase VI: Measuring the performance of the framework to tune its functionalities if needed.

Phase VII: Documenting and publishing results in an international conferences and journals and finally document the overall result as my dissertation.

III. RELATED WORK

In [1] an enhancement of business process security criteria is introduced by defining an extensible layer added to the framework proposed in [2] that allows hashing the token values and performs an on-demand or periodic security checks of any alteration or modifications incidents that may occur in the Business Process Execution Language (BPEL) files and its corresponding Web Services Description Language (WSDL).

A new ontology that presents a new policy is proposed in [4] that integrates the workflow between business process and non-functional description. A connectional security model is presented in [5] to classify and determine an aggregation of security requirements across different services provided by organizational business processes. A service improvements is introduced in [6] through integrating organizational functional and non-functional business service registries.

The framework in [2] applies a new technique to validate the enterprise-based security ontology criteria imposed by the enterprise, and introduces a security tag to be inserted in both the BPEL and its corresponding WSDL files. This provides a correlation between the infrastructure components performance such as firewalls and business processes.

A process model-driven transformation approach to enhance security implementations such as XACML configurations was introduced in [8]. The concept of Petri net [9] is used in modeling the system behavior [10, 8].Translation of WS-BPEL process into stochastic petri nets is presented in [8] to develop a QoS- guaranteed software solution. A Petri Net markup Language (PNML) and its syntax and semantics, along with a modular extension as an independent module concept for petri net is introduced in [10].

IV. THE BPSE FRAMEWORK

This section introduces the layers of the BPSE framework and their functionalities. Figure (2) illustrates the layers' components. The first layer illustrated in section IV consists of two main engines, the validating and security ontology engines. The function of the first layer is to produce a securely adapted validated or a rejected BPEL and WSDL at the pre-production phase. The second layer illustrated in section V is the Security Enhancement Engine (SEE) which is composed of three components; the first one is the Generator Module (GM), the second one is the Monitoring Module (MM) and the third one is the Hashing Module (HM).

V. THE SECURITY ONTOLOGY AND VALIDATING ENGINES

The security ontology engine performs two actions, a normal action in which it generates an XML formatted policy script from the WSS, and a feedback action fired by the validation engine that verifies the token value (P) existed in the BPEL file against a proposed formula (1) of the calculated lower and upper values of the Enterprise-based infrastructure security (QoS) coefficients.

$$P = \sum_{m=1}^{n} Q_{tp(m)} \qquad (1)$$

Where $C_{2i-1} \leq Q_{tp(i)} \leq C_{2i} \; \forall \; i=\{1,2,3\}$

Table (I) shows an example of WSS policy script while table (II) shows the (QoS) coefficients.

TABLE I. WSS POLICY EXAMPLE

Item	Existence
Authentication	1
Message Integrity	0
Message Confidentiality	1

TABLE II. INFRASTRUCTURE SECURITY COMPONENTS (QOS) COEFFICIENTS

QoS coefficient	Belongs to	Interval	
		Lower	Upper
Q1	Firewall	C1	C2
Q2	Wireless Security	C3	C4
Q3	IDP	C5	C6

The lower and upper values (C2i-1& C2i) of the Enterprise-based infrastructure security(QoS) coefficients are grabbed by calculating the min. and the max. of the (QoS) throughput (Qtp) for a given time interval (Timeg) provided by the equations,

$$Q_{tp} = \frac{Reqs}{Ts} \qquad (2)$$

$$Time_g = n*T_s \quad n \geq 1 \qquad (3)$$

Where Reqs represents the total number of completed requests and the Ts represents the unit time, and

$$\text{Minimum}[Q_{tp(i)}] = C_{(2i-1)} \;\&\; \text{Maximum}[Q_{tp(i)}] = C_{(2i)} \quad \forall\, i=\{1,2,3\}$$

This is achieved by sending complete system requests (i.e. Login requests) within a unit time (Ts); for a given time interval (Timeg); through the three main appliances of the infrastructure security (firewall, wireless security and IDP) in the enterprise. Existence of the appliances is determined based on the values of the coefficients, for instance if there is no IDP then the values of C5 and C6 will be zeros.

The Validation engine; as the first layer's second component; performs three actions; the first one is to validate the WSDL file against the formatted XML policy script (see table3), the second action is to validate the BPEL file attributes by matching the token value presented in the newly introduced attribute "wsTokenValue" within the <plnk: partnerLinkType> node, against the value presented in the newly introduced attribute "bpelTokenValue" within the <partnerLink> node. The third action is verifying the token value by calling the feedback action of the security ontology engine. The validation engine; therefore; generates either a validated BPEL or a rejected BPEL with a consistency value that provides a valuable indication of why the file has been rejected or accepted; for instance; a BPEL file could be rejected because it has a wrong provided token value due to incorrect interval values of one or more of the (QoS) coefficients, in which indicates that those coefficients values were not calculated for this enterprise infrastructure or the infrastructure security component performance at the time of calculating the token value is changing and needs revision by the network personnel.

The Consistency value is composed of two parts {Part I. Part II}:

A. Part 1:
{WSS_AuthenticationValue,WSS_MessageIntegrity Value,W SS_MessageConfidentialityValue}

B. Part 2:
{ Qtp(1) _IndicatorValue, Qtp(2) _IndicatorValue, Qtp(3) _IndicatorValue}

The values contained in part one could be either 1 or 0, while the values contained in part two are shown in table III below

TABLE III. $Q_{TP(M)}$ VALUE INDICATORS (PART 2 OF CONSISTENCY VALUE)

#	Expression	$Q_{tp(m)}$ value Indicator	feedback
1	A<>0, B<>0,C<>0	Pv1,Pv2,Pv3	Token value (P) is correct
2	A=0, B=0,C=0	Pv4,Pv5,Pv6	Token Value(p) is incorrect due to incorrect coefficients values for firewall, Wireless Security and IDP appliances
3	A=0,B<>0,C<>0	Pv3	Q3 correct
4	A<>0,B=0,C<>0	Pv1	Q1 correct
5	A<>0,B<>0,C=0	Pv2	Q2 correct
6	A=0,B=0,C<>0	Pv5	Q2 Wrong➔ Wireless Security coefficient values are not within given intervals
7	A=0,B<>0,C=0	Pv4	Q1 wrong➔ Firewall coefficients' values are not within given intervals
8	A<>0,B=0,C=0	Pv6	Q3 Wrong➔ IDP coefficients' values are not within given intervals

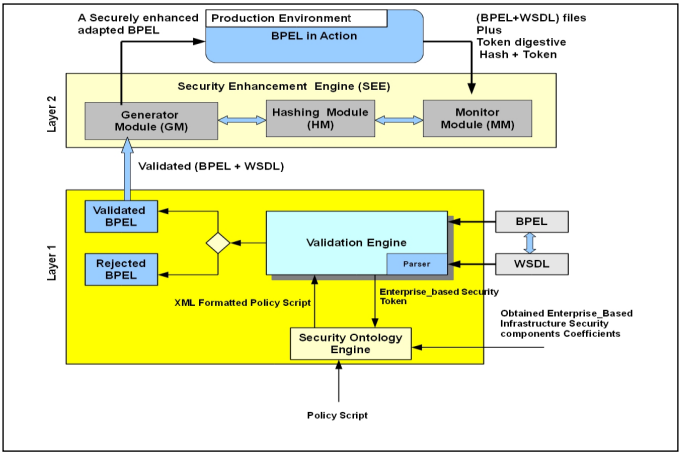


Figure 1. The BPSE Framwork Layers

VI. The Security Enhancement Engine(SEE)

The SEE is designed as a separate layer to provide fully customizable and expandable pool in which can contain any newly introduced security ontology criteria in business process. The input to the SEE is the output from the first layer, and the output of the SEE is a securely enhanced adapted BPEL and its corresponding WSDL files. At the pre-production phase the GM adds two digestive hash values (one is for the token value and one for the file contents) as new tags inside both the BPEL and its corresponding WSDL files as shown in <DigestiveTokenValue>, <DigestiveFileValue> attributes; this hash is generated by a developed fusion hash algorithm provided by the Hash Module (HM), and then the GM produces a securely enhanced adapted files. At the production phase, the Monitoring Module (MM) monitors and verifies that for a specific BPEL file; being called for periodic security checks and/or on-demand security checking; its content or its web services files' contents were not altered or modified. This is done by checking the token digestive hash values inside these files along with checking the digestive hash values for their contents. The Hashing Module produces the digestive hash for both the token values and files' contents, and provides it to both the GM and the MM at pre-production and production phases respectively. Hashing algorithm – Fusion Hashing (FH) – that is being developed is to be used by the Hashing Module (HM) using a combination of the Whirlpool hashing algorithm and the NMACA approach [12]. This will overcome the use of common non-secret key message authentication techniques such as MD5, SHA-1, and Whirlpool where an alteration of both token value, token's digest value and file contents won't be detected by the MM. Therefore; the Fusion Hashing (FH) algorithm detects any alteration either in the token value, digestive hash value, file contents and/or all.

VII. The Behavior of the Business Process Security Enhancement (BPSE) Framwork

Figure (1) illustrates the layers of the BPSE framework. The main components of the first layer are; the validating and security ontology engines. The behavior of the first layer takes place when required inputs are fed simultaneously to it. The inputs are as follow; the WSS policy which reflects the standard WSS (Web Service Security) the enterprise imposes; the interval values of the Enterprise-based infrastructure security coefficients obtained from the (QoS) of the enterprise infrastructure security components[1,2] and the enterprise BPEL file and its corresponding WSDL files that are needed to be validated.

The behavior includes an execution of the following internal functions for both engines; parsing the BPEL file and its corresponding WSDL files, validating the WSS standards presented in the WSDL files, and finally validating the value of enterprise-based token generated from Enterprise-based infrastructure security coefficients. The result of this layer is either a securely adapted validated or a rejected BPEL and its corresponding WSDL files.

The main components of the second layer (Security Enhancement Engine-SEE) [1] are the Generator Module (GM), the Monitoring Module (MM) and the Hashing Module (HM). The input to the first component (GM) is the output from the first layer. The behavior of this layer is divided into two parts, the first part is the execution of the GM and HM which takes place once the input from the first layer is fed to it, while the second part is the execution of MM and HM, this

takes place when a verification for alteration or modification of a specific BPEL file; being called for periodic security checks and/or on-demand security checking; and its content or its web services files' contents is requested.

The behavior of the first part involves the execution of the following functions, parsing the validated BPEL file and its corresponding WSDL by the GM to extract the token value and file contents size to be sent to the Hashing Module (HM). The HM in turn produces digestive hash values for them; using a developed fusion algorithm [1]; then it sends them back to the GM. Finally the GM inserts the digestive hash values into the BPEL file and its corresponding WSDL to produce a securely enhanced adapted BPEL and its corresponding WSDL. The behavior of the second part involves the execution of the following functions, Parsing the specific BPEL file and its corresponding WSDL files; being called for periodic security checks and/or on-demand security checking with respect to any modifications or alteration; to extract the token value, its hash value, file content's size and its hash value to be sent to the FH module.

The FH module in turn generates hashing values for the sent values of both token and file contents size and sends them to the MM module, finally the last function is performed by the MM module by checking the matching between both values' hashing presented in the files and the newly generated ones from the HM, if a match occurs then no modification or alteration has happened into the BPEL file or its corresponding WSDL otherwise, an indication of a security concern is raised.

VIII. Modeling the Behavior of the Framwork

In this section the behavior of the framework explained in section VI is modeled using Petri Net and modular PMNL. The following definitions are used to describe our Petri Net model.

Definition 1: the Petri Net structure for the BPSE framework is a tuple N = {P, T, F}, where: P is a set of places (components), T is a set of transitions, and F is a set of flows such that $F \subset (P \times T) \cup (T \times P)$.

Definition 2: given N= {P,T,F} as a petri net of BPSE framework such that BPSE consists of different interconnected layers that is formed of different modules. A petri net module U = {β, λ, α} Where $\beta \subseteq P$, $\lambda \subseteq T$, and $\alpha \subseteq F$.

Definition 3: given N= {P,T,F}, a set of input(I) and output (O) transitions are defined as I={i | i = (p,t) ϵ F } and O={o | o=(t,p) ϵ F} respectively.

Definition 4: given the marked petri net of BPSE framework as a tuple of N= {P, T, F, M_0, M_f}, where M_0 and M_f are the initial and final marking of any petri net graph, respectively.

Using these definitions the petri net of our BPSE framework is given in figure (2), where petri net modules U1, U2 and U3 represent layer1, layer2-GM&HM and layer2-MM&HM respectively.

Definition 5: given a Petri net N and a marking M, a Reachability Tree RT (N) illustrates the changes in all M's in N.

As an example, figure (3) illustrates the reachability tree of Module U2: RT (U2).

Tables IV, V and VI shows the Modeling of the petri net modules U1, U2 and U3 behaviors using modular PMNL, while table VII shows the interchangeable usage schema between U1 and U2, where an instance of U1 (P4) is in U2 and this means U2 uses U1.

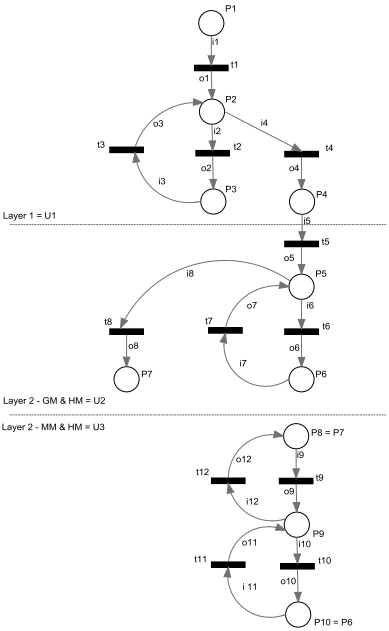


Figure 2. The BPSE Petri Net Model

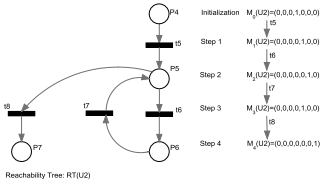


Figure 3. The Reachability Tree of Petri Net Module2: RT(U2)

TABLE IV. MODULAR PMNL REPRESENTATION OF U1 BEHAVIOR

```
<module name="U1"> Table1
<interface>
<exportPlace id="p1" ref="p1"/>
<exporttPlace id="p4" ref="p4"/>
</interface>
<transition id="t1"/>
<transition id="t2"/>
<transition id="t3"/>
<transition id="t4"/>
<arc source="p1" target="t1"/>
<arc source="t1" target="p2"/>
<arc source="p2" target="t2"/>
<arc source="t2" target="p3"/>
<arc source="p3" target="t3"/>
<arc source="t3" target="p2"/>
<arc source="p2" target="t4"/>
<arc source="t4" target="p4"/>
</module>
```

TABLE V. MODULAR PMNL REPRESENTATION OF U2 BEHAVIOR

```
<module name="U2">
<interface>
<exportPlace id="p4" ref="p4"/>
<exportPlace id="p7" ref="p7"/>
</interface>
<transition id="t5"/>
<transition id="t6"/>
<transition id="t7"/>
<transition id="t8"/>
<arc source="p4" target="t5"/>
<arc source="t5" target="p5"/>
<arc source="p5" target="t6"/>
<arc source="t6" target="p6"/>
<arc source="p6" target="t7"/>
<arc source="t7" target="p5"/>
<arc source="p5" target="t8"/>
<arc source="t8" target="p7"/>
</module>
```

TABLE VI. MODULAR PMNL REPRESENTATION OF U3 BEHAVIOR

```
<module name="U3">
<interface>
<exportPlace id="p8" ref="p7"/>
<exportPlace id="p8" ref="p7"/>
</interface>
<transition id="t9"/>
<transition id="t10"/>
<transition id="t11"/>
<transition id="t12"/>
<arc source="p8" target="t9"/>
<arc source="t9" target="p9"/>
<arc source="p9" target="t10"/>
<arc source="t10" target="p10"/>
<arc source="p10" target="t11"/>
<arc source="t11" target="p9"/>
<arc source="p9" target="t12"/>
<arc source="t12" target="p8"/></module>
```

TABLE VII. INTERCHANGEABLE USAGE SCHEMA BETWEEN U1&U2

```
<instance id="n1">
<exportPlace parameter="p4" instance="u1" ref="p4"/>
</instance><instance id="n3"><exportPlace
parameter="p4" instance="u2" ref="p4"/></instance>
```

IX. CASE STUDY

The framework is implemented in a banking environment that has its own automated business processes in BPEL format generated by oracle BPA Suite; its network has infrastructure security components which include a SRX 3600 Juniper firewall, a Wireless security component for Wireless access points Cisco1520, and a juniper IDP250 Intrusion Detection and Prevention Appliance, the basic configuration of the server that was used is 2xQuad Core 3.2 GHZ with 64 GB RAM running windows 2008 R2 with SP2. The implementation has been performed as follow:

At Pre-production Phase:

1. Obtaining the bank security ontology criteria, Policy Script.
2. Obtaining the bank BPEL file that is needed to be validated.
3. Obtaining the values for the (QoS) throughput coefficients, that have been calculated using JMeter (a tool that can perform N requests per second and measures the percentage of completion) to send 40 requests per second for a unit time (Ts) of 900 seconds and for a given time interval (Timeg) of 14400 seconds. This step has been applied for the other infrastructure components existing in the banking environment (wireless security and IDP). The actual intervals calculated for all the appliances were:

$$1996 \leq Q_1 \leq 2342$$
$$1807 \leq Q_2 \leq 2113$$
$$1363 \leq Q3 \leq 1765$$

4. Apply correct WSS security policy for the enterprise to this file, adding different correct token values for BPEL and WSDL.
5. The above inputs are then fed into layer 1 of the BPSE framework to be processed and obtain a validated BPEL files each with validated tokens and its corresponding validated WSDL files' sizes.
6. The validated BPEL and WSDL were fed to the GM at layer 2 to produce securely enhanced BPEL with digestive hash value for each token and file contents using proposed FH algorithm.
7. The time consumed by whirlpool and FH algorithm to generate the digestive hash is measured. The time complexity of the FH algorithm at this phase is O (n) where n is the file size.
8. In the preparation step of the FH algorithm the input is padded to have blocks of 512bits and since the

maximum expected token length is within the range of 0-256 bits, then the padded input will always have one block of 512 bits. As a result of that the time consumed by the FH algorithm to generate the digestive hash for tokens is fixed and was measured as ~0.002 microsecond.

9. For BPEL and its corresponding WSDL files, table (VIII) and Figure (4) illustrates the time in T*10000 unit (T=10 microsecond); consumed by whirlpool & FH; and different validated file sizes in KB {200-10000}. Figure (4) shows the comparison in time between applying whirlpool to create the files' contents digestive values and applying FH algorithm to create the same values. The result shows there is a tiny difference in time, and suggests that using our developed FH algorithm provides enhanced security without a noticeable degrading in performance.

TABLE VIII.
TIME VALUES FOR APPLYING WHIRLPOOL AND FH TO AN EXPERIMENTAL FILES' SIZES
(PRE-PRODUCTION PHASE)

File Size(KB)	Whirlpool algorithm	Fusion algorithm
200	6.467	22.4501
400	7.769	32.0727
600	10.565	42.7509
800	15.025	51.99
1000	18.531	63.2324

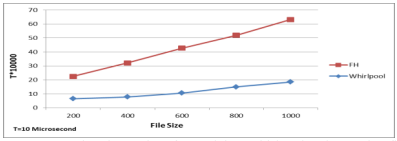


Figure 4: Comparing time values for applying Whirlpool and FH using files' sizes

At Production Phase:

1. A periodic security check and/or on-demand security check is called on the securely enhanced BPEL
2. The securely enhanced BPEL and its corresponding WSDL files are fed into the MM at layer 2.
3. The Token values, their digestive hash values, and files' contents digest values are extracted from the WSDL files by the MM.
4. The Token values and their digestive hash values are extracted from the WSDL files by the MM.
5. The MM calls the HM.
6. The FH algorithm is used by the HM to generate a corresponding digestive hash values for alteration checking of passed files and tokens. The time complexity of the FH algorithm at this phase is O (n) where n is the file size.
7. In the preparation step of the FH algorithm the input is padded to have blocks of 512bits and since the maximum expected token length is within the range of 0-256 bits, then the padded input will always have one block of 512 bits. As a result of that the time consumed by the FH algorithm to generate the digestive hash for tokens is fixed and was measured as ~0.002 microsecond.
8. For BPEL and its corresponding WSDL files, table (IX) and Figure (5) illustrates the time in T*10000 unit (T=10 microsecond); consumed by GM and MM; and different file sizes in KB {200-1000}. Figure (5) shows the comparison in time between the FH used separately by the GM at pre-production phase and using FH in conjunction with a digestive hash comparison steps (CT) at production phase. The result shows there is a tiny difference in time, and suggests that using our developed FH algorithm provides enhanced security without a noticeable degrading in performance.

TABLE IX.
TIME VALUES FOR USING FH AT PRE-PRODUCTION AND PRODUCTION PHASES RESPECTIVELY

File Size(KB)	Fusion algorithm	Fusion algorithm + CT
200	22.4501	22.4551
400	32.0727	32.0777
600	42.7509	42.7559
800	51.99	51.995
1000	63.2324	63.2374

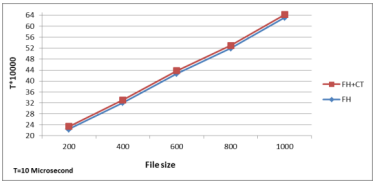


FIGURE 5: PERFORMANCE OF FH AT PRE-PRODUCTION AND PRODUCTION PHASES.

X. CONCLUSION AND FUTURE WORK

In this paper, a novel framework was introduced, it is composed of two layers, the first layer is to validate a BPEL and its corresponding WSDL files against the enterprise security ontology criteria, and produce a validated BPEL. The second layer at the pre-production phase enhances the securely adapted BPEL by adding a digestive hash using the Generator Module. The behavior of the business processes security enhancing framework layers has been modeled using Petri Net and extended our model by using the modular PMNL. The schematic diagram that describes the behavior in modules is defined. The framework through a real case study is implemented in a banking environment. Currently and as a part of a future work phases VI and VII are being carried out in

which a formal model of the Business Process Security Enhancement framework will be defined where new parameters that measure its performance are analyzed and improvements could be devised.

MY PUBLICATIONS

[1] Ahmed A. Hussein, "An Enhanced Security Approach for Securing and Validating Enterprise Business Process," Softwaew Measurement News,Journal of the software metrics Community, vol. 17, Number 1,pp. 17–29, March 2012.

[2] Ahmed A. Hussein, Ahmed Ghoneim, and Reiner R. Dumke,"An Approach for Securing and Validating Business Processes Based on a Defined Enterprise Security Ontology Criteria", (CCIS) Series of Springer LNCS), Volume 189, pp. 54-66 , 2011.

[3] Ahmed A. Hussein, Ahmed Ghonein,Samir Mougy, Reiner Dumke,"A Service Oriented Integration and Adaptation Framework for Organizational Services Orchestration", WORLDCOMP/SWWS'10, DBLP SWWS 2010: 105-111

REFERENCES

[4] Dong Huang, "Semantic Descriptions of Web Services Security Constraints", In SOSE '06: Proceedings of the Second IEEE International Symposium on Service-Oriented System Engineering, Pp. 81 – 84, 2006.

[5] Michael Menzel; Christian Wolter; Christoph Meinel, "Towards the Aggregation of Security Requirements in Cross-Organisational Service Compositions", LNCS, pp. 297 – 308, 2008.

[6] Frédérique Biennier, Régis Aubry, Mathieu Maranzana, "Integration of Business and Industrial Knowledge on Services to Set Trusted Business Communities of Organisations", PRO-VE, pp. 420-426, 2010.

[7] Y. Christian Wolter, Michael Menzel, AndreasSchaad, Philip Miseldine, Christoph Meinel, "Model-driven business process security requirement specification,Journal of Systems Architecture",Secure Service-Oriented Architectures (Special Issue on Secure SOA), pp. 211-223, ISSN 1383-7621, April 2009.

[8] Dario Bruneo, Salvatore Distefano, Francesco Longo, Marco Scarpa, "QoS Assesment of WS-BPEL Processes through non-Makrovian Stochastic Petri Nets", Proceeding of 2010 IEEE international symposium on Parallel & Distributed Processing (IPDPS), Atlanta, USA, ISBN: 978-1-4244-6442-5, April 19-23.

[9] De Backer, Manu Snoeck, Monique, "Deterministic Petri net languages as business process specification language", K.U.Leuven - Departement toegepaste economische wetenschappen, DTEW Research Report 0577, pp.1-21, 2005.

[10] Michael Weber and Ekkart Kindler, "The Petri Net Markup Language", Petri Net Technology for Communication-Based Systems – Advances in Petri Nets, LNCS volume 2472, pp. 124-144, 2003.

[11] Marc Stevens, Arjen Lenstra, and Benne de Weger, "Chosen-prefix collisions for MD5 and colliding X.509 certificates for different identities", EUROCRYPT 2007 (Moni Naor, ed.), LNCS, vol.4515, Springer, 2007, pp. 1–22.

[12] Khaled S. Alghathbar and Alaaeldin M. Hafez, "The Use of NMACA Approach in Building a Secure Message Authentication Code", , International Journal of Education And Information Technologies Issue 3, Volume 3, 2009

[13] Robert Neumann, Konstantina Georgieva, Ayman Massoud, Anja Fiegler, Florian Muhß, Detlef Günther, Ahmed Hussein, Reiner R. Dumke,"Quality Based Digital Eco-System", Internal Research Report, SML@b, University of Magdeburg, Germany, September 2011

Design and Analysis of Visualization and Browsing Methods for Spatial Neuroanatomical Atlases

Anja Kuß
Capgemini Deutschland GmbH
Potsdamer Platz 5
10785 Berlin
Email: Anja.Kuss@gmx.de

Bernhard Preim
University of Magdeburg
Universittsplatz 2
D-39106 Magdeburg
Email: preim@isg.cs.uni-magdeburg.de

***Abstract*—In biomedical informatics, digital surface-based neuroanatomical atlases serve as reference frames for relating data from different experiments and data modalities. They support neurobiologists in visualizing and analyzing their data. In this paper, we present results and open tasks of our work which addresses visualization and interaction approaches to improve the atlases' use in education, data exploration, and result presentation. A first step in this work was to define initial representative user stories. From the stories, we found that visual atlas browsing and generating meaningful atlas visualizations are time-consuming tasks requiring anatomical and application expert knowledge. To address these problems, we developed a visualization approach that generates intuitive visualizations for high-level user queries. This method is based on semantic information stored in an ontology. We further evaluated the effectiveness of techniques that emphasize the spatial relationships between filamentous structures and neighboring or surrounding volumetric objects. To do so, we performed an initial qualitative user study which was refined by a subsequent quantitative study. Presently, we review and refine our initial user stories and convert them into detailed use cases. We further analyze the results of a second run of our qualitative user study.**

I. Introduction

Digital surface-based brain atlases are reference frames in which data from different experiments and data modalities can be related. They facilitate visualization of biomedical data and form a basis for investigating the relation between morphology, function, and genetic factors.

Neuroanatomical atlases are addressed in several communities and research projects. For example, [1] constructed an atlas for the human brain, [2] developed an atlas for the mouse's nervous system and a honeybee brain atlas was built by [3]. The community is constantly growing and so is the amount of collected data.

The early tasks in this field of digital neuroanatomy were the atlases' generation and the automatic integration of additional data [4], [5]. Still, a lot of effort is put into continuous data acquisition [6]. For a while now, some focus is shifting to browsing, analyzing and visualization the collected atlas data [6]. However, the currently available tools and research results cover quite general use case and require technical or biological expert knowledge.

The work presented in this paper is based on the project *Digital Neuroanatomy, Data Visualization, and Modeling*[7]. Here, we focus on scenarios, visualization and interaction approaches to improve the atlases' use in education, data exploration, and result presentation. We aim at defining specific use cases as well as reducing the needed expert knowledge and selecting and evaluating adequate visualizations methods. The used data contains surface reconstructions ranging from coarse brain structures (neuropil) down to nerve cells (neurons).

The paper's structure is as follows: First, we introduce our aims and a brief research plan in sections II and III. These aims are supported by the related work presented in section IV. Further, we present the so far completed work and the remaining planning in sections V and VI.

II. Specific Aims

The overall goal of this work is to facilitate the access to surface-based anatomical atlases. This section presents the specific aims which, in our opinion, will support this overall goal. Before we define those specific aims, we back-up them with general user stories.

Abstract User Stories

We outline two abstract user stories which we observed regularly during the project's [7] first stage. These user stories are, for example, typical for the visual exploration of the honeybee brain atlas.

U_1 In the first user story, a student wants to learn about the brain anatomy represented in the atlas. She wants to understand the hierarchical organization of the brain structures and their substructures. In more detail, she wants to learn about different neuron types, in which neuropil certain neuron types can be found, which neuropil are connected by which neurons, and which neurons meet in a certain neuropil.

U_2 In the second user story, a neurobiologist wants to visually present or compare data. For example, a neuron has been acquired using a new imaging technique. The neurobiologist wants to visually compare it to neurons of the same type but acquired with another imaging technique. She probably also wants to know in which neuropils the new neuron is located, Finally, she wants to create a visualization to show her results to a colleague or cooperation partner.

Specific Aims

Based on the abstract user stories, we define the following specific aims:

- A_1 To develop a visualization technique that combines semantic atlas information and the concept of focus+context visualization to create meaningful visualizations.
- A_2 To analyze the effectiveness of selected visualization techniques that emphasize the spatial relationships between filamentous structures, such as nerves, and neighboring or surrounding volumetric objects, such as coarse brain structures.
- A_3 To analyze, describe and refine the selected user stories U_1 and U_2 into documented use cases.

Our work is successful if a user can create and browse meaningful visualizations of neurobiological atlas data with *less expert knowledge but without increasing the effort*. For us, a meaningful visualization contains all *structures of interest*, additionally presents *context* and uses an *appropriate technique* for the underlying data. We think that such a visualization is *intuitive* and can be *interpreted easily*. We further think that a good visualization creation process pays attention to user interface usability issues. A plan which outlines how we intend to achieve these aims is presented in the following section.

III. Plan of Research

The first step is to review the work that is especially related to the above formalized aims. A summary of related work is presented in section IV. As another precondition, we collect representative data from our cooperation partners and select a visualization framework to be able to test our approaches.

In the next step, we analyze how semantic information could be presented in an ontology to support browsing and visualizing of atlas data with less expert knowledge. An appropriate knowledge representation for our neuroanatomical atlas data should contain a hierarchical description of structures and the relations between them. A visualization framework should be able to read and process such a representation. Based on our analysis results, we develop a prototypic ontology for selected data from our research project.

Afterwards, we design an approach that creates understandable focus+context visualization using the before mentioned semantic information scheme. The approach should define the structures to be visualized and apply an appropriate visualization technique. Its user input should be intuitive and preferably small. We experimentally implement the approach within our selected framework using our prototypic ontology.

We then evaluate selected visualization techniques which are applied to generate results in our semantic based visualization approach. We want to show that our preferred techniques simplify the understanding of specific relations between structures. Different visualization evaluation methods will be pursued to compare different set-ups and application fields. Here, an initial requirement would be to find a representative number of participants who will take part in the evaluation studies.

Finally, we review the prototypic implementation within a module based visualization framework in terms of usability. We do this by refining the previously introduced user stories U_1 and U_2. The focus will be on components for atlas querying and visualization where aims A_1 and A_2 form examples. Specific, detailed and documented use cases are the intended result.

IV. Related Work

Today, a variety of digital atlases for several species and imaging modalities have been generated. Currently, atlases are available, for example, for the human brain [1] and the nervous system of the mouse which offers access to gene expression data [8], [2], [9], [10]. Invertebrate brain atlases like the brain atlases of the fruit fly brain [11], [12], the moth brain [13], and the honeybee brain [3], [14] also emerged.

In the beginning, the most challenging task was the generation of digital atlases. Pipelines comprising preferably automatic methods for image preprocessing, segmentation, registration, and image storage were developed [4], [5]. Even though the atlas creation methods still require improvements, now, the focus is on the actual usage of the continuously growing atlases. The users want to browse, analyze, query, describe, share, and visualize their collected data [6], [15]. Associated with these wishes are several requirements: *application and use case dependent approaches*, intelligent data management, data enrichment with *semantically meaningful information* that can be searched, intuitive *querying tools*, and methods for the efficient *creation of meaningful visualizations*, but also the *evaluation* of available methods and tools.

A. Use Case Design

To support atlas exploration, several tools have been developed or are current work in progress. Software is available which offers general database, data annotation, and image analysis functionality such as the Open Microscopy Environment (OME) [16]. Exemplary guidelines on how to design anatomy information systems are presented in [17].

Others concentrate on the specific data contained in the atlases they are working with. For example, The Mouse Brain Image Visualizer (MBIV) [18] and the Allen Brain Atlas [19] provide techniques for browsing and analyzing 3d data of the mouse nervous system. These tools enable slicing as well as spatial querying and viewing of annotated image data.

A big step forward in interactive atlas querying has been made in the *BrainGazer* system of [20]. The system provides browsing tools which enable visual interaction and querying of semantic based data and spatial relationships in neurobiological volume data.

Relevance and Significance: Despite all these contributions, from the usability and visualization point of view, atlas browsing is still in its infancy. Especially, conceptual user interface design as, for example, described in [21] plays a minor role. As a result, the available tools tend to support many different and quite general use cases instead of supporting a small amount of detailed ones.

Other research fields, such as microbiology [22] or medical visualization [23], present approaches in which the frame-

work's focus is on specific use cases or scenarios. [23] developed systems for medical training and planning based on the concept of *scenario based design* which was described by [24]. Applying this concept to digital atlas browsing could improve the atlas frameworks' usability, applicability and further their acceptance. We will use the design approach for the user scenarios presented in section II.

B. Ontologies to Enrich Digital Atlases with Semantics

One way to address the problem of data management and data enrichment semantics is the application of ontologies. Ontologies formalize knowledge and can be used to describe biomedical structures and relations among them. Providing regimentations of terminology, they can support reusability and integration of data [25].

The interest in the application of ontologies is rising and a high number has already been built and published [26]. Guidelines for the creation of neuroanatomical ontologies were discussed by [27]. Approaches for modeling detailed information about functional and structural neuroanatomical parts have been presented in [28], [29].

Relevance and Significance: So far, ontology development in neurobiology concentrates on representing complete sets of structures and comprehensive terminologies. Of minor interest has been the ontologies' relation to atlas data and its use for the exploration of atlas data. It is unclear how ontologies should be designed to support the browsing of digital atlases and how they could be used, for example, for the creation of visualizations.

C. Semantics-based Visual Querying of Atlas Data

First approaches integrate ontologies into atlas browsing frameworks. The EMAP JAtlasViewer [15] and The Smart Atlas (BIRN) [9] offer hierarchical trees of structures and substructures from ontologies. These structures can be selected and their related data can be displayed in image viewers. A more sophisticated approach is provided by the BrainGraph tool, a part of the Mouse BIRN Atlasing Toolkit [30], which displays the structures as nodes in a graph. Opening individual nodes accesses more information, such as functional connectivity and anatomical relations. Obviously, techniques for the visualization of trees and graphs are required and, thus, some approaches have already been proposed [15], [31].

Relevance and Significance: Often, much knowledge about the atlas browsing system or the contained data is required to successfully navigate through the atlas. In some applications special query languages have to be used. Ontology visualization and especially visual ontology interaction are limited or missing. An approach which successfully integrates browsing and visualization approaches could reduce the currently existing drawbacks of digital atlas exploration.

D. Visualization of Structures in Surface-based Brain Atlases

In neurobiological applications, such as anatomical atlases, visualizations of filamentous structures together with neighboring structures are often used. Examples are presented in the use cases in Section II.

Focus+Context Visualization: We consider the problem of adequately presenting filaments together with neighboring structures as a focus+context visualization problem. Focus+context visualizations distinguish between objects of interest and their surroundings. To create such visualizations, a classification of objects into focus and context is required. This discrimination can either be binary or smooth [32]. While the selection of one specific focus object in atlas-based applications is performed by the user (e.g. by selecting a structure in a list of objects or by directly picking in the visualization), the determination of other important entities has to be implemented by the system.

Methods to interactively adapt focus+context visualizations of volume data to user input have been presented in [33], [34]. The importance values used to weight object visibility and rendering style are assumed to be given or set to "high" for the selected object and "low" for all others. Assignment of importance values that do not depend on the spatial relations but on semantic information have been described as cue methods by Kosara [35].

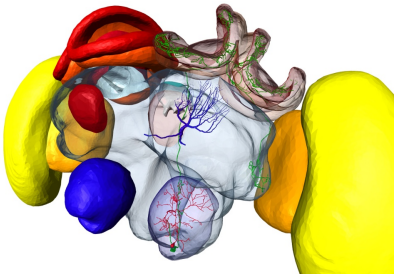


Fig. 1. Example for manually created visualization of anatomical atlas data of the honeybee brain. Most coarse structures (neuropils) are presented opaque. Transparency is used to expose finer structures (neurons)

Visualization of Filament-Surface Relationships: Once the distinction between focus objects and context objects has been completed, suitable visualization techniques should be applied. Especially, methods which realistically present filaments and methods which solve the occlusion problem of filaments and their context structures are required. In general, the methods for filament visualization try to find a compromise between speed and quality [36].

Recently, more and more knowledge from perception theory is exploited to improve the visualizations. The most obvious and also the fastest way to visualize filamentous structures is the rendering of simple polylines. To improve the perception of depth and spatial relations polylines can be illuminated [37]. Better suited for a realistic visualization are line rendering using tubes or convolution surfaces [38], approximation of tubes using textured triangle strips or rectangular primitives [39], [40]. [41] assign different hues to line elements and use halos for a better distinction of adjacent and overlapping lines.

Presenting filament visualizations together with their surrounding or neighbouring context structures depends to a very large extent on solving the occlusion problem. Techniques for occlusion reduction are, for example, cut-away views, ghosted views, and exploded views [42]. Unfortunately, these techniques tend to remove parts of the data, resulting in visualizations that increase the viewer's mental load. Using transparency, all relevant structures could be displayed together with unmodified shape. An example visualization is shown in figure 1. However, understanding the shape and location of a transparent surface is difficult, because ordinary shape and depth cues of shading and occlusion are less pronounced [43]. To overcome these limitations, [43], [44] display ridge and valley lines and further textures on transparent surfaces. [45] propose a view-dependent transparency rendering method based on rules defined in books for technical or scientific illustrations.

Relevance and Significance: Simplified generation of meaningful visualizations is hardly contained in current atlas exploration systems. The visualization generation process should become easier and less time consuming. To optimize the resulting visualizations, enhanced techniques for the display of tube-like rendered filaments together with transparently rendered context structures are required. These visualization techniques could improve the presentation and communication of research results. They could further support the understanding of spatial relationships between structures in neuroanatomy.

E. User Studies to Evaluate Visualizations and Frameworks

User studies that evaluate the perceptual quality of visualization methods or the usability of an application are becoming an important part of the visualization method and application development process. We can distinguish between *qualitative* and *quantitative* studies.

Qualitative User Studies: Qualitative studies explore subjective parameters, such as personal preferences or taste [46], [21]. They, for example, use questionnaires or interviews to ask participants if they prefer one visualization method or application design to another.

A good example for a qualitative study of visualization methods is presented in [47] where medical doctors and laymen were asked to compare and evaluate medical illustrations based on hybrid visualization techniques. Compared to visualization evaluation, qualitative studies of user interfaces have an older history and are more complex. Questionnaires for interface evaluation vary between the interface's development stages and its application field. In [48] several questionnaire types for interface usability testing were compared. The researchers found that simplest questionnaires yielded among the most reliable results.

Quantitative User Studies: In contrast, participants of quantitative studies have to perform specific tasks under varying conditions [46], [21]. The number of correct responses for a task or the capturing of the users' pattern of activity can shed light on a method's advantage or an interface's usability.

An example for a visualization evaluation experiment is presented in [49], who evaluated four emphasis techniques for structures in medical visualizations. They further describe guidelines for experimental user studies. The value that 3D visualizations convey to students in reaching their anatomical learning objectives was analyzed by [50]. The authors suggest a potentially beneficial effect on learning from 3D visualizations. Computer-assisted and text-based learning were compared in the study of [51]. The results provide evidence that computer-assisted instruction can be superior to traditional textbook learning. The effect of explicit guidance for exploring spatial relations was evaluated in [52]. Their results indicate that a virtual jigsaw of anatomical structures improves the understanding of spatial relations from 3D illustrations.

Relevance and Significance: To our knowledge, neither the perception of filament trajectories through surrounding objects nor the usability of tools supporting digital atlas browsing have been analyzed in user studies.

V. Finished Work

Major parts of aim A_1 and A_2 have already been completed. They have been published in several articles or abstracts. This section summarizes their major contribution.

A. Ontology-Based Visualization of Hierarchical Neuroanatomical Structures

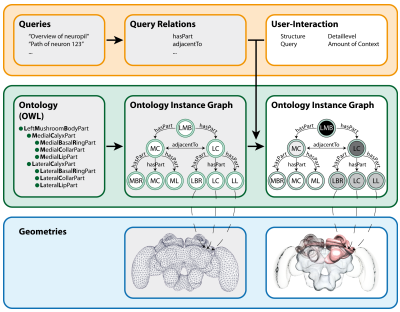


Fig. 2. Work flow of our ontology-based visualization approach. First an ontology graph storing semantic information is created. Afterwards high-level queries are defined on the relations of the ontology. The user can ask for a visualization of a structure. This query is evaluated and the vertices of the ontology graph receive query-dependent importance values. Geometries are linked to the graph vertices and are visualized using importance-dependent parameters. Text boxes having a white background show processes during run-time. The example geometries show parts of the bee brain.

We developed an approach to generate query-dependent visualizations of surface-based anatomy atlas data. In this approach, expert knowledge is formalized in an ontology and on predefined queries. Using the formalized knowledge, we combine a spreading activation approach for computing focus+context structures with a specific level-of-detail strategy

for hierarchical structures. The resulting visualization contains the selected structure and appropriate context structures. Because our reconstructed data is linked to instances of the ontology, the visualization is created almost fully automatic.

The core steps of the method are:

1) An expert develops an ontology with a specific structure, suitable for deriving visualizations. This ontology is linked to the available geometries.
2) An expert defines high-level visualization queries that specify a set of relevant relations.
3) A user selects a focus object and a visualization query.
4) A graph algorithm generates query-dependent importance values for each structure.
5) These importance values are mapped to visualization parameters such as transparency.
6) The user can control the level of detail of the visualization.

Figure 2 illustrates the work flow. Steps one and two form preprocessing steps in which the semantic information is created. The interactive generation of visualizations represented by step three to six takes place at run-time.

To test the visualization technique, we exemplary developed a first ontology outline that describes the hierarchical organization of the honeybee brain's structures. Its design enables the users to relate their reconstructed data to the ontology's semantic information. Our ontology fulfills the following requirements: It is restricted to a specific domain, the hierarchical representation of the honeybee brain's anatomical structures including basic neuron types and basic spatial relations. The ontology also handles structures at different scales and data incompleteness. This work has been published in [53], [54]. A visualization result is shown in figure 3.

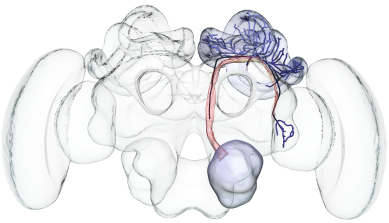


Fig. 3. Example ontology-based visualization of a honeybee brain's nerve. The nerve's input and output regions and the tract are emphasized. A lower transparency is used for the context structures.

B. Effective Techniques to Visualize Filament-Surface Relationships

We developed visualization techniques that accentuate intersections and trajectories of tubes representing nerves through transparent structures representing neuropils. We implemented explicit visualization of intersection points, tube segment coloring according to surrounding objects, color modulation according to depth, and depth-based contrast enhancement [55], [56]. The techniques were evaluated in two subsequent user studies. Both studies are described in more detail in [57].

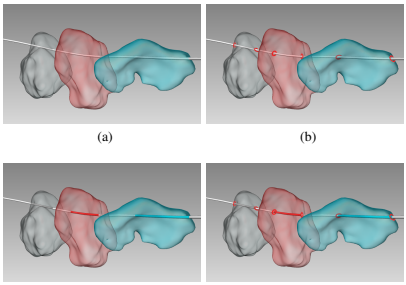


Fig. 4. Example test scenes from the confirmatory study. 4a) Baseline visualization. 4b) Intersection points of tube and volumetric structures are displayed using rings. 4c) Tube coloring by surrounding material. 4d) Combined intersection display and coloring by material.

1) Exploratory User Study: In a first exploratory unsupervised user study, we asked participants to rate and compare visualizations that used the above mentioned accentuation techniques. To do so, the participants filled out a printed questionnaire which consisted of 13 tasks. Each task was composed of several questions regarding one or multiple static visualizations of a bee brain models's substructure. A questions underlying the ratings or comparisons was, for example: *How well can you identify if lines lie inside, outside or behind an object?* The study's result met most of our expectations. All evaluated methods were thought to reduce the perception problems described above. The most preferred techniques were the coloring of lines and the marking of line-surface intersections by glyphs.

2) Confirmatory User Study: We designed a second study to recheck and strengthen our findings. The study concentrated on the techniques that had been preferred in the exploratory study, namely material coloring, intersection glyphs, and the combination of both. We limited the task to identifying a line's trajectory. This reduced the complexity of the study while concentrating on a commonly performed task in neuroscience applications. We used static scenes, because we were mainly interested in the influence of the emphasizing techniques on perception. Using static scenes should also allow us to infer how suitable the techniques are for 2D media, like print or web. 48 participants tested a standard visualization (see figure 4a) and visualizations that resulted from adding either one or both of the emphasizing techniques (see figures 4b, 4c and 4d). In the scenes, the participants had to specify whether a filament runs through a certain structure or not.

Our study results show that the visualization techniques support the user in determining whether a filament runs through

a transparent structure. Both accuracy and speed are improved by the techniques. The best performance is achieved for the combination of coloring and glyphs. In figure 5, the study's average number of correct answers and average response times together with their 95% confidence intervals are displayed for all evaluated visualization techniques.

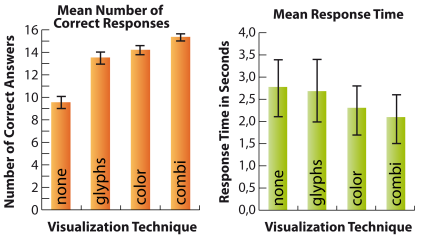


Fig. 5. Quantitative user study's results showing the number of correct answers and response times for all participants. 5a) Average number of correct responses with 95% confidence intervals for each tested visualization technique. 5b) Average response times together with their 95% confidence intervals for each tested visualization technique.

VI. Outlook

Several parts of aim A_3 and a last milestone for aim A_2 are still work in progress. We want to outline the plan for the remaining work in this section.

A. Use Case Design

Often, frameworks for processing biological or medical data are module based: data objects, for example microscopy images, are processed using one or multiple selected modules. Those systems enable variable work flows and, thus, less regulate the users procedures. Usually, detailed use cases play a less important role. Example frameworks are Amira [58], MeVisLab [59] and SCIRun [60] where Amira was used for the underlying project of this work.

During the project, we regularly met with our cooperation partners from the faculty of neurobiology at the Free University of Berlin. Our partners were students, research assistants as well as professors. In those meetings, we repeatedly observed the user stories (U_1, U_2) and the missing components (aim A_1 and A_2) outlined in section II. Aim A_1 and A_2 were prototyped in Amira.

Watching our cooperation partners using their common and new components, we became convinced that a more specific framework design would be better suited for the neurobiologists' work. Therefore, we want to work out our user stories into detailed use cases. These use cases should be formed in such a way that they could be used as a template for application development.

We plan to perform two interview sessions with a student and a research assistant from neurobiology. In these sessions, we want to refine our user stories using the approach described in [23].

B. User Study's Second Run Result Analysis

The second study's procedure (see section V-B was used for another run using a changed set-up. This second run took place during an open day for the *Lange Nacht der Wissenschaft*. 78 participants finished the study. The set-up differed in the following parts: the study was performed in an open area where also other experiments and presentations took place. Thus, quietness, optimal lighting and the participant's seating could not be controlled as in the first run. Further, a wider age range (7 years to 75 years) than in the first run (24 years to 55 years) took part in the study.

The second run's result analysis is an open task and will be done in the upcoming weeks. Besides the analysis, which was also done for the study's first run, we will compare both runs with each other. One question would be if the suboptimal set-up produced significantly different results than the idealized set-up. Another interesting point for discussion would be whether one set-up is more realistic than the other. We will further focus on analyzing the results within and between different age groups.

The results containing correct answers, response times and the participants' additional information are available in structured csv-files. For the analysis, no specific tools are required.

VII. Summary

In this paper, we presented approaches to support the work with neuroanatomical atlases. The main goals are to develop and evaluate visualization approaches that reduce needed expert knowledge and, further, to define specific use cases. The detailed design and description of use cases for initially presented user stories has not been finished yet. Additionally, the results of a second study run evaluating a line visualization approach has still to be analyzed.

References

[1] A. Toga and P. Thompson, "Maps of the Brain," *The Anatomical Record Part B: The New Anatomist*, vol. 265, no. 2, pp. 37–53, 2001.

[2] A. MacKenzie-Graham, E.-F. Lee, I. Dinov, M. Bota, D. Shattuck, S. Ruffins, Y. Heng, F. Konstantinidis, A. Pitiot, Y. Ding, G. Hu, R. Jacobs, and A. Toga, "A Multimodal, Multidimensional Atlas of the C57BL/6J Mouse Brain," *J. Anatomy*, vol. 204, no. 2, pp. 93–102, 2004.

[3] R. Brandt, T. Rohlfing, J. Rybak, S. Krofczik, A. Maye, M. Westerhoff, H.-C. Hege, and R. Menzel, "Three-Dimensional Average-Shape Atlas of the Honeybee Brain and Its Applications," *The Journal of Comparative Neurology*, vol. 492, no. 1, pp. 1–19, 2005.

[4] A. Ku, H.-C. Hege, S. Krofczik, and J. Brner, "Pipeline for the Creation of Surface-Based Averaged Brain Atlases," in *Proceedings of the WSCG 2007*, vol. 15, 2007, pp. 17–24.

[5] A. Maye, T. H. Wenckebach, and H.-C. Hege, "Visualization, reconstruction, and integration of neuronal structures in digital brain atlases," *International Journal of Neuroscience*, vol. 116, no. 4, pp. 431–459, 2006.

[6] J. Boline, E.-F. Lee, , and A. W. Toga, "Digital atlases as a framework for data sharing," *Frontiers in Neuroscience*, vol. 2, no. 1, pp. 100–106, 2008.

[7] A. Kuß, H.-C. Hege, and M. Gensel, "Digital neuroanatomy, data visualization, and modelling," 2006. [Online]. Available: http://www.zib.de/de/projekte/projektarchiv/projektarchiv-detail/article/digineuro.html

[8] R. Baldock, J. Bard, A. Burger, N. Burton, J. Christiansen, G. Feng, B. Hill, D. Houghton, M. Kaufman, J. Rao, J. Sharpe, A. Ross, P. Stevenson, S. Venkataraman, A. Waterhouse, Y. Yang, and D. Davidson, "EMAP and EMAGE: A Framework for Understanding Spatially Organized Data," *Neuroinformatics*, vol. 1, no. 4, pp. 309–325, 2003.

[9] M. Martone, I. Zaslavsky, A. Gupta, A. Memon, J. Tran, W. Wong, L. Fong, S. Larson, and M. Ellisman, *Anatomy Ontologies for Bioinformatics*, ser. Computational Biology. Springer, 2008, ch. 13, pp. 267–286.

[10] L. Ng, S. Pathak, C. Kuan, C. Lau, H. Dong, A. Sodt, C. Dang, B. Avants, P. Yushkevich, J. Gee, D. Haynor, E. Lein, A. Jones, and M. Hawrylycz, "Neuroinformatics for genome-wide 3-d gene expression mapping in the mouse brain," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 382–393, 2007.

[11] W. Pereanu and V. Hartenstein, "Digital Three-Dimensional Models of Drosophila Development," *Curr. Opin. in Genetics & Development*, vol. 14, no. 4, pp. 382–391, 2004.

[12] K. Rein, M. Zoeckler, M. T. Mader, C. Gruebel, and M. Heisenberg, "The Drosophila Standard Brain," *Current Biology*, vol. 12, no. 3, pp. 227–231, 2002.

[13] P. Kvello, J. Rybak, B. B. Lofaldli, A. Ku, and H. Mustaparta, "A digital, three-dimensional standard brain of the moth, heliothis virescens," in *Frontiers in Neuroinformatics. Conference Abstract: Neuroinformatics 2008*, 2008.

[14] J. Rybak, A. Ku, H. Lamecker, S. Zachow, H.-C. Hege, and R. Menzel, "The Digital Bee Brain: A platform for integrating neurons into a common 3D space using semi-automic and automated methods," *Frontiers in Systems Neuroscience. Special Topic: Digital Brain Atlases.*, 2009, (accepted).

[15] A.-S. Dadzie and A. Burger, "Providing visualisation support for the analysis of anatomy ontology data," *BMC Bioinformatics*, vol. 6, no. 2, p. 74, 2005.

[16] D. A. Schiffmann, D. Dikovskaya, P. L. Appleton, I. P. Newton, D. A. Creager, C. Allan, I. S. Nthke, and I. G. Goldberg, "Open microscopy environment and findspots: integrating image informatics with quantitative multidimensional image analysis," *BioTechniques*, vol. 41, no. 2, pp. 199–208, 2006.

[17] J. F. Brinkley, B. A. Wong, K. P. Hinshaw, and C. Rosse, "Design of an anatomy information system," *IEEE Computer Graphics and Applications*, vol. 19, no. 3, pp. 38 – 48, 1999.

[18] R. Bennett, Y. Ma, A. Siram, M. McGuigan, and H. Benveniste, "Mouse brain image visualizer (mbiv)," 2005. [Online]. Available: http://vis7.bnl.gov/MouseAtlas/Visualization/main.html

[19] *Allen Brain Atlas*, Allen Institute for Brain Science, 2006.

[20] S. Bruckner, V. Soltszov, E. Grller, J. Hladvka, K. Bhler, J. Y. Yu, and B. J. Dickson, "Braingazer - visual queries for neurobiology research," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1497–1504, 2009. [Online]. Available: http://www.cg.tuwien.ac.at/research/publications/2009/bruckner-2009-BVQ/

[21] B. Shneiderman and C. Plaisant, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 5th ed. Addison Wesley, 2010.

[22] S. J. Fernstad, J. Johansson, S. Adams, J. Shaw, and D. Taylor, "Visual Exploration of Microbial Populations," in *2011 IEEE Symposium on Biological Data Visualization (BioVis)*, J. Kennedy and J. Roerdink, Eds., 2011, pp. 127 – 134.

[23] J. Cordes, J. Dornheim, and B. Preim, "Szenariobasierte entwicklung von systemen fr training und planung in der chirurgie," *iCom*, vol. 8, no. 1, pp. 5–12, 2009.

[24] M. B. Rosson and J. M. Carroll, "Scenario-based design," in *The Human-Computer Interaction Handbook*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 2003, pp. 1032–1050.

[25] D. L. Rubin, S. E. Lewis, C. J. Mungall, S. Misra, M. Westerfield, M. Ashburner, I. Sim, C. G. Chute, H. Solbrig, M.-A. Storey, B. Smith, J. Day-Richter, N. F. Noy, and M. A. Musen, "National center for biomedical ontology: advancing biomedicine through structured organization of scientific knowledge," *OMICS A Journal of Integrative Biology*, vol. 10, no. 2, pp. 185–198, 2006.

[26] R. B. Albert Burger, Duncan Davidson, Ed., *Anatomy Ontologies for Bioinformatics. Principle and Practice*, ser. Computational Biology. Springer, 2008.

[27] M. Bota and L. W. Swanson, "BAMS Neuroanatomical Ontology: Design and Implementation," *Frontiers in Neuroinformatics*, vol. 2, 2008.

[28] J. M. Niggemann, A. Gebert, and S. Schulz, "Modeling Functional Neuroanatomy for an Anatomy Information System," *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 671–678, 2008.

[29] L. L. Fong, S. D. Larson, A. Gupta, C. Condit, W. J. Bug, L. Chen, R. West, S. Lamont, M. Terada, and M. E. Martone, "An ontology-driven knowledge environment for subcellular neuroanatomy," in *CEUR Workshop Proceedings*, 2007, pp. 1613–0073.

[30] "Mouse birn atlasing toolkit (mbat)," http://cms.loni.ucla.edu/mbat/index.aspx, 2009.

[31] D. Gilbert, M. Schroeder, and J. van Helden, "Interactive visualization and exploration of relationships between biological objects," *Trends in Biotechnology*, vol. 18, no. 12, pp. 487–494, 2000.

[32] H. Hauser, "Generalizing focus+context visualization," Ph.D. dissertation, Vienna University of Technology, Mar. 2004.

[33] I. Viola, M. Feixas, M. Sbert, and M. E. Grller, "Importance-driven focus of attention," *IEEE TVCG*, vol. 12, no. 5, pp. 933–940, 2006.

[34] P. Rautek, S. Bruckner, and M. E. Grller, "Interaction-Dependent Semantics for Illustrative Volume Rendering," in *Proc. Eurographics / IEEE VGTC Symposium on Visualization*, vol. 27, no. 3, May 2008, pp. 847–854.

[35] R. Kosara, "Semantic depth of field - using blur for focus + context visualization," Ph.D. dissertation, Institute of Computer Graphics and Algorithms, Vienna University of Technology, 2001. [Online]. Available: http://www.cg.tuwien.ac.at/research/publications/2001/Kosara-thesis/

[36] O. Mallo, R. Peikert, C. Sigg, and F. Sadlo, "Illuminated lines revisited," in *Proceedings of IEEE Visualization 2005*, 2005, pp. 19–26.

[37] M. Zöckler, D. Stalling, and H.-C. Hege, "Interactive Visualization of 3d-Vector Fields using Illuminated Stream Lines," in *VIS '96: Proceedings of the 7th conference on Visualization '96*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1996, pp. 107–ff.

[38] S. Oeltze and B. Preim, "Visualization of vasculature with convolution surfaces: method, validation and evaluation," *IEEE Transactions on Medical Imaging*, vol. 24, no. 4, pp. 540 –548, 2005.

[39] D. Merhof, M. Sonntag, F. Enders, C. Nimsky, and Guenther Greiner, "Hybrid Visualization for White Matter Tracts using Triangle Strips and Point Sprites," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1181–1188, 2006.

[40] C. Stoll, S. Gumhold, and H.-P. Seidel, "Visualization with Stylized Line Primitives," in *IEEE Visualization 2005 (VIS 2005)*, C. T. Silva, E. Gröller, and H. Rushmeier, Eds., Minneapolis, USA, 2005, pp. 695–702.

[41] V. Interrante and C. Grosch, "Strategies for Effectively Visualizing 3D Flow with Volume LIC," in *Proceedings of IEEE Visualization*, 1997, pp. 421–424.

[42] I. Viola and M. E. Grller, "Smart visibility in visualization," in *Proceedings of EG Workshop on Computational Aesthetics Computational Aesthetics in Graphics, Visualization and Imaging*, 2005, pp. 209–216.

[43] V. Interrante, H. Fuchs, and S. Pizer, "Enhancing Transparent Skin Surfaces with Ridge and Valley Lines," in *Proceedings of IEEE Visualization*, 1995, pp. 52–59.

[44] ——, "Illustrating Transparent Surfaces with Curvature-Directed Strokes," in *VIS '96: Proceedings of the 7th conference on Visualization '96*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1996, pp. 211–ff.

[45] J. Diepstraten, D. Weiskopf, and T. Ertl, "Transparency in Interactive Technical Illustrations," *Computer Graphics Forum*, vol. 21, no. 3, pp. 317–325, 2002. [Online]. Available: http://www.eg.org/EG/CGF/volume21/issue3/abstracts/CGF591.HTML

[46] B. Preim, *Entwicklung interaktiver Systeme: Grundlagen, Fallbeispiele und innovative Anwendungsfelder*. Springer-Verlag Berlin Heidelberg, 1999.

[47] C. Tietjen, T. Isenberg, and B. Preim, "Combining silhouettes, surface, and volume rendering for surgery education and planning," in *Proceedings of EUROGRAPHICS - IEEE VGTC Symposium in Visualization*, K. W. Brodlie, D. J. Duke, and K. I. Joy, Eds., 2005, pp. 303–310.

[48] T. S. Tullis and J. N. Stetson, "A Comparison of Questionnaires for Assessing Website Usability," in *Proceedings of UPA 2004*. 140 N. Bloomingdale Road, Bloomingdale, IL 60108-1017: UPA, 2004.

[49] A. Baer, F. Adler, D. Lenz, and B. Preim, "Perception-based evaluation of emphasis techniques used in 3d medical visualization," in *Vision, Modeling, and Visualization Workshop (VMV)*, 2009.

[50] H. Petersson, D. Sinkvist, C. Wang, and r. Smedby, "Web-based interactive 3D visualization as a tool for improved anatomy learning," *Anatomical Sciences Education*, vol. 2, no. 2, pp. 61–68, 2009.
[51] A. K. Qayumi, Y. Kurihara, M. Imai, G. Pachev, H. Seo, Y. Hoshino, R. Cheifetz, K. Matsuura, M. Momoi, M. Saleem, H. Lara-Guerra, Y. Miki, and Y. Kariya, "Comparison of computer-assisted instruction (CAI) versus traditional textbook methods for training in abdominal examination (Japanese experience)," *Medical Education*, vol. 38, no. 10, pp. 1080–1088, 2004.
[52] F. Ritter, B. Berendt, B. Fischer, R. Richter, and B. Preim, "Virtual 3D Jigsaw Puzzles: Studying the Effect of Exploring Spatial Relations with Implicit Guidance," in *Proceedings of Mensch & Computer*, Hamburg, September 2002, pp. 363–372.
[53] A. Ku, S. Prohaska, and J. Rybak, "Using ontologies for the visualization of hierarchical neuroanatomical structures," in *Frontiers in Neuroinformatics. Conference Abstract: 2nd INCF Congress of Neuroinformatics*, 2009.
[54] A. Ku and H.-C. Hege, "Knowledge Representation for Digital Atlases," in *Conference Abstracts from the Workshop of the Working Group Ontologies in Biomedicine and the Life Science*, 2009.
[55] V. J. Dercksen, M. Gensel, and A. Ku, "Visual accentuation of spatial relationships between filamentous and voluminous surface structures," in *Eurographics/IEEE Symposium on Visualization 2009, Conference Abstract*, 2009.
[56] M. Gensel, "Visualisierungsmethoden zur Verdeutlichung der rumlichen Beziehungen zwischen linien- und flchenartigen Strukturen am Beispiel neurobiologischer Daten," Diploma Thesis, Free University of Berlin, 2009.
[57] A. Ku, B. Meyer, V. J. Dercksen, M. Gensel, and S. Prohaska, "Effective Techniques to Visualize Filament-Surface Relationships," in *Eurographics/ IEEE-VGTC Symposium on Visualization (2010)*, G. Melanon, T. Munzner, and D. Weiskopf, Eds., 2010, (submitted).
[58] "Amira. visualize - analyze - present," 2012. [Online]. Available: http://www.amira.com/
[59] "Mevislab. medical image processing and visualization," 2011. [Online]. Available: http://www.mevislab.de/home/about-mevislab/
[60] SCIRun: A Scientific Computing Problem Solving Environment, Scientific Computing and Imaging Institute (SCI). [Online]. Available: http://www.scirun.org

Chancen und Potentiale der Altersbestimmung latenter Fingerspuren mittels kontaktloser Sensorik

Ronny Merkel
Arbeitsgruppe Multimedia & Security
Otto-von-Guericke Universität Magdeburg
Universitätsplatz 2, 39106 Magdeburg
Deutschland
ronny.merkel@iti.cs.uni-magdeburg.de
Betreuende Hochschullehrerin: Prof. Dr.-Ing. Jana Dittmann

Abstrakt—**Die Altersbestimmung latenter Fingerspuren am Tatort ist eine seit vielen Jahrzehnten ungelöste Herausforderung für forensische Experten. Im Rahmen der vorliegenden Arbeit wird ein Dissertationsvorhaben beschrieben, welches die Eignung kontaktloser, optischer Sensorik zur Erstellung von Fingerabdruck-Zeitreihen und deren Weiterverarbeitung mittels digitaler Verarbeitungsmethoden zur Altersbestimmung untersucht. Neben der Definition einer grundlegenden Verarbeitungskette und eines Vorgehensmodells wird im Rahmen bisheriger Arbeiten das Binary Pixel Merkmal als ein erstes charakteristisches Alterungsmerkmal präsentiert und anhand aufgestellter Arbeitspakte evaluiert. Erste Testergebnisse zeigen eine Genauigkeit von ca. 70 - 80% für die Einteilung von Fingerabdrücken in solche jünger oder älter als 5 Stunden, was bereits eine starke Verbesserung bisheriger Ansätze darstellt. Weitere Merkmale und Sensoriken sollen in zukünftigen Arbeiten des Dissertationsvorhabens untersucht und in geeigneter Weise mit dem Binary Pixel Merkmal kombiniert werden, um die Klassifikationsgenauigkeit weiter zu verbessern.**

Schlüsselworte - Tatortforensik, Altersbestimmung, latente Fingerspuren, Binary Pixel, kontaktlose Aufnahmeverfahren, Chromatischer Weißlichtsensor

I. Kontext und Forschungslücke

Die Aufklärung von Verbrechen und anderen kriminellen Handlungen ist eine wichtige und verantwortungsvolle Aufgabe einer jeden Gesellschaft. Dabei besteht die besondere Schwierigkeit im Auffinden und geeignetem Sichern von Spuren, sowie der Identifizierung potentieller Täter und der Rekonstruktion von Tathergängen. Hierzu werden oftmals latente Fingerspuren genutzt, welche von Menschen unbewusst durch das Berühren von Gegenständen hinterlassen werden und meist nicht auf den ersten Blick erkennbar sind. Solche Spuren sind an vielen Tatorten zu finden und erleichtern forensische Untersuchungen durch die Identifikation von Personen, welche sich am Tatort aufhielten.

Ein wichtiger Aspekt der Nutzung latenter Fingerspuren in kriminalistischen Untersuchungen ist die Altersbestimmung. Dabei bezeichnet das Alter eines Fingerabdrucks die Zeit, welche vom Hinterlassen der Spur bis zu deren Erfassung durch forensische Experten vergeht. Eine solche Altersbestimmung bietet die Möglichkeit, Fingerspuren einem Tatzeitpunkt zuzuordnen, um somit gängige Argumente zu entkräften, welche die Anwesenheit eines Verdächtigen am Tatort einräumen, diese jedoch auf Zeitpunkte außerhalb der Tatzeit beschränken. Des Weiteren bietet das Alter von Fingerspuren das Potential, Tathergänge zu rekonstruieren, indem die Berührung von Objekten, auf denen ein Fingerabdruck hinterlassen wurde, in eine zeitliche Reihenfolge gebracht wird. Auch die Sequenzierung überlappender Fingerspuren und somit die Bestimmung der Reihenfolge der Berührung eines Objektes von unterschiedlichen Personen kann so ermöglicht werden.

Trotz des enormen Nutzens, welcher aus der Altersbestimmung latenter Fingerspuren erwächst und welcher forensischen Experten seit mehreren Jahrzehnten bekannt ist, konnte diese Forschungsherausforderung bisher nicht gelöst werden. Der Titel einer Veröffentlichung von Kasey Wertheim im Jahr 2003 [1] trägt den Titel „Fingerprint age determination – is there any hope?“ (Altersbestimmung von Fingerabdrücken – gibt es irgendeine Hoffnung?) und beschreibt anschaulich die Problematik dieser Forschungslücke.

Durch die signifikante Verbesserung kontaktloser Sensorik in den Bereichen der Oberflächenmessung, Mikroskopie und Spektroskopie kann diese Frage nun erneut aufgerollt und mit digitalen Bildverarbeitungsmethoden (z.B. aus der Biometrie, der Mustererkennung oder der Signalverarbeitung) auf hochaufgelösten Fingerspuren untersucht werden. Dabei ergeben sich die folgenden Forschungsfragen, deren Beantwortung grundlegend für eine mögliche Lösung dieser Herausforderung ist:

- Unter welchen Voraussetzungen ist eine Altersbestimmung latenter Fingerspuren mit zerstörungsfreier Sensorik möglich?
- Welche Mustererkennungsansätze sind sinnvoll und zielführend?
- Welche Einschränkungen sind zu erwarten?
- Welche Merkmale sind geeignet mit welcher Genauigkeit / Fehlerrate?
- Wie performant würde ein solches System arbeiten (bezüglich Scanzeit, Rechenzeit, Skalierbarkeit, etc.)?

Die Beantwortung dieser Fragen soll im Mittelpunkt der hier beschriebenen Arbeit stehen. Dabei ist es das Ziel, das Potential kontaktloser Sensorik aufzuzeigen und mögliche Vorgehensweisen zu beschreiben, um eine Altersbestimmung von Fingerspuren für kriminalistische Untersuchungen zu erschließen.

Als potentielle Merkmale zur Altersbestimmung werden dabei Eigenschaften von Fingerspuren angesehen, welche sich charakteristisch über die Zeit verändern. Dazu können neben klassischen Merkmalen wie Rillen, Poren oder Minutien auch statistische Merkmale wie Kontrast, Rauheit oder Mittelwert aber auch Merkmale der chemischen Zusammensetzung gehören.

Im Folgenden wird zunächst der Stand der Technik in Kapitel 2 zusammengefasst. Kapitel 3 beschäftigt sich mit dem Design eines Konzeptes zur Evaluierung des Potentials kontaktloser Sensorik zur Altersbestimmung von Fingerspuren. Dabei werden Ziele, Methoden, Qualitätskriterien und Fehlerraten sowie konkrete Arbeitspakete formuliert. In Kapitel 4 werden eigene Forschungsergebnisse bezüglich des vielversprechenden Binary Pixel Alterungsmerkmals präsentiert, welche erste Rückschlüsse zur Beantwortung der gestellten Forschungsfragen zulassen. Weitere Schritte für zukünftige Arbeiten werden ebenfalls diskutiert.

II. Stand der Technik

Bisherige Arbeiten zur Altersbestimmung latenter Fingerspuren lassen sich nach ihrer Erfassungsmethode aufteilen. Dabei wird im Rahmen dieses Beitrags zwischen der kontaktbehafteten und somit den Fingerabdruck verändernden bzw. zerstörenden sowie der kontaktlosen und somit (meist) zerstörungsfreinen Erfassung unterschieden. Kontaktbehaftete Methoden beinhalten jegliche Art von Sichtbarmachung eines Fingerabdrucks durch Vorbehandlung, wie beispielsweise dem Einfärben, Berußen oder Vorbehandeln mit Ninhydrin oder Cyanoacrylat. Weiterhin beinhalten sie die zerstörungsbehaftete Aufnahme des Fingerabdrucks, wie beispielsweise das Abziehen mit Klebeband. Kontaktbehaftete Verfahren verändern daher potentiell sowohl die Morphologie eines Abdrucks, als auch dessen chemische Zusammensetzung. Kontaktlose Verfahren sind meist strahlungsbasiert und können bei sehr energiereicher Strahlung Abbauprozesse in der Fingerspur hervorrufen, welche deren chemische Zusammensetzung verändern. Im visuellen und infraroten Bereich sind sie jedoch relativ energiearm und können hier als zerstörungsfrei betrachtet werden.

A. *Kontaktbasierte Erfassungsmethoden*

Kontaktbasierte Erfassungsmethoden verändern einen Fingerabdruck entweder durch die Vorbehandlung mit bestimmten Substanzen zur Sichtbarmachung oder durch physikalische Veränderungen bei der Abnahme. Fotografierte Fingerspuren gelten daher ebenso als kontaktbasiert, wenn sie vorher durch Einbringung einer Substanz sichtbar gemacht wurden.

Bisherige Untersuchungen zum Alterungsverhalten von Fingerspuren beschränken sich hauptsächlich auf die Anwendung kontaktbasierter Erfassungsmethoden. Baniuk et al. untersuchten in [2], ob sich das Alter einer Fingerspur durch manuellen Vergleich eines menschlichen Beobachters mit Fingerabdrücken unterschiedlichen Alters aus einer Referenzdatenbank abschätzen lässt, jedoch ohne zuverlässige Ergebnisse zu erzielen. Popa et al. untersuchten in [3] u.a. die Rillenbreite, Talbreite sowie die Anzahl und Eigenschaften von Poren eingefärbter Fingerabdrücke über einen Zeitraum von 180 Tagen. Dabei stellten sie signifikante Veränderungen dieser Elemente fest, konnten allerdings aufgrund der hohen Variabilität keine zuverlässigen Korrelationen zwischen Alter einer Spur und Merkmalsausprägung herstellen. Jörg Ähnlich untersuchte im Rahmen seiner Diplomarbeit [4] das Fluoreszenzverhalten von Fingerschweiß, welcher hierzu größtenteils in Küvetten abgefüllt und mit Laserlicht bestrahlt wurde. Es zeigten sich charakteristische Veränderungen des Fluoreszenzverhaltens für bestimmte Substanzen. Seine Forschungen werden durch diverse Untersuchungen zur chemischen Zusammensetzung von Fingerabdrücken ergänzt, welche beispielsweise von Mong et al. [5], Wolstenholme et al. [6] und De Paoli [7] durchgeführt wurden. Untersuchungen zur chemischen Zusammensetzung von Fingerabdrücken nutzen meist die Massenspektrometrie als Methode zur Elementbestimmung, welche Fingerabdruckmaterial auflöst und somit physisch zerstört. Die Arbeiten bestätigen eine hohe Variabilität der chemischen Zusammensetzung von Fingerspuren, welche eine zuverlässige Altersbestimmung erschwert.

Die hohe Variabilität in der chemischen Zusammensetzung von Fingerspuren sowie die Vielzahl unterschiedlicher Einflüsse auf das Alterungsverhalten, welche neben den Umwelteinflüssen (z.B. Temperatur, Luftfeuchte, Wind, UV-Strahlung) auch die Art des Schwitzens (z.B. ekkriner vs. talghaltiger Schweiß, Schwitzen durch Sport oder hohe Umgebungstemperaturen), Scanparameter (z.B. Auflösung und Größe der Messfläche) , unterschiedliche Oberflächen und Art der Aufbringung (z.B. Anpressdruck, Anpressdauer, Verschmieren, Kontamination des Fingers) umfassen, können als Ursachen dafür gesehen werden, dass die vorgestellten Ansätze bisher keine verwertbaren Ergebnisse zur Altersbestimmung von Fingerspuren hervorbringen konnten. Ein weiterer wichtiger Grund ist jedoch auch in der kontaktbehafteten Erfassung zu sehen, welche die wiederholte Aufnahme eines spezifischen Fingerabdrucks und somit die Erstellung von Zeitreihen unterbindet. Weiterhin werden vorbehandelte Fingerabdrücke meist fotografiert, was zu einer nur mäßig guten Auflösung der Spur führt. Digitale Verarbeitungsmethoden (z.B. aus der Biometrie, Mustererkennung oder Signalverarbeitung) werden in sehr geringem Maß zur Auswertung eingesetzt, obwohl sie ein großes Potential darstellen.

B. *Kontaktlose Erfassungsmethoden*

Kontaktlose Erfassungsmethoden beschränken sich bisher hauptsächlich auf die Analyse der chemischen Zusammensetzung von Fingerspuren mittels Spektroskopie. Crane et al. zeigten in [8], dass Fourier-Transform Infrarot Spektroskopie (FTIR) die Möglichkeit bietet, Fingerabdrücke auch auf

schlecht reflektierenden Oberflächen sichtbar zu machen, welche mit bloßem Auge oft nicht erkannt werden können. Antoine et al. [9] und Williams et al. [10] nutzten ebenfalls FTIR-Sensorik zur Untersuchung von Unterschieden in der chemischen Zusammensetzung zwischen Fingerabdrücken von Kindern und Erwachsenen und beobachteten ausgewählte Substanzgruppen über die Zeit. Dabei konnten sie feststellen, dass sich bestimmte Substanzen in Fingerabdrücken mit der Zeit quantitativ verändern (z.B. Ester, bestimmte Fette oder Eiweiße), während andere relativ konstant bleiben (z.B. Salze). Allerdings wurden keine konkreten Altersbestimmungsansätze untersucht, was mit der hohen Variabilität in der chemischen Zusammensetzung zwischen unterschiedlichen Fingerabdrücken begründet werden kann. Über die Analyse der chemischen Zusammensetzung hinausgehende Untersuchungen mittels kontaktloser Erfassungsmethoden sind nicht bekannt.

Im Rahmen des vorgestellten Dissertationsvorhabens werden kontaktlose Erfassungsmethoden speziell für die Aufnahme morphologischer Fingerabdruckmerkmale untersucht. Ein solcher Ansatz ist neu und bietet das Potential einer späteren Kombination mit chemischen Methoden, wobei jedoch die Bildgebung und digitale Auswertung von Fingerabdruckbildern im Vordergrund steht. Der große Vorteil einer kontaktlosen Erfassung kann dabei in der wiederholten Aufnahme einer Fingerspur in regelmäßigen Zeitabständen und somit der Erstellung von Zeitreihen gesehen werden. Ein solches Vorgehen ist mit kontaktbehafteten Verfahren nicht möglich, da hier die Fingerspur nach der ersten Erfassung verändert ist und somit kein weiteres Mal erfasst werden kann. Die Erstellung von Zeitreihen bietet hingegen die Möglichkeit, charakteristische Veränderung einer bestimmten Fingerspur über die Zeit zu beobachten, auszuwerten und charakteristische Alterungsmerkmale aus ihr zu extrahieren.

Weiterhin bieten kontaktlose Erfassungsmethoden oftmals den Vorteil, deutlich hochaufgelöstere Daten bereitzustellen, als dies bei herkömmlichen Aufnahmeverfahren, wie etwa dem Fotografieren nach einer Vorbehandlung des Fingerabdrucks, der Fall ist. Die hochaufgelösten Daten können im Anschluss mit digitalen Verarbeitungsmethoden untersucht werden, um Störungen zu entfernen, Fingerabdrücke zu segmentieren oder Alterungsmerkmale aus ihnen zu berechnen.

III. Design eines Konzepts zur Evaluierung des Potentials kontaktloser Sensorik zur Altersbestimmung

Um das Potential kontaktloser Sensorik zur Altersbestimmung von Fingerspuren zu bewerten, müssen zunächst konkrete Ansätze definiert und praktisch evaluiert werden. Dabei ist es aufgrund der extremen Komplexität der möglichen Zusammenhänge nicht realistisch, alle potentiell nutzbaren Sensoriken, alle auf ihnen definierbaren Merkmale, alle möglichen veränderlichen Parameter sowie alle auftretenden Einflüsse im Detail zu untersuchen. Vielmehr ist die Untersuchung ausgewählter und vielversprechender Kombinationen von Sensorik, Merkmalen, Messparametern, deren Abhängigkeiten von unterschiedlichen Einflüssen sowie Klassifikationsstrategien beispielhaft zu untersuchen, um daraus verallgemeinernde Schlüsse über die prinzipielle Machbarkeit der Altersbestimmung sowie deren Genauigkeit und Grenzen zu ziehen. Weiterhin kann ein im Rahmen solcher Untersuchungen effiziertes Vorgehen später auf beliebige andere Sensoriken und Merkmale angewandt werden, um in einem Fusionsansatz die Leistung der Altersbestimmung zu verbessern. Das Hauptziel der hier vorgestellten Arbeit besteht daher in einer Abschätzung des Potentials kontaktloser Sensorik anhand ausgewählter Beispiele sowie der Beschreibung eines allgemeinen Vorgehens bei der Extraktion charakteristischer Merkmale.

Im Folgenden wird ein Überblick über die Teilziele des Dissertationsvorhabens sowie angewandter Methoden zu deren Realisierung beschrieben. Es werden weiterhin Qualitätskriterien zur Leistungsbewertung und zum Vergleich definiert.

A. Ziele

Die folgenden fünf Teilziele werden als maßgeblich für die Beantwortung der in Kapitel 1gestellten Forschungsfragen angesehen:

1. Eine Mustererkennungskette von der Aufnahme einer Fingerspur bis zur finalen Altersabschätzung und deren Bewertung ist zu definieren und allgemeingültig zu beschreiben, sodass sie auf eine Vielzahl von Sensoren, Vorverarbeitungs-, Segmentierungs- und Merkmalsextraktionsmethoden sowie verschiedene Klassifikationsstrategien angewendet werden kann.
2. Mindestens drei unterschiedliche und möglichst weitgehend voneinander unabhängige Merkmale sind zu definieren und zu untersuchen.
3. Unterschiedliche Einflüsse auf die ausgewählten Merkmale und die zugehörigen Sensoriken sind bezüglich ihrer Relevanz zu untersuchen und durch Ausschluss oder Einbezug entsprechend zu berücksichtigen.
4. Eine Diskussion und Auswahl bzw. Definition von Klassifikationsstrategien, Qualitätskriterien und Fehlerraten zur praktischen Altersabschätzung und deren Bewertung ist vorzunehmen.
5. Ausgewählte Merkmale und zugehörige Klassifikationsstrategien sind anhand einer statistisch signifikanten Testmenge und für verschiedene, an Tatorten vorkommende Oberflächen zu bewerten und zu vergleichen.

B. Methoden

Im Folgenden werden die wichtigsten zur Umsetzung des Vorhabens ausgewählten Methoden vorgestellt:

a. Zeitreihen: Eine der wichtigsten Methoden zur systematischen Evaluation des Alterungsverhaltens von Fingerspuren ist die Erstellung von Zeitreihen, welche eine präzise Beobachtung der Veränderungen innerhalb eines spezifischen Fingerabdrucks über die Zeit ermöglicht. Die wiederholte Messung von Fingerspuren in regelmäßigen Zeitintervallen ist nur mit kontaktloser, zerstörungsfreier Sensorik möglich und stellt somit einen erheblichen Neuwert dar.

b. Sensorik: Unterschiedliche Erfassungsdomänen sind zu untersuchen, um verschiedene Merkmale zu evaluieren, wie 2D-Intensitäts oder 3D-Topographiedaten eines Chromatischen Weißlichtsensors (CWL, [11]) aber auch anderer Sensoren, wie beispielsweise Mikroskope oder Spektroskope.

c. Vorverarbeitung: Nach der Erfassung von Fingerspuren durch einen Sensor sind diverse Verfahren zur Normalisierung, Segmentierung und Maskierung von Umwelteinflüssen erforderlich, welche Fingerabdruckbilder für die Merkmalsextraktion vorbereiten.

d. Merkmalsextraktion: Die Extraktion von Merkmalen aus Fingerabdruckbildern erfolgt bezüglich spezifischer, sich über die Zeit charakteristisch verändernder Eigenschaften des Fingerabdrucks, welche meist für bestimmte Zeitintervalle charakteristisch sind (z.B. Kurzzeitalterung bis zu einigen Tagen, Langzeitalterung über Wochen, Monate und Jahre). Dabei sind Merkmale meist auf spezifische Anwendungsszenarien beschränkt (genutzte Sensorik, Oberflächenmaterial, Umweltbedingungen, etc.).

e. Klassifikatoren & Merkmalsselektion: Durch den experimentellen Vergleich werden besonders charakteristische Merkmale und Klassifikatoren bestimmt, welche eine möglichst hohe Leistung bei der Altersabschätzung erzielen.

f. Mustererkennungskette und Vorgehensmodell: Die beschriebenen Methoden sind in eine zeitliche Reihenfolge zu bringen und als Mustererkennungskette bzw. Handlungsabfolge zu beschreiben, welche die Anwendung der Werkzeuge auf unterschiedliche Sensoriken, Merkmale und Klassifikationsstrategien erlaubt.

g. Fusion: Die Fusion wird als wichtiges Werkzeug betrachtet, um die Ergebnisse einzelner Kombinationen aus Sensor, Merkmal und Klassifikationsstrategie zu einem gemeinschaftlichen Altersbestimmungsansatz zusammenzuführen. Erst durch die Kombination unterschiedlicher Ansätze scheint eine ausreichende Gesamtleistung erreichbar zu sein

h. Statistische Signifikanz: Testergebnisse können nur dann als allgemeingültig angesehen werden, wenn sie auf einer Testmenge von statistischer Signifikanz evaluiert wurden.

i. Unterschiedliche Oberflächenmaterialien: Für eine praktische Anwendung von Altersbestimmungsansätzen am Tatort ist die Untersuchung unterschiedlicher, häufig am Tatort vorkommender Oberflächenmaterialien eine zwingende Voraussetzung.

j. Vergleich und Bewertung: Anhand definierter Qualitätskriterien und Fehlerraten kann eine objektive Bewertung der unterschiedlichen Sensoren, Merkmale und Klassifikationsstrategien erfolgen.

C. Qualitätskriterien und Fehlerraten

Um die Güte unterschiedlicher Sensoriken, Merkmale und Klassifikationsstrategien auf einer objektiven Grundlage zu vergleichen, ist die Definition von Qualitätskriterien und Fehlerraten von großer Bedeutung. Die folgenden Kriterien sollen dabei im Rahmen des vorliegenden Dissertationsvorhabends zur Anwendung kommen:

i. Korrelationskoeffizient: Der Pearson Korrelationskoeffizient ist ein weitbekanntes Maß zur Bestimmung der Korrelation zwischen Variablen. Er soll hier genutzt werden, um die Ähnlichkeit der Alterungskurve eines Merkmals (der Veränderung der Merkmalswerte über die Zeit) zu ihrer nächstliegenden idealen mathematischen Funktion zu bestimmen. Er ist daher ein Maß dafür, wie linear, logarithmisch oder exponentiell die Veränderung von Merkmalswerten ist.

ii. Regression: Über den Korrelationskoeffizienten hinaus kann mit Hilfe der Regression eine lineare, logarithmische oder exponentielle mathematische Funktion bezüglich der experimentellen Alterungskurve approximiert werden.

iii. Klassifikationsgenauigkeit: Für eine jede Zeitklasse kann die relative Häufigkeit der korrekt in diese Klasse eingeordneten Fingerabdrücke sowie die Menge der falsch eingeordneten Abdrücke bestimmt werden. Hieraus ergeben sich die bekannten Fehlerraten wahr positiv, falsch positiv, wahr negativ und falsch negativ. Weiterhin kann die Klassifikationsgenauigkeit eines Gesamtsystems als die relative Häufigkeit korrekt klassifizierter Elemente bestimmt werden.

iv. Kappa: Das statistische Maß des kappa wird aus der Klassifikationsgenauigkeit berechnet und bereinigt diese um die Wahrscheinlichkeit des Ratens sowie eine ungleiche Verteilung der Testmenge. Es kann daher als normierte Klassifikationsgenauigkeit angesehen werden.

v. Minimum, Maximum, Varianz, Standardabweichung: Statistische Maße von Wahrscheinlichkeitsverteilungen sind ebenfalls zum Vergleich geeignet, beispielsweise zur Bestimmung der Variabilität eines Merkmals.

vi. Reproduzierbarkeit: Erkenntnisse müssen für klar definierte Anwendungsbereiche und Sensoriken reproduzierbar sein.

D. Arbeitspakete

Das hier beschriebene Dissertationsvorhaben kann in 8 Arbeitspakete (AP) aufgeteilt werden, welche im Folgenden erläutert werden:

AP1: Literaturrecherche: Zunächst ist eine intensive Literaturrecherche durchzuführen, um den Stand der Technik zu erfassen, potentiell geeignete Sensorik zu identifizieren sowie mögliche Merkmale und Klassifikationsstrategien zu definieren. Der größte Teil dieses APs ist zu Beginn des Dissertationsvorhabens durchzuführen, allerdings ist eine parallele Auswertung neu hinzugekommener bzw. neu entdeckter Literatur während der gesamten Durchführung des Vorhabens von Bedeutung.

AP2: Design der Mustererkennungskette: Im Anschluss an eine intensive Literaturrecherche ist das Design einer Mustererkennungskette von der Aufnahme einer Zeitreihe am Sensor bis zur finalen Auswertung der Leistung einer Altersabschätzung von Bedeutung. Alle weiteren Arbeitspakte müssen sich an einer solchen Verarbeitungskette orientieren und diese schrittweise implementieren.

AP3: Identifikation von mindestens drei unterschiedlichen Merkmalen sowie zugehöriger Sensorik und Klassifikationsstrategien: Im Rahmen diverser Experimente sind in diesem Arbeitspaket drei charakteristische aber weitgehend voneinander unabhängige Kombinationen aus Sensor, Vorverarbeitung, Merkmal und Klassifikationsstrategie zu definieren.

AP4: Einflüsse auf die Alterung: Für die in AP3 definierten Merkmale sind unterschiedliche Einflüsse aus Umwelt, Schweißzusammensetzung, Oberflächen, Scanparametern und Art der Aufbringung hinsichtlich ihrer Relevanz zu untersuchen und auszuschließen bzw. zu berücksichtigen.

AP5: Definition von Klassifikationsstrategien und Erstellung eines statistisch signifikanten Testsets: Soweit in AP3 noch nicht vorgenommen, müssen in AP5 zu untersuchende Klassifikationsstrategien spezifiziert und ein diesbezüglich passendes Testset von statistisch signifikanter Größe erzeugt werden.

AP6: Auswertung, Vergleich und Anpassung der Klassifikationssysteme: Im sechsten Arbeitspaket sind die definierten Klassifikationssysteme anhand des in AP5 erstellten statistisch signifikanten Testsets und für unterschiedliche, häufig an Tatorten vorkommende Oberflächen zu evaluieren, unter Nutzung der definierten Qualitätskriterien und Fehlerraten zu vergleichen und ggf. Anpassungen für eine Optimierung der Genauigkeit vorzunehmen.

AP7: Fusion: In diesem Arbeitspaket sind mögliche Fusionsstrategien zu erörtern und praktisch bzgl. einer potentiellen Steigerung der Genauigkeit zu evaluieren.

AP8: Dokumentation und Anfertigung der Dissertationsschrift: Als paralleler Vorgang zu allen durchzuführenden Arbeitspakten ist eine detailgenaue Dokumentation der erzielten Ergebnisse sowie die Anfertigung und iterative Verbesserung der Dissertationsschrift vorzunehmen.

IV. Bisherige Erkenntnisse und weitere Arbeiten

Im bisherigen Verlauf des Dissertationsvorhabens wurde eine umfangreiche Literaturrecherche durchgeführt (AP1) und basierend auf ihren Erkenntnissen sowie unterschiedlichen Vortests eine Mustererkennungskette zur Altersbestimmung mittels kontaktloser Sensorik abgeleitet (AP2), welche im Wesentlichen die Erfassung von Fingerspuren am Sensor, deren Vorverarbeitung, Merkmalsextraktion, die Gewinnung von Match-Scores und eine finale Einteilung in ausgewählte Zeitklassen beinhaltet (siehe auch [12]). Ein allgemeines Vorgehensmodell für die Erstellung von Alterungsansätzen unterschiedlicher Anwendungsszenarien, Sensoriken, Merkmale und Klassifikationsstrategien wurde in [13] beschrieben und soll als Ausgangsbasis für alle untersuchten Merkmale dienen.

Für 2D-Intensitäts- sowie 3D-Topographiedaten des Chromatischen Weißlichtsensors (CWL, [11]) wurde im Rahmen des AP3 das Binary Pixel Merkmal [14] definiert. Das Merkmal zeigt eine charakteristische logarithmische Alterung über mehrere Stunden und Tage hinweg (Kurzzeitalterung) und wurde im Rahmen der APs 4 - 7 bezüglich seiner Abhängigkeiten, möglicher Klassifikationsstrategien, Genauigkeit, Fehlerraten und Fusionsmöglichkeiten untersucht. Das Binary Pixel Merkmal wird im nächsten Abschnitt detaillierter beschrieben. Weitere erforschte Merkmale beinhalten zudem die Analyse der Anzahl und Größe von Korrosionsartefakten auf korrodieren Oberflächen (z.B. Kupfermünzen) für die Langzeitalterung über mehrere Wochen und Monate [13].

A. *Das Binary Pixel Merkmal*

Im Folgenden wird das im Rahmen des Dissertationsvorhabens identifizierte Binary Pixel Merkmal beschrieben und die im Rahmen der Arbeitspakete durchgeführten Untersuchungen präsentiert.

AP3

Im Rahmen des AP3 wurde das Binary Pixel Merkmal als für die Altersbestimmung vielversprechend identifiziert [14]. Für die Berechnung des Merkmals wird ein Fingerabdruckbild zunächst normalisiert und binarisiert. Dabei kommen unterschiedliche Methoden zum Einsatz, wie beispielsweise dynamische Normalisierung, statische Normalisierung, Schwellwertbinarisierung, Otsu-Binarisierung oder Hintergrund-Maskierung (siehe auch [15]). Anschließend wir die relative Häufigkeit der Hintergrundpixel bestimmt, welche ein Maß für den Kontrast des Bildes darstellt. Mit zunehmendem Alter einer Spur verringert sich dieser Kontrast, indem zum Fingerabdruck gehörige Pixel in Hintergrundpixel übergehen, bedingt durch unterschiedliche Abbauprozesse sowie die Verdunstung von Wasser. Die Zunahme der relativen Häufigkeit von Hintergrundpixeln, welche das Binary Pixel Merkmal darstellt, vergrößert sich dabei mit einer logarithmischen Tendenz. Mit Hilfe der Regression und anderer Methoden können hier mathematische Alterungsfunktionen approximiert werden [16], [17], deren Ähnlichkeit zur experimentell bestimmten Alterungskurve mit Hilfe des Pearson Korrelationskoeffizienten objektiv angegeben werden kann.

AP4

Einflüsse aus Umwelt, Schweißzusammensetzung, Oberflächenmaterial, Scanparametern und Aufbringungsfaktoren wurden in [14] zusammengetragen. Die Einflüsse beinhalten ca. 50 Untereinflüsse, von welchen ein Großteil auf ihre Relevanz bezüglich des Binary Pixel Merkmals untersucht wurden.

Der Einfluss der Schweißzusammensetzung ist sehr komplex und sehr stark von der initialen Zusammensetzung des Fingerabdrucks abhängig. Im Rahmen von [15] und [18] wurde der Schweiß der ekkrinen Drüsen (welche sich auf der Fingerkuppe befinden und kaum talghaltige Substanzen absondern), Talgdrüsen (fetthaltiger Schweiß, vornehmlich an Stirn, Wange und Nacken zu finden) sowie Mischschweiß aus täglicher Arbeit (durch Berührung von Körperregionen beider Schweißdrüsen sowie diverser Gegenstände des alltägliche Lebens) analysiert. Weiterhin wurde das Alterungsverhalten von Fingerabdrücken nach dem Schwitzen unter heißen bzw. kühlen Umgebungsbedingungen sowie nach Sport und dem Konsum von Alkohol untersucht. Der Einfluss der Schweißzusammensetzung auf das Alterungsverhalten scheint signifikant zu sein,

allerdings ist die Analyse der chemischen Zusammensetzung als Haupteinfluss für die Vorhersage des Alterungsverhaltens mit dem optischen CWL Messgerät nur sehr bedingt möglich und erfordert weitere Untersuchungen mit spektroskopischen kontaktlosen Verfahren.

Der Einfluss unterschiedlicher Umweltbedingungen wurde in [15] und [18] analysiert und bestätigt einen klaren Einfluss von Temperatur, Luftfeuchte, UV-Strahlung und Wind auf das Alterungsverhalten von Fingerabdrücken. Während hohe Temperaturen, UV-Strahlung und Wind mit geringer Geschwindigkeit die Abbauprozesse und damit die Alterung zu beschleunigen scheinen, führt eine hohe Luftfeuchte zur Verlangsamung der Alterung bis hin zu einer Absorption von Wasser der Umgebungsluft durch den Fingerabdruck und somit zu einer Verjüngung.

Die Analyse von 10 unterschiedlichen, häufig an Tatorten vorkommenden Oberflächen und drei verschiedenen beispielhaften Samples pro Oberfläche stand im Mittelpunkt der Arbeiten zu [16]. Hier konnte gezeigt werden, dass für die Oberflächen Glas, Möbel, Smartphone-Display, Autotür, Furnier und Festplattenplatter die logarithmische Alterungstendenz des Binary Pixel Merkmals reproduziert werden kann, während sich die Oberflächen Schere, Steckdosendeckel, Kupfermünze und CD-Hülle als herausfordernd darstellen. Als Ursache für die unterschiedliche Güte der Binary Pixel Alterungstendenz scheint die Grauwertverteilung zwischen Fingerabdruck- und Hintergrundpixeln im Histogramm verantwortlich zu sein. Während sich das Binary Pixel Merkmal besonders gut anwenden lässt, wenn die Grauwerte der Fingerabdruckpixel von denen der Hintergrundpixel klar trennbar sind (was eine zuverlässige Messung des Kontrastverlustes ermöglicht), stellen Fingerabdruckpixel mit ähnlichen Grauwerten wie die Hintergrundpixel eine Herausforderung dar. Eine Altersbestimmung ist in diesem Fall zwar prinzipiell möglich, allerdings können Schwankungen z.B. der Umwelteinflüsse hier leicht zu Veränderungen der Hintergrundpixel führen, welche die Binary Pixel Alterungstendenz überlagern und damit stören.

Unterschiedliche Scanparameter wurde in [13] und [18] untersucht. Dabei zeigte sich die Auflösung in Kombination mit der Größe der zu erhebenden Messfläche als wichtigstes Kriterium, während die Position der zu messenden Fläche innerhalb eines Fingerabdruckes zufällig gewählt werden kann. Weiterhin zeigten die Experimente, dass ein minimaler Punktabstand von 20 µm in Kombination mit einer Messfläche von 4 x 4 mm als untere Schranke für eine erfolgreiche Reproduktion des Binary Pixel Merkmals angesehen werden kann.

Unterschiedliche Aufbringungsfaktoren wie Anpressdruck, Anpressdauer, Verschmieren eines Fingerabdrucks sowie die Kontamination des Fingers mit Hautcreme und Speiseöl wurden in [19] untersucht. Dabei stellte sich heraus, dass Anpressdruck, -dauer und Verschmieren des Fingerabdrucks keinen über die natürliche Variabilität von Fingerspuren hinausgehenden Einfluss auf das Alterungsverhalten zu haben scheinen. Die Kontamination des Fingers zeigte, dass Wasser enthaltende Substanzen wie Hautcreme die Alterungsgeschwindigkeit einer Fingerspur deutlich erhöhen, während die Kontamination mit Speiseöl keine charakteristische Veränderung der Alterungsgeschwindigkeit hervorbringt. Anhand von Tropfenscans wurden diese Ergebnisse verifiziert, in dem eine sehr hohe Alterungsgeschwindigkeit für einen Wassertropfen gezeigt wurde, während Hautcreme als Emulsion von Wasser und Fett nur eine mittlere Geschwindigkeit besaß und Öl als weitgehend wasserfreie Substanz eine sehr gering Geschwindigkeit des Alterns zeigte.

AP5

Im Rahmen des AP5 wurden in [12] zwei grundlegende Klassifikationsstrategie diskutiert und verglichen. Dabei wurde ein formelbasierter Altersberechnungsansatz definiert, welcher aus drei wiederholten Messungen am Tatort (mit einem zeitlichen Abstand von $t \leq 1h$) das Alter einer Fingerspur potentiell berechnen kann. Aufgrund des logarithmischen Zusammenhangs führen hier jedoch kleinste Änderungen der Messwerte (im Bereich 10^{-3} bis 10^{-4}) bereits zu starken Verzerrungen des berechneten Alters. Ein solcher Ansatz scheint daher im Moment nicht umsetzbar.

Als weitere Strategie zur Altersabschätzung wurde in [12] und [18] ein maschinenlernbasiertes Verfahren untersucht, welches einen Fingerabdruck in die zwei Zeitklassen [0h, 5h] und [5h, 24h] unterteilt und mit einer ausreichend großen Menge von Fingerabdrücken trainiert wird. Dabei wurde eine Gesamtmenge von 770 Fingerabdrücken von 70 unterschiedlichen Testpersonen (50 Männer, 27 Frauen, Alter 21 - 73 Jahre) für die Evaluation aufgenommen und deren statistische Signifikanz gezeigt [18].

AP6

Die Variabilität des Alterungsverhaltens von Fingerspuren zwischen unterschiedlichen Personen (inter-person), innerhalb einer Person (intra-person), zwischen unterschiedlichen Fingern einer Person (inter-finger) sowie zwischen unterschiedlichen Regionen eines Fingerabdruck (intra-finger) wurde in [18] praktisch ermittelt und zeigt die größte Variabilität zwischen den Fingerabdrücken unterschiedlicher Personen sowie denen einer bestimmten Person zu unterschiedlichen Zeitpunkten. Die Variabilität ist jedoch für einen Innentatort gering genug, um eine Klassifikationsgenauigkeit von ca. 70 - 80% für das in AP5 beschriebene, statistisch signifikante Testset zu erzielen. Eine solche Klassifikation wurde bisher nur für die gut reflektierende Oberfläche eines Festplattenplatters analysiert.

AP7

In [12] wurde ein erster Fusionsansatz untersucht, welcher das Binary Pixel Alterungsmerkmal für 2D-Intensitätsdaten und 3D-Topographiedaten des CWL-Sensors kombiniert. Dabei wurde eine Fusion auf Sensorebene, Vorverarbeitungsebene, Merkmalsebene, Match-Score-Ebene und Klassifikationsebene durchgeführt. Eine Kombination unterschiedlicher Ausprägungen des Binary Pixel Merkmals für 2D-Intensitätsdaten, 3D-Topographiedaten sowie weiteren statistischen Merkmalen wie Durchschnitt, Standardabweichung, lokale Varianz, Gradienten, Rauheit und Kohärenz wurden in [18] analysiert. Alle Experimente kamen dabei zu dem übereinstimmenden Ergebnis, dass die Genauigkeit der Fusionsansätze keine signifikante Verbesserung im Vergleich zu den Einzelgenauigkeiten erzielen konnte. Grund für dieses Verhalten ist hauptsächlich in der Erfassungsdomäne zu sehen. Da alle bisher evaluierten Merk-

male auf Daten des CWL-Sensors beruhen, sind sie alle in hohem Maß voneinander abhängig. Zusätzliche Sensoren scheinen daher für die Untersuchung weiterer Merkmale als sinnvoll.

AP8

Die gewonnenen Erkenntnisse wurden umfassend dokumentiert und werden im Rahmen der Dissertationsschrift verwertet.

B. Beantwortete Forschungsfragen und nächste Schritte

Die im Rahmen des Dissertationsvorhabens zusammengestellten Arbeitspakete wurde für das Binary Pixel Merkmal als erstes von drei zu untersuchenden Merkmalen weitgehend durchlaufen. Weiterhin wurden die Rahmenbedingungen für eine erfolgreiche Evaluation weiterer Merkmale bzw. Sensoriken in Form einer grundlegenden Literaturrecherche sowie der Definition einer Mustererkennungskette und eines Vorgehensmodells zur Erstellung von Altersbestimmungsansätzen geschaffen. Die bisherigen Untersuchungen haben gezeigt, dass sich die gestellten Forschungsfragen zunächst für ausgewählte Merkmale mit zugehöriger Sensorik und Klassifikationsstrategie beantworten lassen. Mit der zunehmenden Evaluierung weiterer Merkmale und Sensoriken aber auch dem zunehmend besseren Verständnis der Relevanz unterschiedlicher Einflussfaktoren können die Ergebnisse im Laufe der Zeit iterativ weiter verbessert werden.

Für das Binary Pixel Merkmal basierend auf 2D-Intensitäts- und 3D-Topographiedaten eines CWL-Sensors lässt sich die Forschungsfrage nach den notwendigen Voraussetzungen zur Altersbestimmung mit der Möglichkeit der Erstellung von Zeitreihen, also der zerstörungsfreien und wiederholten Aufnahme von Fingerspuren am Tatort beantworten. Weitere Voraussetzungen bildet die minimale Scaneinstellung von 20 µm in Kombination mit einer Messfläche von 4 x 4 mm (untere Schranke), die Beschränkung auf Oberflächenmaterialien, welche eine gute Trennung der Fingerabdruck- und Hintergrundgrauwerte aufweisen sowie die Beschränkung auf einen Innentatort, um grobe Störeinflüsse von Temperatur, Luftfeuchte, Wind und UV-Strahlung zu minimieren. Die Einschränkung auf nur wenige Zeitklassen bei der Klassifikation ist eine weitere Limitierung.

Als zielführender Mustererkennungsansatz wurde das Binary Pixel Merkmal als Veränderung der Pixelwerte eines normalisierten und binarisierten Fingerabdruckbildes identifiziert. Die hohe Variabilität der Alterungskurven bzw. Alterungsgeschwindigkeiten, bedingt durch die vielen, in komplexen Beziehungen zueinander stehenden Einflussfaktoren, stellt eine deutliche Limitierung des Ansatzes dar. Die Klassifikationsgenauigkeit kann jedoch durch ein verbessertes Verständnis und eine verbesserte Berücksichtigung solcher Einflüsse sowie die Fusion von Merkmalen unterschiedlicher Sensorik iterativ gesteigert werden. Von den untersuchten Merkmalen hat sich das Binary Pixel Merkmal durch seine gute Performanz deutlich gegenüber anderen statistischen Merkmalen hervorgehoben [18]. Dabei erzielten die 2D-Intensitätsdaten bessere Ergebnisse als die 3D-Topographiedaten. Weitere Merkmale basierend auf unterschiedlichen Sensoren sind zu evaluieren.

Ein praktischer Altersabschätzer basierend auf dem Binary Pixel Merkmal und kontaktloser CWL-Sensorik kann für den in AP5 beschriebenen Zweiklassenansatz eine Klassifikationsgenauigkeit von ca. 70 - 80% erreichen. Dabei liegt die noch nicht optimierte Scanzeit im Bereich $0 < t \leq 10$h, kann jedoch durch Bestimmung der unteren Schranke sowie dem Einsatz eines Flächensensors noch deutlich reduziert werden. Der momentan evaluierte Messaufbau erlaubt weiterhin die parallele Untersuchung von bis zu 20 Fingerspuren. Die Rechenzeit zur Bestimmung einzelner Merkmalswerte liegt im Sekundenbereich und ist somit vernachlässigbar gering.

Mit Hilfe des Binary Pixel Merkmals konnte das Potential kontaktloser Sensorik zur Altersbestimmung latenter Fingerspuren gezeigt und anhand einer statistisch signifikanten Testmenge eine erste Genauigkeitsabschätzung vorgenommen werden. Weiterführende Arbeiten haben daher das Ziel, zusätzliche Sensoriken für die Altersbestimmung zu erschließen und neue Merkmale zu definieren, zu evaluieren und zu selektieren. Dabei stellt die vergleichsweise hohe Erfassungszeit von Zeitreihen eine große Herausforderung dar, da für neue Sensoriken auch neue Testsets von signifikanter Größe erstellt werden müssen. Potentiell zur Altersbestimmung geeignete Sensoren sind in [12] zusammengefasst. Wurden weitere charakteristische Kombinationen aus Sensorik und Alterungsmerkmalen identifiziert, können diese zu einem gemeinsamen Altersbestimmungsansatz fusioniert werden, um eine finale Abschätzung des Potentials kontaktloser Sensorik zur Altersbestimmung latenter Fingerspuren vorzunehmen.

ACKNOWLEDGMENT

Teile dieser Veröffentlichung entstanden aus dem Forschungsvorhaben „Digitale Fingerspuren (Digi-Dak)“ mit der Projektnummer FKZ:13N10818, welches vom Bundesministerium für Bildung und Forschung (BMBF) gefördert wird.

REFERENZEN

[1] Wertheim, K.: Fingerprint Age Determination: Is There Any Hope?, J. of Forensic Identification 53(1), pp.42-49, 2003.

[2] Baniuk, K.: Determination of Age of Fingerprints, Forensic Science International, (46) 1990, pp.133-137, 1990.

[3] Popa, G., Potorac, R., Preda, N.: Method for Fingerprints Age Determination, http://www.rjlm.ro/doc/127849931710-methodforfingerprints age determination.pdf, 15.05.2012.

[4] Ähnlich, J.: Altersbestimmung von daktyloskopischen Spuren mit Hilfe der Laser-Fluoreszenzspektroskopie, Diplomarbeit, Universität Hannover, 2001.

[5] Mong, M., Petersen, C. E., Clauss, T. R. W: Advanced fingerprint analysis project fingerprint constituents, Pacific Northwest National Laboratory, Richland, WA 99352, report PNNL-13019, Sept 19994.

[6] Wolstenholme, R., Bradshaw, R., Clench, M. R., Francese, S.: Study of latent fingermarks by matrix-assisted laser desorption/ionisation mass spectrometry imaging of endogenous lipids, Rapid Commun. Mass Spectrom. 23: 3031–3039, 2009.

[7] De Paoli, G., Lewis, S. A., Schuette, E. L., Lewis, L. A., Connatser, R. M., Farkas, T.: Photo and Thermal-Degradation Studies of Select Eccrine Fingerprint Constituents, Journal of Forensic Science Vol. 55, Issue 4, 962-969, 2010.

[8] Crane, N. C., Bartick, E. G., Schwartz Perlman, R., Huffman, S.: Infrared Spectroscopic Imaging for Noninvasive Detection of Latent Fingerprints, Journal of Forensic Science, Vol. 52, No. 1, 2007.

[9] Antoine, K. M., Mortazavi, S., Miller, A. D., Miller, L. M.: Chemical Differences Are Observed in Children's Versus Adults' Latent Fingerprints as a Function of Time, Journal of Forensic Science Vol. 55, No. 2, 2010.

[10] Williams, D. K., Brown, C. J., Bruker, J.: Characterization of children's latent fingerprint residues by infrared microspectroscopy: Forensic implications, Forensic Science International Vol. 206, Is. 1, pp. 161-165, 2011.

[11] Fries Research Technology, http://www.frt-gmbh.com/en/, 15.02.2012.

[12] Merkel, R., Gruhn, S., Dittmann, J., Vielhauer, C., Bräutigam, A.: General fusion approaches for the age determination of latent fingerprint traces: results for 2D and 3D binary pixel feature fusion, Proc. SPIE 8290, 82900Y, 2012.

[13] Merkel, R., Krapyvskyy, A., Leich, M., Dittmann, J., Vielhauer, C.: A first framework for the development of age determination schemes for latent biometric fingerprint traces using a chromatic white light (CWL) sensor, Proceedings of SPIE Security + Defence 2011, 2011.

[14] Merkel, R., Dittmann, J.: Resolution and Size of Measured Area Influences on the Short- and Long-Term Aging of Latent Fingerprint Traces Using the Binary Pixel Feature and a High-Resolution Non-Invasive Chromatic White Light (CWL) Sensor, Proceedings of IEEE ISPA 2011, 2011.

[15] Merkel, R., Breuhan, A., Hildebrandt, M., Vielhauer, C., Bräutigam, A.: Environmental impact to multimedia systems on the example of fingerprint traces aging behavior at crime scenes, Proceedings of SPIE Photonics Europe, 2012.

[16] Merkel, R., Bräutigam, A., n Kraetzer, C., Dittmann, J., Vielhauer, C.: Evaluation of binary pixel aging curves of latent fingerprint traces for different surfaces using a chromatic white light (CWL) sensor, Proceedings of the Thirteenth ACM Multimedia Workshop on Multimedia and Security. pp. 41-50, DOI: 10.1145/2037252.2037262, Niagara Falls, NY, USA, Sept. 29-30, Publish-er: ACM, New York, NY, USA, 2011.

[17] Merkel, R., Dittmann, J., Vielhauer, C.: Approximation of a Mathematical Aging Function for Latent Fingerprint Traces Based on First Experiments Using a Chromatic White Light (CWL) Sensor and the Binary Pixel Aging Feature, In B. de Decker et al. (Eds.): CMS 2011, LNCS 7025, pp.59-71, IFIP International Federation for Information Processing, 2011.

[18] Merkel, R., Gruhn, S., Dittmann, J., Vielhauer, C., Bräutigam, A.: On non-invasive 2D and 3D Chromatic White Light image sensors for age determination of latent fingerprints, Forensic Sci. Int. (2012), http://dx.doi.org/10.1016/j.forsciint.2012.05.001, 2012.

[19] Merkel, R., Dittmann, J., Vielhauer, C.: How Contact Pressure, Contact Time, Smearing and Oil/Skin Lotion Influence the Aging of Latent Fingerprint Traces: First Results for the Binary Pixel Feature using a CWL Sensor, Proceedings of IEEE Intl. Workshop on Information Forensics and Security - WIFS'11, Foz do Iguaçu, Brazil, 2011.

[20] Hildebrandt, M., Dittmann, J., Pocs, M., Ulrich, M., Merkel, R., Fries, T.: Privacy preserving challenges: New Design Aspects for Latent Fingerprint Detection Systems with contact-less Sensors for Preventive Applications in Airport Luggage Handling, In C. Vielhauer et al. (Eds): BioID 2011, LNCS 6583, pp. 286-298, Springer-Verlag Berlin, 2011.

User Interfaces for Exploratory Search

Towards generalized design patterns for complex information retrieval tasks

Marcus Nitsche
Data and Knowledge Engineering Group
Faculty of Computer Science, Otto-von-Guericke-University
Magdeburg, Germany
marcus.nitsche@ovgu.de

Abstract—**While ad-hoc searches are well supported by current search engines, complex search tasks are not. There exist nearly no tools that match users' needs in managing search results, when they are trying to satisfy complex information needs, in a way that users are able to search for and filter information in domains they might be not familiar with. Exploratory search is a promising interaction paradigm that tries to tackle these problems. While information retrieval techniques in general are quite good developed, exploratory search systems often lack in ergonomically designed user interfaces, e.g. they do not provide easy-to-use interactions for dynamic query reformulation, do not allow contextual change of user's perspective and do not provide adequate result overviews. Users engaged in an exploratory search need to search sequentially or/and perform parallel searches, and often switch between sub-searches and different modes. This activity strains a user's working memory capacities and increases her or his workload. This PhD thesis aims to address the ergonomic design of user interfaces and user experience for exploratory search tasks by providing generalized design patterns for the implementation of such applications.**

User Interface Design, Exploratory Search, Design Pattern, Web Search, Information Retrieval

I. INTRODUCTION

In 1945 Vannevar Bush envisioned the *MEMEX* system [1], a theoretical machine that empowers its user to read documents of a large, electronic library. Bush's concept covered the option to follow associative information links that users can create or follow. Among other breakthrough innovations, which will be presented in more detail in a thesis subchapter, Bush's theoretical concept lead to the development of the world's biggest network – the Internet – and it's most important service: the World Wide Web (WWW). The problem how to access a huge information space like the WWW is still relevant to research as it was back in 1945.

While digital information is permanently growing and information density is constantly rising (information overload, e.g. [2]), the fraction of *relevant information* is still hard to identify. The main idea of Web 2.0 even boost this problem: Users are not longer only consumers, they become producers of information. Keeping a minimum of quality becomes harder in times of easy content producing online communities, blogs and wikis. Search engines offer users just rare opportunities to filter these growing and complex information spaces.

Often users just like to have a quick answer for a simple structured query. Those "ad-hoc"- or "lookup"-searches are sufficiently supported by existing (web) search systems. But these systems do not provide sufficient support for complex research and investigative searches. In an increasing number of cases users are interested in finding answers on complex information retrieval tasks [3]. Users often do not know exactly what they are searching for, how they should formulate a query and - while reviewing search results - users start to rethink their initial query and like to reformulate it since they have learned something new about (e.g.) an unknown topic. In these cases users need to be supported in an *exploratory search* process [4].

II. RESEARCH GOALS AND QUESTIONS

Providing users with an ergonomic user interface, which supports them in querying and reviewing search results is a strongly required and challenging research task. According to a research study of 2005[1] at least 90% of all Internet users (ca. 2.06 Billion people) are using search engines. That is why the development of ergonomic user interfaces for exploratory search scenarios becomes more and more relevant to support users in finding relevant information.

The general goal of this thesis is to identify user interface (UI) and user experience (UX) design patterns for the development of exploratory search interfaces (XSI). These generalized design patterns aim to support users in complex information retrieval scenarios. Therefore this PhD thesis tackle the problem of finding answers to the following research questions:

Question 1: ***How can users be better supported in ad-hoc search scenarios?***

What do users require when conducting searches? What are the current trade-offs they need to accept? What techniques are applicable that support users in managing search results? How to support users in mobile search scenarios? How to prevent users from unintended context and mode switches?

Question 2: ***What are special requirements in exploratory search scenarios?***

[1] http://www.pewinternet.org/Reports/2005/How-Women-and-Men-Use-the-Internet.aspx (30.05.2012)

What is exploratory search? Are the existing definitions suitable? What are complex information retrieval tasks in this context? What are realistic search scenarios for exploratory search tasks? Why are these scenarios not sufficiently supported by user interfaces of existing search engines? Which functionality is needed to support exploratory search?

Question 3: ***Do patterns exist that describe general building blocks for designing exploratory search user interfaces?***

What are UID and UxD patterns? Can evaluated user interface paradigms and techniques be generalized in order to create building blocks for designing XSIs?

To provide a structured foundation to answer these research questions the following cases are made. The first three statements picture the *characteristics* of exploratory search:

Thesis 1: ***Exploration is not a stand-alone information access method****. Aspects of exploration can enhance information access methods like browsing, navigation and search: So that* ***there exist methods like exploratory browsing, exploratory navigation and exploratory search.***

Explanation: This statement targets the orthogonal character of the term "exploration". Exploration, used as a stand-alone information access method, is often explained as a process that is a combination of browsing, navigation and search. Therefore it should rather be seen as an extension of basic information access methods than as a single one.

Example: Planning a journey is one example for an *exploratory search* task. It basically includes common search processes. Additionally, aspects of exploration come into play. Repeating the same task later still follows exploratory goals, but is more characterized by following a familiar link structure and therefore equals *exploratory navigation*. Searching in familiar information spaces for new insights can be widely compared to *exploratory browsing* (e.g. browsing the own music collection).

Thesis 2: *Currently existing web search user interfaces (e.g. Google Search[2]) do* ***not match the requirements of complex exploratory search tasks****.*

Explanation: This claim aims at missing support in building up and managing search trails, missing support in saving intermediate results, lack of providing overviews and further important aspects that would support exploratory search tasks. Since UID and UxD are key issues and central aspects in supporting users efficiently, it is crucial that current user interfaces of Web search engines are neither effective, nor efficient and not satisfying for a user. Therefore they do not fit the ISO, EN and DIN standard norm definitions of ergonomic systems [5].

Example: Finding new music that match one's personal interest and taste by using a standard search engine like Google can be a frustrating and time consuming task. Google for example does not support users in gaining an overview about possible places. In addition single results are not interconnected to each other, even if implicit or explicit cross-links exist.

Thesis 3: *User interface elements for exploratory search* ***can be classified and generalized toward abstracted design patterns****.*

Explanation: While identifying patterns is a common method in computer science to ensure abstractions for reuse and cost reduction - e.g. [6] - this method has been rarely used in human-computer interaction. One reason for this is that designing an ergonomic user interface is usually depending on many constraints that need to be taken into account (user's task, application environment, characteristics of target group, technological limitations, etc.). Therefore it is not clear, whether it is possible to create generalized user interface patterns that meet the requirements of exploratory search tasks.

Example: Possible UI patterns for XSI might describe interface element solutions for query (re-) formulation, reviewing results (overviews and single entries), methods for changing the perspective, management of search trails, saving of intermediate search results.

After presenting characterizing claims of exploratory search systems, next some *UI and UX related* cases are made:

Thesis 4: ***Direct manipulations of work items*** *like (sub) queries and document representations support users in understanding interconnections between those objects.*

Explanation: By direct interaction users express their intention. But while direct manipulation of user interface elements is an often-applied aspect in graphical user interfaces (GUI); it is still not common to directly manipulate work items like (sub) queries and document representations in research fields like information retrieval (IR). The claim addresses the (implicit) goal of a user to learn about the structure of an information space without directly manipulating its content. It is comparable to changing a perspective.

Example: By directly manipulating the position of document representations in result visualizations, users might recognize immediate effects on other parts of the document collection. By this users have the ability to learn more about the inner structure of this information net. Thus, the rephrasing of queries respectively the filtering of data spaces becomes an easier task for them. E.g., marking a search result as "unsuitable" may lead to fading out further results that are similar. Often users will recognize a change in ranking order, position, size or general appearance of other results.

Thesis 5: *The appropriation of* ***mechanisms for managing intermediate search results*** *improves the usability of an information retrieval system.*

Explanation: Often long-term searches like exploratory tasks need to be supported by offering some kind of external memory [7]. Ideally, these mechanisms for saving and managing intermediate results are embedded and strongly connected to the user interface of the IR system. Using

[2] http://www.google.com

external tools often leads to mode switches and work context changes that lower cognitive resources of users [8, 9]. Usually IR systems do not support users by an embedded solution for managing intermediate search results. I.e., today interesting search results need to be bookmarked, downloaded, printed or transformed in other formats to compare them to each other and to gain an overview about a certain topic.

Example: Using external personal information management (PIM) tools lead to higher cognitive workloads and decreases user's attention on the actual task [10]. E.g., saving homepages of suitable places for spending one's holidays to compare them later on is usually not possible to do while parallel browsing through the result list.

Thesis 6: *Providing transparency and traceability of adaptive exploration methods by* ***choosing dynamic visualization techniques*** *invigorate user's trust in IR systems.*

Explanation: Dynamic visualizations adapt oneself to user's interaction and provide different views on the same data set. Adapting visualizations dynamically towards user's personal needs ensure users to be in control of the interactive system; even if adaptive methods are used. Since cognitive processes during investigative search tasks should be intuitively supported by the interaction flow, dynamic visualization support user's trust in the reliability of the IR system and its delivered results.

Example: While monitoring packages in a logistics application, users can choose from different visualizations to have different views on the same data. For example a user takes a look at a map, a geographically referenced point of view. The user might see that the payloads have been delivered successfully. But if he takes a look at a time-referenced timeline-visualization he might notice that single payloads have been delivered too late. This illustrates the importance of providing dynamic views on the same data in exploratory searches.

Thesis 7: *In exploratory search tasks it is more important to support users with* ***reliable navigation and orientation concepts*** *than providing them with liberty of action.*

Explanation: This statement addresses mainly the challenge of result visualization. Often classic IR systems provide users with a result list. Exploratory search interfaces on the other hand often use graph-based visualizations or maps, e.g., [11, 12]. Even if graphs provide ergonomic overview visualizations, users often feel lost in those information nets. The claim predicates that result visualizations like lists or tables provide more reliable navigation and orientation support than graph layouts.

Example: One example is the fixed alignment structure a list or table has. Users can rely to a certain degree that a specific search result will be found at a rank respective position; even if they repeat the same search some days after. Depending on the chosen layout algorithm a graph-based visualization might place this certain result at a different position. In addition to this the starting point for a user to investigate the result space is fixed in lists and tables: usually on top. Also the further navigation is linear. This supports more efficiently users in finding answers to the basic questions in orientation and navigation: Where do I am? Where do I come from? Where can I go next? [13]

If these theses would be proven by corresponding findings, this would mean for computer science and especially for the research on exploratory search that the development of XSIs could be improved to create more effective, efficient and satisfying to use search systems. If the thesis succeeds in its goal to develop general UI & UX design patterns for exploratory search, the development of exploratory search interfaces could even be partly done automatically. Finally this saves time, money and ensures health of users and developers of such systems.

III. Vision: Potential Users, Scenarios and Use Cases

Potential users of exploratory search interfaces are all web users that search for information that cannot be found obviously in one (of many) retrievable document(s). Exploratory search includes in most cases the comparing and / or combination of at least two documents containing required information pieces. Furthermore users of desktop searches and huge information systems like enterprise resources planning (ERP) software need to be supported in those complex information retrieval processes. Other potential user groups are lawyers and librarians that - for example - search for patents [14]. Also a casual exploratory search scenario like music retrieval is an often-mentioned field of application, e.g. [15]. For example, a user does not know anything about classical music, but likes to find music of this genre that he likes. The goal group is versatile. Characteristics that they all share are a vague understanding of an appropriate answer to their information need, therefore a missing understanding of how to formulate a query and the awareness that the search might take more time than a usual web search.

Typical scenarios are travel-planning scenarios, research on specific topics (e.g., figuring out the patent status of a specific idea) and searches that aim to fulfil emotional needs (e.g., finding likeable music). Possible **Use Cases** that are well addressed by the characteristics of an exploratory search are for instance:

(1) Given a user, who is interested in learning more about the current economic crisis. Since the user heard of a possible connection, he likes to find out what connection exists between the financial situation of today's *Greece* and the ending of the *2nd World War*. Instead of searching for the two terms separately one after the other, the user uses an exploratory search system, where he is able to formulate the specific aspects he is interested in. The user, since he does not know much about this specific topic, is mainly interested in getting an overview about it. After this he is interested in reviewing cross connections between entries. Furthermore he likes to mark potential interesting sources to pull them together. Especially he is not interested in using several tools for resolving this task.

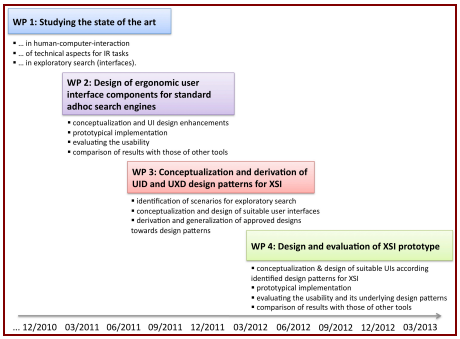


Figure 1. Timetable of the PhD thesis project.

(2) Another possible scenario might be a worker who comes up with a concept to improve a technical device. He likes to protect his rights on this idea and therefore wants to patent it. Before writing a patent application he needs to clarify whether this idea is really new or not. Furthermore he is interested in similar patents and patent applications to classify his own invention appropriately. For this complex task he is using a patent database. An ideal XSI for this task could support him in offering cross-references to similar patents. While getting an overview about the topic he learns more about the structure of patent categories and applied patents in his field of work.

IV. Work Packages and Research Methodologies

The following four key work packages (WP) have been derived from the previous mentioned theses and the planned thesis structure. See Fig. 1 for a timetable, which covers these main work packages and Section VI for more details on the planned thesis structure.

WP 1: *Studying the state of the art ...*

Task 1.1: *... in human-computer-interaction.*
Task 1.2: *... of the design of information retrieval systems.*
Task 1.3: *... in exploratory search (interfaces).*

Milestone M1: Basic overview of related work created. Future directions of state of the art identified and focused.

This work package covers the analysis of thesis relevant research fields to present selected related work and to put this PhD thesis in context.

WP 2: *Design of ergonomic user interface components for standard ad-hoc search engines (e.g. integrated tools for search result management and query refinement).*

Task 2.1: *Conceptualization and design of suitable user interface enhancements.*
Task 2.2: *Prototypical implementation.*
Task 2.3: *Evaluating the usability.*
Task 2.4: *Comparison of results with those of other tools.*

Milestone M2: Understanding the needs of users that search for information. Successfully transfer research methods from HCI to IR. Improvements in usability aspects are verified.

In the second work package the aim is to conceptualize and develop re-designs and enhancements concerning standard ad-hoc search support. This leads to a better understanding of the limits of standard IR systems and motivates the research of XSIs from a practical perspective.

WP 3: *Conceptualization and derivation of UI and UX design patterns for XSI.*

Task 3.1: *Identification of scenarios for exploratory search.*
Task 3.2: *Conceptualization and design of suitable UIs.*
Task 3.3: *Derivation and generalization of approved designs towards design patterns.*

Milestone M3: Set of user interface and user experience design patterns for the development of exploratory search user interfaces identified.

To support exploratory search systems appropriately you first need to understand typical scenarios of exploratory search. What are good concepts and designs of XSIs? Finally the third work package covers the identification of abstracted concepts for the XSI design, so called generalized design patterns.

WP 4: *Design and evaluation of an XSI prototype.*

Task 4.1: *Conceptualization and design of suitable user interfaces according to identified design patterns for exploratory search user interfaces (see WP 3).*
Task 4.2: *Prototypical implementation.*
Task 4.3: *Evaluating the usability and, by this, its underlying design patterns.*
Task 4.4: *Comparison of results with those of other tools.*

Milestone M4: Successful evaluation of a practical XSI implementation, based on theoretical UID and UxD patterns.

The last work package aims to proof the previously identified design patterns by applying them during the development of an exploratory search prototype. This prototype needs to be tested by user studies and the results will be compared to other tools that might be applicable for solving the same user tasks.

Applied **research methodologies** include, but are not limited to, user centred design, rapid prototyping, end user involvement, user studies, within- and between-subject design experiments, formative and summative evaluation methods like (paper) prototype testing, questionnaires, lab experiments and user observations. The **achievement of the research goals** and **validation** of realized systems will be checked towards usability study results and theoretical & practical comparative studies between **realized tools** and common used systems, like major web search engines. The **quality** will be checked permanently during the development since a formative evaluation design is chosen. Since the identified user group is quite common and widespread it might be not that hard to find **test persons** for user studies. If certain fields of application lend themselves to **cooperate with experts** of that special domain, this possibility will be used whenever possible. Used **programming languages** include Java, C++,

Objective C, Adobe Flex, HTML 5, JavaScript and CSS. It is the overall goal to realize prototypes that are platform independent. Therefore web-based applications, written in HTML 5, are favoured. For software modelling UML diagrams will be applied.

V. Exploratory Search in Related Work

While a lot of concepts and systems of the past are focusing on good system performance measurements, user needs and capabilities often have been left out. One problem for example is the trade-off between the technical need for fine configuration of an IR system on the one hand and the lack of terminological understanding of users on the other hand. So this work is located at the intersection of multiple research fields such as information science, information retrieval (IR), information management, human computer interaction (HCI), information visualization, design and psychology.

One object of this thesis is also the transfer of methods from these disciplines to use benefits of research not only from computer science. Each of these disciplines get into trouble when explaining user's difficulties in doing investigative research with standard search engines. But their combination might lead to ergonomic solutions. This idea of constructing bridges between those disciplines motivates several CASE studies. In the following the state of the art of the three main research fields will be discussed in more detail.

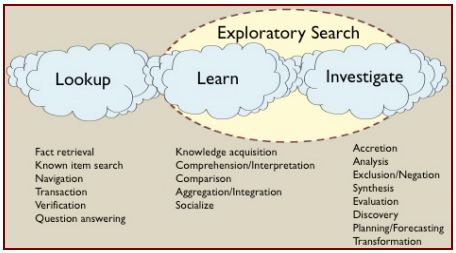


Figure 2. Exploratory search as an enhancement of standard IR activities [4].

Among other definitions of "exploratory search", which will be discussed in more detail in the final thesis, the following definitions represent three different views on this search concept. While the first one addresses the problem by figuring out possible starting points, the second categorizes various information access methods in different categories:

Definition 1: "Exploratory Search", according to White et al. [16], is a kind of information exploration. Furthermore it is the representation of searcher activities, which are either:

a) unfamiliar with the domain of information surrounding their need
b) unsure about the ways to achieve their goals (either the technology or the process)
c) or even unsure of their goals in the first place.

They further state that the process covers a broader class of activities that goes beyond typical information retrieval, such as investigating, evaluating, comparing, and synthesizing. Therefore, users use a combination of querying and browsing strategies to foster learning and investigation. In the past, exploratory search scenarios were often motivated by questions like "what if users do not know how to formulate a question?" or "what if a user is more interested in getting an overview about a certain topic than a specific answer?".

Definition 2: "Exploratory Search" differentiates exploratory search activities from simple lookup activities by positioning exploratory search as an enhancement of standard IR activities [4], see Fig. 2.

"Exploratory search can be used to describe an information-seeking problem context that is open-ended, persistent, and multifaceted; and to describe information seeking processes that are opportunistic, iterative, and multi-tactical. In the first sense, exploratory search is commonly used in scientific discovery, learning, and decision-making contexts. In the second sense, exploratory tactics are used in all manner of information seeking and reflect seeker preferences and experience as much as the goal." [4]

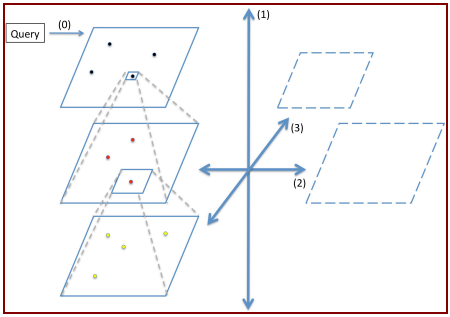


Figure 3. Dimensions of exploratory search activities according Laurence Noël, adapted based on [17].

One can state that definition 1 is a *problem-oriented* and definition 2 a *method oriented* approach to formally describe exploratory search. Next to these two *declarative* definitions, Noël presents in [17] a *constructive* way to describe what exploratory search is and how it is bringing together different dimensions of interaction and information access methods: We assume that there might be a query input (0), She further states their exist three layers of interaction in an information space like presented in Fig. 3. (1) is the vertical axis for filtering (zoom In & zoom Out) operations. (2) shows the horizontal axis that symbolizes similar items and neighbourhood relationships. The third axis (3) is called transversal axis and offer ways into other domains. The author states that these are the basic degrees of freedom a user has when navigating, browsing or searching a data space exploratory.

Often the term “exploratory search” get mixed with terms like “discovery” and “exploration”, so you can find a lot of overlapping definitions. Furthermore important aspects like “context of use”, “personal and dynamic views”, “serendipity”, possible user-driven “change of perspective” and even the word “search” in sense of “querying” are not taken into account. To clarify these overlapping concepts for this PhD thesis project, I decided to set up an own definition:

Exploratory search is a highly dynamic (learning) process in which a user accesses an information space to get an overview about a topic, or to gain a vague understanding, or to get an answer towards complex information needs.

Thereby the process expands a simple lookup by using techniques of exploration. Furthermore users usually look at a (sub) set of information through a specific view angle, which might change during the investigation process. This personal and dynamic view – more generally known as the context of use – allows changes of perspective in order to (re) formulate or to refine an initial query. User’s overall goals are to learn (about content and structure of an information net), to investigate, to understand, or to conceptualize (about) their initial information need by building up a personal mental map or model. Thereby the acts of “searching”, “browsing” and “exploring” are often more important than the actual find. Success in this context does not necessarily mean to find certain information.

The evaluation of XS-Systems is difficult. Since exploratory search tasks and goals are often undefined and unpredictable it is difficult to evaluate those systems with typical IR measurements. Measuring *accuracy*, for example, is not applicable, since *correct* answers cannot be identified if *summarizing a domain*. Since *researching* a topic is a common use case, *time efficiency* should not be a goal of designing exploratory search systems. Furthermore the setting is hard to describe to study participants since giving them well defined tasks that can immediately prevent them from exhibiting exploratory behaviour.

VI. Core Concept and Thesis Structure

The concept is reflected by the chosen thesis structure. The chapters are conceptualized as slightly overlapping sections. First a problem motivation and general discussion of the topic will be given. Then interaction methods for accessing digital information (2.) in general will be examined to understand the fundamental differences between them. Also the basic requirements for ergonomic support of exploratory search are supposed to be identified in this chapter. After a discussion and presentation of possible solutions towards the problem of supporting users in standard ad-hoc search scenarios (3.) a chapter about search result management (4.) examines possibilities that could be used to manage user’s personal information appropriately. In chapter 5 the dynamic navigation through information spaces will be discussed. This is a main enhancement of classic web search since here users will be actively supported in navigating through their fields of interest. Each of these main chapters is divided into a theoretical part and a practical one, where use cases show the applicability of the theoretical base. Here the user-centred development and evaluation of prototypes are described. In chapter 6 UID and UX patterns, which derive from previous usability testing, are discussed. These theoretical building blocks are supposed to be used in chapter 7, where the development and evaluation of a prototype for exploratory search are presented. In final chapter 8 the research findings and future work will be discussed.

After a close look to the state-of-the-art in these research fields, developed concepts respectively systems and common methods were examined to proof the possibility of combining concepts of various disciplines. Especially HCI methods are taken into account for improving the usability of IR systems. Selected case studies and applications will be considered to proof concepts of ergonomic and usable interfaces that combine iterative filtering, ranking and exploration methods for standard desktop interfaces and mobile devices like tabletops or mobile phones. Based on the experience of conducted user studies with the proposed prototype systems, a summative evaluation need to be conducted to prove the combination of several techniques. An outlook and a discussion of open research questions will close the thesis.

VII. Prototyping

Preliminary work includes the development of prototypes of mobile user interfaces for tabletops (Fig. 7, [24]) and mobile phones (Fig. 4, [19]), desktop applications for ad-hoc / known item search scenarios (Fig. 6), control room applications for context dependable searching, filtering and exploration of information items from a logistics scenario (Fig. 5, [21]), alternative search result visualization and studies about annotation of screen objects [26], which might be helpful for managing search results. Furthermore algorithms for graph layout methods were examined in [25], possible platforms for multi-user support were compared to each other in terms of information security [22] and context-aware retrieval systems in fields of cross-lingual text retrieval were studied [23] to build in context support in XSIs.

Recently some work on PIM in context of user’s awareness in changing hierarchical information structures (semi-) automatically [20] and theoretical considerations about the term exploration [18] were published.

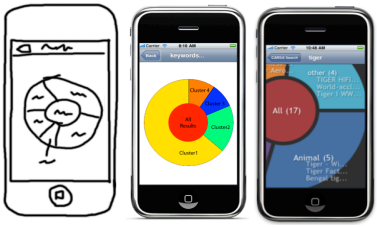


Figure 4. Prototype of a user interface for interactive filtering search results on a mobile device, screenshot derived from [29].

In the following some prototypes will be presented shortly to demonstrate the consideration of different aspects of the design of exploratory search user interfaces:

Fig. 4 presents the design steps from a rough sketch, over a paper mock-up towards a running prototype of a mobile search interface, which supports categories, two-stage semantic zoom and does provide an integrated web browser to prevent users from context switches.

In Fig. 5 a screenshot from an application for logistics is shown. The left picture shows a list of deliveries with some detailed information on the content, followed by a filter area for dragging screen items to build up a complex filter for setting constraints for a search query towards packages with specific characteristics.

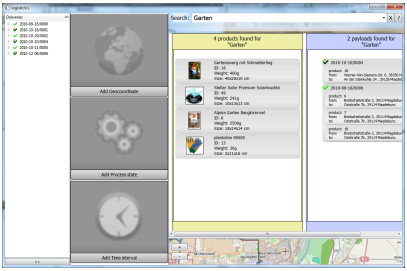


Figure 5. Interactive retrieval, filtering and exploration of information items in a logistics scenario, see also [21].

The redesign of a desktop-based interface for an information retrieval framework, called CARSA, is presented in Fig. 6. Here users are supported in reviewing results from a web search query by visual feedback. But contrary to other solutions, which use techniques to create thumbnails of web pages, the content of retrieved web documents become browsable & explorable in our solution when enlarging them.

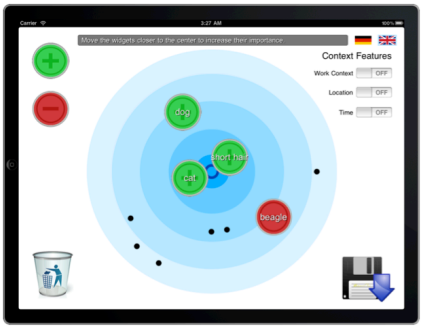


Figure 6. Tablet based user interface for interactive search and filtering, screenshot derived from [30].

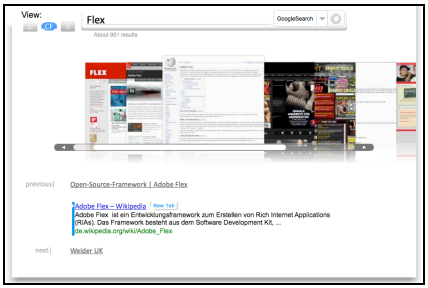


Figure 7. Adobe Flex based user interface for ad-hoc search scenarios.

This "in-frame-browsing", which (1) prevents users from unnecessary context switches (no need for site switch, new tab or a second program to navigate to the selected source of interest), (2) the decision, whether a specific search result is of value for the user can be made much more easier and reliable and (3) recognizing content seen before is also easier by a visual representation of a website.

VIII. Preview: On Generalized UID and UxD Patterns

Although this is the main task of the later thesis, I like to give a short preview on generalized *user interface and user experience design patterns* in exploratory search tasks:

Trivial *UID patterns* in a web search UI are the text-based *query input field* and the *search button*. Both have different characteristics concerning edit ability, degree of efficiency or their general way of use. Furthermore, some kind of logic between these elements exists. While an entry field can be used without necessarily using a search button (a.k.a. *InstantSearch*, [27]), a search button fulfils its functionality only be previously receiving any kind of input. Next to dependencies between *user's task*, current *information access strategy or interaction paradigm* and the *selection of ergonomic UI components*, also *technical aspects* need to be taken into account when trying to realize a certain concept. If, for example, complex pre-processing steps like a re-ranking of retrieved results are necessary, *InstantSearch* might be not an adequate solution since the response time might be extended and this is used to be seen as non-ergonomically.

Generalized *UxD patterns* can be derived out of the application and interaction of UI patterns. A certain UxD pattern might be *searching for fun*, like used in casual search [15]. This specific *user experience* influences what users think of the interface. While UX is usually measured *while* or *after* using an interactive system [28], here the idea is to proof whether there exist patterns to create a certain experience *before using* an XSI.

IX. Conclusion

The problem of finding information in huge data spaces like the Internet is presented in I. Since users often do not know

how to formulate a query and are also often unfamiliar with the domain they search in, the research field of *exploratory search* came up. The goal of this thesis is to identify generalized user interface and user experience design patterns for these kind of complex information retrieval tasks to facilitate the development of *exploratory search user interfaces (XSIs)*. To achieve this goal several prototypes are and will be created to study different aspects of exploratory search (e.g. to tackle the problem of query formulation [29]). The quality of those prototypes has been and will be measured by conducting usability studies (see also *IV – Work Packages and Research Methodologies*). Here also an overview about the project and the current status has been provided. Next steps will be the derivation of *UI and UX design patterns* and their application during the development of an ***integrated Exploratory Search Interface Prototype***, which will be evaluated in a formative way to ensure a high degree of usability.

The potential scientific and practical benefit of this thesis has been illustrated in *III – Vision: Potential Users, Scenarios and Use Cases*. Especially when considering a potential user group of 2.06 Billion users, the practical benefit becomes clear.

The mentioned prototypes are conceptualized taking into account previous work in this field (see references). Since this is a quite new field of study there has been conducted just a few studies in the intersection of *exploratory search* and *UID*.

ACKNOWLEDGMENT

I would like to give Prof. Dr. Andreas Nürnberger props for his support and the supervising of this PhD project. I thank my friendly colleagues of the Data and Knowledge Engineering Group for their helpful advices and I am also obliged to supervise so many talented students over the past four years. Sincere thanks are given to them all. The work presented here was partly supported by the German Ministry of Education and Science (BMBF) within the ViERforES II project, contract no. 01IM10002B.

REFERENCES

[1] Bush, V.: As We May Think. The Atlantic Monthly, 176, 1, pp. 641-649, 1945.
[2] Wurman, R. S.: Information Anxiety. New York: Doubleday, 1989.
[3] White, R. W. and Roth, R. A.: Exploratory search: Beyond the Query-Response paradigm. In: Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers, 2009.
[4] Marchionini, G.: Exploratory search: from finding to understanding. Commun. ACM 49, 4, pp. 41-46, 2006.
[5] International Standards Organization. ISO 9241: Ergonomics of human-system interaction, 2008.
[6] Beck, K.: Implementation Patterns. Addison-Wesley, 2008.
[7] Van Kleek, M.: Effort, memory, attention and time: paths to more effective personal information management. Ph.D.-Thesis, MIT, 2011.
[8] Johnson, A., Proctor, R. W.: Attention: Theory and Practice. Sage Publication, Thousand Oaks, 2004.
[9] Zijlstra, F. R. H.: Efficiency in Work Behaviour: a Design Approach for Modern Tools. Delft: Delft University Press, 1993.
[10] Tungare, M.: Mental Workload in Personal Information Management: Understanding PIM Practices Across Multiple Devices. PhD thesis, Virginia Tech, 2009.
[11] Goldenberg, S.: Exploratory Search in WorkTop. Master-Thesis, Brown-University, Providence, Rhode Island, USA, 2012.
[12] Dörk, M., Carpendale, S., Williamson, C.: Fluid Views: A Zoomable Search Environment. In: Proceedings of the International Conference on Advanced Visual Interfaces (AVI), ACM, 2012.
[13] Nielsen, J.: Usability Engineering. Morgan Kaufmann, San Francisco, 1993.
[14] Wilson, M. L., Kules, B., schraefel, m.c., Shneiderman, B.: From Keyword Search to Exploration: Designing Future Search Interfaces for the Web. Found. Trends Web Sci. 2, 1, pp. 1-97, 2010.
[15] Wilson, M.L. and Elsweiler, D.: Casual-leisure Searching: the Exploratory Search scenarios that break our current models. In Proc. HCIR'10, New Brunswick, NJ, USA, pp. 28-31, 2010.
[16] White, R.W., Kules, B., Drucker, S.M., and schraefel, m.c.: Supporting Exploratory Search, Introduction to Special Section of Communications of the ACM, Vol. 49, Issue 4, pp. 36–39, 2006.
[17] Noël, L.: From semantic web data to inform-action: a means to an end. In: Proc. of *Semantic Web User Interaction at CHI 2008 Exploring HCI Challenges,* Ed. Duane Degler, 543, pp. 1-7, 2008.
[18] Gossen, T., Nitsche, M., Haun, S., Nürnberger, A.: Data Exploration for Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications. In: Bisociative Knowledge Discovery, Springer Lecture Notes in Computer Science Volume 7250/2012, pp. 287-300, 2012.
[19] Nitsche, M., Nürnberger, A., Bade, K.: An ergonomic user interface supporting information search and organization on a mobile device. In: Personal information management in a socially networked world, 2012.
[20] Bade, K., Nitsche, M., Nürnberger, A.: Effective data mining support for personal information management. In: Personal information management in a socially networked world, 2012.
[21] Heydekorn, J., Nitsche, M., Dachselt, R., Nürnberger, A.: On the interactive visualization of a logistics scenario - requirements and possible solutions. In: IWDE 2011: proceedings of the 2nd International Workshop on Digital Engineering 2011, Magdeburg, 2011.
[22] Nitsche, M., Dittmann, J., Nürnberger, A., Vielhauer, C., Buchholz, R.: Security-relevant challenges of selected systems for multi-user interaction. In: Workshop on Adaptive multimedia retrieval, Berlin [u.a.], Springer, pp. 124-134; LNCS; 6535, 2011.
[23] Ahmed, F. A., Nürnberger, A., Nitsche, M.: Supporting arabic cross-lingual retrieval using contextual information. In: Multidisciplinary information retrieval, Heidelberg [u.a.], Springer, pp. 30-45, Lecture Notes in Computer Science; 6653, IRFC, 2 (Vienna), 2011.
[24] Nitsche, M., Nürnberger, A.: Supporting vague query formulation by using visual filtering. In: LWA 2011, Magdeburg, 2011.
[25] Haun, S., Nitsche, M., Nürnberger, A.: Interactive visualization of continuous node features in graphs. In: Workshop on Explorative Analytics of Information Networks at ECML PKDD 2009, EIN 2009, Bled, pp. 98-106, 2009.
[26] Nitsche, M., Kindsmüller, M. C., Arend, U., Herczeg, M.: Social adaptation of ERP software - tagging UI elements. In: Online communities and social computing. Berlin [u.a.], Springer, pp. 391-400, Lecture Notes in Computer Science; 5621, 2009.
[27] Bast, H. & Weber, I.: Type less, find more: fast autocompletion search with a succinct index. In Proc. of 29th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 364–371, 2006.
[28] Rodden, K., Hutchinson, H., Fu, X.: Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the 28th international conference on Human factors in computing systems* (CHI '10). ACM, New York, NY, USA, pp. 2395-2398, 2010.

SUPERVISED THESES

[29] Schemmer, M.: Development of an ergonomic User Interface for supporting Information Search and Organization on a Mobile Device, Study Thesis, 2010.
[30] Schemmer, M.: An ergonomic user interface for vague query formulation, Diploma Thesis, 2011.
[31] Müller, R.: Konzeption und Entwicklung eines User Interfaces zur Exploration technischer Forschungsberichte, Bachelor Thesis, 2011.
[32] Scheil, K.: Trail-based Interaction for Exploratory Search, Master's Thesis, 2012.
[33] Grope, T.: Entwicklung einer Web-basierten, interaktiven Nutzungsschnittstelle zur Suche und Filterung, Bachelor's Thesis, 2012.

From Medical Images to Finite Element Models - Decision Support For Joint Replacement Surgery

Heiko Ramm (né Seim)
Medical Planning Group, Zuse Institute Berlin (ZIB), Berlin, Germany
Supervisor: Prof. Bernhard Preim, University of Magdeburg, Magdeburg, Germany
Email: ramm@zib.de

***Abstract*—This work presents a PhD project that aims at improving the preoperative planning of total joint replacement (TJR) surgery based on patient specific anatomical data. The major aim of this thesis is to develop a workflow that is able to provide decision support for the planning of TJRs by making the prediction of the postoperative outcome easily accessible. A combination and improvement of methods from medical image analysis, geometry processing and biomechanics is proposed that is suitable to achieve this goal in an almost automatic manner.**

I. Introduction

Osteoarthritis (OA) is a degenerative joint disease affecting at least 20% of the population over the age of 60, leading to severe functional limitations and pain. In most cases the treatment of choice for advanced OA is a total or partial joint replacement, where an endoprosthesis has to be selected, positioned and fixed within the patient's joint to restore the best possible joint functionality in terms of longevity and pain reduction. Annually, approximately 550,000 hip (THR) and 250,000 knee joint (TKR) replacements are performed in the EU [1].

In today's clinical routine, the planning of a total joint replacement (TJR) happens in most parts intraoperatively and is heavily based on the experience of the surgeon. Available planning systems in orthopaedics are limited to presenting the medical image data, usually X-ray, to the surgeon for inspection and allow for simple visual overlay of 2-dimensional (2D) implant templates (cf. Figure 1). More desirable would be a method that allows for an objective preoperative assessment of a planned intervention in terms of the expected joint functionality considering the individual patient's characteristics. This includes musculoskeletal anatomy, height, weight or the individual joint reaction forces. Questions that could be answered by such a system are for example: What is the functional outcome if the implant is implanted according to a certain plan [2]? How does the stress distribution in the bone look like if another type of implant is selected? Where to place the bone cuts to increase the implant to bone interface?

However, such a system is currently not available. The reasons for this lack of suitable methods for computer assisted planning in orthopaedics are manifold. First, recent and ongoing advances in medical imaging have introduced an increasing amount of high quality and high resolution medical image data [3]. This data has to be processed in order to extract the relevant information, i.e. anatomical measures, geometric models, etc. In clinical routine this is not feasible without automated methods. Second, advances in biomechanical computation have only recently gained a level, where they provide a valuable source of additional information to the surgeon that even allows for prediction of functional outcome [4], [5], [6].

Fig. 1. Typical workflow for the preoperative planning of TJRs in today's clinical routine. Blue: computer assisted steps.

The goal of this dissertation is to improve decision support for joint replacement surgery by extending the traditional preoperative therapy planning of such procedures in the following ways:

- consider subject specific 3-dimensional (3D) musculoskeletal information for therapy planning and allow for (guided) virtual implantation considering the subject's anatomy,
- provide easy access to state of the art biomechanical methods, like finite element analysis (FEA), to predict postoperative functional outcome of the implanted joint that enables the surgeon to verify/correct an existing therapy plan, and
- provide automatic and seamless integration of the planning steps to avoid additional manual effort and make therapy planning objectively verifiable

Within this work an extended planning workflow is proposed that reflects these goals (see Figure 2). There, additional (automatic) steps are added to the current therapy planning pipeline that will lead to increased knowledge available before surgery.

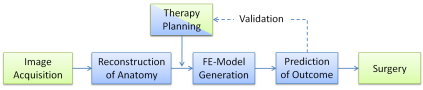


Fig. 2. Desired workflow for preoperative planning of TJRs.

The first goal of the improved planning workflow is to

Heiko Ramm is supported by the EU-FP7 Project MXL (ICT-2009.5.2).

make 3D patient specific anatomical models available to the planning pipeline (Anatomical Reconstruction). Second, computer assisted guidance will be added to the traditional therapy planning, namely the positioning of the implant. Based on patient specific anatomical features, a therapy plan will be generated that can be modified or accepted (Virtual Implantation). Based on the generated therapy plan geometric models will be generated that are suitable for FEA (FE-Model Generation). By application of state-of-the-art FEA methods patient specific outcome predictions can then be performed to provide valuable decision support before surgery (Prediction of Outcome), which is currently not available.

Contribution: The goal of this work is to implement the above proposed workflow for the planning of TJRs for the three big joints in the human body, namely the knee, the hip and the shoulder. To reach this goal suitable methods have to be found for each planning step and if necessary extended allowing full integration into the automatic pipeline. This work is sub-divided into the following work packages (WPs):*Anatomical Reconstruction*, *Virtual Implantation* and the *FE-Model Generation*. Also an essential part of the planning workflow, the *Prediction of Outcome* using numerical simulation lies outside the scope of this work. However, it will be used to verify the above mentioned methods. The specific goals of each WP are given below.

Anatomical Reconstruction:

1) Investigate and discuss existing methods to reconstruct 3D anatomical models from tomographic or planar medical image data.
2) Develop method, suitable for automatic reconstruction of 3D anatomical models of the knee, the hip and the shoulder from tomographic data.
3) Propose an approach to reconstruct 3D anatomical shape information from 2D projection images using shape and intensity information.
4) Evaluate reconstruction accuracy compared to an expert-defined gold-standard.

Virtual Implantation:

1) Investigate and discuss existing approaches for computer-aided positioning/planning of TJRs.
2) Identify suitable approaches to automatically position an implant in a 3D model of the joint.
3) Develop methods to automatically extract key anatomical landmarks and axes of the human skeletal system.
4) Provide strategies to automatically select and position an implant within the joint anatomy by using the selected anatomical features.
5) Evaluate strategies for implant positioning by means of expert feedback (e.g. questionnaire).

FE-Model Generation:

1) Investigate and discuss existing methods for automatic generation of FE models from multiple input objects.
2) Develop a method for automatic generation of FE-ready models based on bone and implant geometry, requiring Boolean operations on the input geometries and correct assignment of material properties.
3) Evaluate meshing method by either large scale FE studies on implanted geometries or evaluation of mesh quality.

The single methods and the proposed pipeline in its entirety will be evaluated with clinical data to demonstrate the potential of the proposed workflow.

II. Related Work

Computer-assisted medical planning can be regarded as generating a blueprint that is later transferred to the operating room or, as the problem will be addressed within this PhD project, as planning an intervention and additionally predicting the expected outcome. For the former computer-assisted methods for planning orthopedic surgery became popular in the late 1990s. Typically based on CT data, those approaches were restricted to position implants in a 3D environment for the knee [7], [8], [9], the hip [10], [11] or the shoulder [12].

Only in recent years the aspect of predicting the functional outcome of a therapy plan became more and more important. Dick et al. [2] presented a method that allows for interactive positioning of a hip prosthesis. Their method instantly provides a visual feedback of the expected strain distribution in the femur (thigh bone). Also on the proximal femur Bryan et al. [6] and on the proximal tibia (shinbone) Galloway et al. [13] perform large scale finite element analyses to study the influences that might lead to early migration of implant components. Both including automatic implant positioning and FE-mesh generation. Krekel et al. [14] introduced a method to predict the range of motion after shoulder arthroplasty by approximating the glenohumeral joint by a ball joint. With the aim of improving the preoperative osteotomy planning of at the upper extremities, Fürnstahl presented his approach for a preoperative planning workflow [15]. While no implants are involved, this work includes the simulation of the expected functional outcome.

The methods currently available for preoperative planning of TJRs only cover parts the complete workflow, e.g., image segmentation, generation of anatomical model, simulation etc. Other aspects are either not addressed or performed manually. In the following sections the single aspects that are to be integrated into an automated planning workflow for TJR that covers anatomical reconstruction, virtual implantation and FE model generation will be discussed.

A. Anatomical Reconstruction

Medical image data like CT or X-ray is routinely used in orthopedics and provides detailed 3D and 2D anatomical shape information. This image data, however, cannot be used directly for planning and biomechanical simulation, it has to be processed to extract the individual anatomy in suitable formats, e.g., geometric models of the bony joint structures. It is therefore the first goal of this work to make this 3D anatomical shape information from tomographic or planar image data readily available to the subsequent planning steps. The focus of this work is the extraction of bony anatomy of the joint structures for which the planning will be performed.

In the last decades higher level approaches for medical image segmentation, like atlas-based registration [16] or approaches based on deformable geometric models [17], [18], have proven to be a reliable tool that offer the necessary robustness for full automation. Atlas-based registration, like applied by Ehrhardt et al. [19] can be used to segment an image in an automatic manner and at the same time reconstruct anatomical landmarks that might be necessary for therapy planning. However, for accurate fine-level segmentation those approaches are still computationally inefficient and typically a post-processing step is necessary to reconstruct a geometric model from the deformed atlas.

As a representation of deformable model approaches, statistical shape models (SSM) like proposed by Cootes and Taylor [20] are widely used for medical image segmentation [21]. SSMs depict the mean shape of an anatomical structure plus its typical geometric variation and can be comprehensively described by only a few parameters (see Figure 3). To reconstruct bony anatomy of joint structures a number of SSM-based approaches have been introduced in the last decade [22], [23]. In most cases user intervention, like manual positioning or manual correction of an automatic segmentation result is required [24] that does not allow for an application in an automatic planning workflow as it is desired within this work. To overcome those limitations the first goal of this work project is to develop an automatic method for anatomical reconstruction that allows for segmentation of different joint regions from 3D medical image data and the generation of 3D geometric models based on SSMs. This requires the generation of set of SSMs of the joints of interest. An automatic method has to be developed to initialize, i.e. position, an SSM in medical image data. Existing SSM-based segmentation strategies are investigated, and if necessary extended, to provide the necessary accuracy for the planning of TJRs.

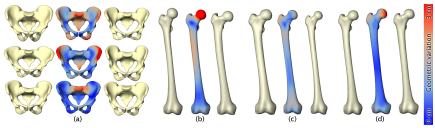


Fig. 3. Statistical shape models of the pelvic bones (a) and the right femur (b-d). For both structures the first three modes of shape variation are displayed with the geometric variation color encoded on the mean surface. First to third shape mode: pelvis(top-bottom), femur (b-d).

Another advantage of SSMs is the possible application to the problem of reconstructing a 3D shape from 2D image data like X-ray, which is widely used for preoperative planning of TJRs in clinical routine. Zheng et al. [25] and Lamecker et al [26] propose SSM-based methods to retrieve the 3D shape of the pelvic bones from single- or multi-planar projection images. Using a similar approach, Baka et al. [27] introduced a method for reconstruction of the proximal femur from bi-planar projection images and evaluated their approach on cadaver datasets with well defined boundaries. The above mentioned methods are purely based on the contour of an object in the X-ray image, thus, rendering the reconstruction of objects that are invariant w.r.t. certain transformations an ill-posed problem, e.g., the bones of the knee joint. Within the scope of this work, it will also be investigate if and how the SSMs used for segmentation from 3D image data can be adopted to the problem of reconstruction of 3D bony anatomy from 2D image data, without losing the generality to be applied in the preoperative planning workflow.

B. Implant Positioning

Having the geometric model of the patient's anatomy available from the previous step in the workflow, the goal is then to select and align the implant component(s). Surgical guidelines or recommendations exist for the three major joints, e.g., the knee [28], the hip [29], or the shoulder [30]. However, the final adjustment is strongly influenced by the experience and preferences of the surgeon. Other parameters that have an influence on the surgery are the type of the selected surgical access or the location of ligaments and tendons.

Typically, the existing surgical guidelines can be summarized as follows: align the implant coordinate system to the joint's coordinate system that is depending on anatomical axis, planes, landmarks, etc. In recent years, several large-scale biomechanical studies have been performed where an implant was positioned following surgical recommendations. Viceconti et al. [11] and Otomura et al. [31] both optimize the position of a manually aligned hip implant by a registration approach. Galloway et al. [13] position tibial trays of different sizes according to automatically extracted anatomical features on the proximal tibia.

To establish a joint's coordinate systems anatomical or geometric landmarks have to be identified (see Figure 4). In recent years a number of works focused on the automatic extraction of such landmarks from medical image data for different application scenarios. With the goal of providing an orthopedic planning environment, Ehrhardt et al. [19] proposed a non-rigid registration approach for detecting pelvic landmarks in CT data. They employ an atlas that includes labeled voxels and landmarks. Geometric landmark localization based on extremal differential properties like ridges, corners or saddles, was introduced by Wörz and Rohr [32]. Izard et al. [33] suggest an algorithm for landmark detection based on a probabilistic model of image intensities. By learning the image intensities in a set of manually landmarked images, they create a tissue-probability map, which is then aligned to new image data by a likelihood maximization approach. A different method, employing techniques from machine learning, was introduced by Dikmen et al. [34]. They present a three-stage system to roughly locate, verify and finally correct the positions of anatomical landmarks based on previously learned spatial relationships and image features.

The methods mentioned above are either limited to a single imaging protocol because they focus on specific image features, or they are restricted to detect only geometric landmarks, whereas, anatomical landmarks are not always distinguishable

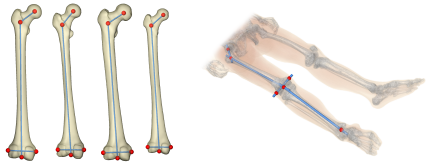


Fig. 4. Model of the femur with landmarks (red) relative to the surface (a). Adaptation of the model to image data allows for direct extraction of the given landmarks (b).

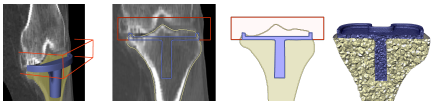


Fig. 5. Typical finite element meshing scenario (from left to right): a implant (blue) and a resection geometry (red) are aligned to a patient's bone (beige), schematic cross-section of this setup, the opaque parts are preserved during meshing, cross section through the final tetrahedral mesh.

by their geometric appearance. One of the goals of this work is to investigate whether existing methods for automatic landmark detection are suitable to establish a joint coordinate system to be applied within the proposed planning pipeline or if a new method has to be developed. Key aspects that need to be considered by such a method are: (1) generality in terms of application to different joint anatomies and (2) integration into the planning workflow in terms of data types (image data, geometric models).

As mentioned before the selection of the joint coordinate system is very surgeon specific and also varies in literature. Therefore, final alignment is the only step in the planning pipeline, where user interaction might be necessary. This step should be supported by the planning framework in such a way that a *good* initial position and size of the implant based on patient specific features is provided. The surgeon would then be able to easily adjust the initial position supported by 3D manipulation and visualization techniques, e.g., 3D snap dragging presented by Bier [35] or interactive distance visualization as presented by Dick et al. [36].

C. FE-Mesh Generation

With the reconstructed individual anatomy and the aligned implant available, finite element mesh generation is the final step in the automatic planning pipeline to generate suitable models for FEA. FEA has become an approved instrument in orthopedic surgery to preoperatively assess the functional outcome of joint replacement procedures as it allows for an objective evaluation of bone-implant compounds, like implant wear [37] or bone stress distributions [6]. When optimizing implant size and position for a single subject or when investigating new implant designs for a large population [13], numerous different implantation configurations have to be analyzed to determine and compare characteristics like the expected strains and stresses during everyday activities. To perform such analyses in the automatic planning pipeline, FE-meshes have to be automatically generated that minimize computation time while maintaining reasonable accuracy of the FEA outcome (compare Figure 5).

The type of finite elements chosen to discretize a domain depends on the application. According to Owen [38], tetrahedral and hexahedral meshes form the majority of unstructured meshes. Since any polyhedron can be subdivided into tetrahedra, tetrahedral meshes are the most general. FEA for orthopedic applications is typically performed using pure tetrahedral meshes [11], [37], [6], [13], because in most cases the user is interested in the complete domain of the joint-implant compound. Other medical applications where FEA is performed, like fluid dynamics, are typically interested in highly anisotropic phenomena, where hybrid meshes with other element types outperform tetrahedral meshes [39].

Another important aspect is the required geometric accuracy of the finite element mesh. To generate finite element meshes in a joint replacement scenario the input is typically given as a combined description of mechanical parts like the implant component including sharp edges and the bone that is more or less a smooth organic shape provided as a reconstructed geometric model or a voxel representation. A method to generate a tetrahedral mesh from this combination, ideally is able to preserve the geometric detail of the input structures where required without generating too much finite elements.

Existing tetrahedral meshing methods (a comprehensive overview is given by Liseikin [40]) lack the robustness or the geometric accuracy to be used for automatic generation of bone-implant meshes. They typically require an explicit fusion of the objects to be discretized before meshing, for instance using a single voxel- or surface-representation. While inconsistencies resulting from object overlaps are resolved easily within a voxel grid like used by Zhang et al. [41], the resulting mesh typically suffers from artificially introduced inaccuracies, for example lost information about sharp edges. Alternatively, a consistent description as a surface triangulation that separates the different objects (or domains) can be computed. This is likely to cause problems where boundary intersections introduce small angles or narrow inter-boundary regions occur. A triangulation of those regions leads to very small or badly shaped triangles. If a meshing approach is used that depends on this boundary representation, like the advancing front method [40], unnecessarily small or badly shaped finite elements will be introduced.

It is therefore one of the goals of this work to develop a method that is suitable to mesh bone-implant compounds in an automatic and robust way using a given therapy plan, e.g., patient specific anatomy and positioned implant. If a suitable method to generate FE-ready meshes can be identified its suitability has to be validated meshing a large set of different implantation settings. Such a method could be applied in any scenario where FEA is performed for a combination of

organic and mechanical parts for which sharp edges have to be preserved.

III. Contribution and Methods

This section presents results that have already been achieved in the specific areas of the planning pipeline, namely anatomical reconstruction, implant positioning, and FE-mesh generation.

A. Anatomical Reconstruction

In [42] we presented a fully automatic method for segmentation of the human pelvic bones from CT datasets that is based on the application of a statistical shape model (see Figure 6). The proposed method was divided into three steps: (1) The averaged shape of the pelvis model is initially placed within the CT data using the Generalized Hough Transform, (2) the statistical shape model is then adapted to the image data by a transformation and variation of its shape modes, and (3) a final free-form deformation step based on optimal graph searching is applied to overcome the restrictive character of the statistical shape representation. The achieved accuracy of the segmentation and an average segmentation time of less than 5 minutes on a standard desktop computer indicated that our method meets the requirements of clinical routine. We showed the generality and accuracy of this approach by successfully adapting it to other joint structures and image modalities [43], [44]. In [45] we have shown that our 3D reconstruction framework can also be applied to assess the post-operative outcome of TJRs in follow-up examinations by fitting rigid re-engineered geometric models of prostheses to CT data of the knee.

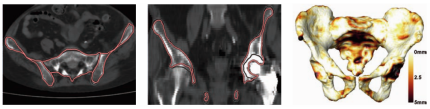


Fig. 6. Automatic segmentation result of the pelvic bones as presented in [42].

Currently we are investigating the application of statistical shape models that are enriched with volumetric intensity information to the problem of 3D shape reconstruction from 2D image data. By simulating realistic X-ray images by projection of this intensity information, we aim at applying image-based error measures like mutual information to improve the robustness of 3D/2D reconstruction where pure silhouette approaches fail.

B. Implant Positioning

As a first step towards the integration of musculoskeletal soft tissue structures that are relevant for orthopedic intervention planning, we presented an approach for reconstructing ligament and tendon attachment sites from 3D medical image data [46] based on SSMs. To show the potential of our method, we manually extracted the surface of 11 distal femora and their corresponding ligament attachment sites in clinical CT data and compared them to the results of our method.

In [47] we presented and compared three different methods for automatic detection of anatomical landmarks in pelvic CT data. The methods exhibited different degrees of generality in terms of portability to other anatomical regions and require a different amount of training data. For our most generic approach only a small set of training landmarks is required. Those landmarks are transferred to the patient specific geometry based on mean value coordinates (MVCs). We evaluated and compared our methods on 100 clinical CT datasets, for which gold standard landmarks were defined manually by multiple observers. In [48], we showed that our landmark extraction approach can easily be integrated into our extraction pipeline because it relies on automatic segmentation that are provided by our frame work.

To establish the joint coordinate systems of the knee bones, i.e. tibia, femur, and patella (knee cap), we recently presented a fully automatic method in [45]. There geometrical as well as anatomical landmarks are identified by making use of our previous approach [47] and geometric reconstruction following surgical guidelines.

C. FE-Mesh Generation

In a first study we introduced a problem specific method for meshing of the automatically implanted proximal tibia [13]. The tibia was first resected, i.e. bone has been removed, and merged with the implant geometry using Boolean-like operation on the input surface meshes and afterwards remeshed to ensure good triangle quality. From this high-quality surface triangulation a tetrahedral mesh was generated using the advancing front approach.

Recently based on an approach by Pons et al. [49] we proposed a method that outputs a merged tetrahedral mesh of the patient's joint anatomy and an arbitrarily positioned implant geometry suitable for finite element analysis [50]. Our approach has several advantages: (1) it avoids error-prone intermediate stages, e.g., data type conversion, (2) it is able to preserve constraints such as sharp edges and (3) it can be fully automated by initially defining a few parameters that describe the desired geometric accuracy and element quality. Based on the meshing of 100 different patient-specific bone-implant configurations at the tibia (shinbone), we show that our approach produces high-quality meshes in all cases automatically. Our approach is currently adopted to the hip and the shoulder joints (see Figure 7).

IV. Discussion and Conclusion

This work presents a new automated workflow to improve the preoperative planning of joint replacement surgery. Methods from medical image analysis, feature detection and geometry processing are combined and extended to provide easy access to biomechanical methods that allow for prediction of preoperative outcome.

Existing methods to extract 3D anatomical models from 3D medical image data have been automated and are currently

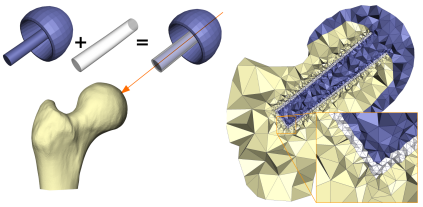


Fig. 7. Automatic generation of an finite element mesh for the hip joint (proximal femur): A cement layer (white), an implant (blue) and the individual bone (beige) are combined.

adopted to the problem of extracting 3D anatomy from 2D image data. Our approach automatically generates geometric models that are suitable for the subsequent workflow. For the second stage of the pipeline, new automatic approaches based on SSMs have been developed within the scope of this PhD project to establish joint coordinate systems based on anatomical landmarks. Using those methods we could show, that a reasonable initial implant position can be achieved following surgical guidelines. As the last part of the workflow that is addressed within this work a method has been developed to robustly generate FE-meshes from bone-implant compounds. The single aspects of the pipeline have been validated individually, each offering high accuracy and robustness w.r.t. full automation.

At the moment, the major aspect of this work that is still missing is the validation of the fully integrated workflow. Although, some aspects have already been validated in combination, e.g., implant positioning and meshing, we still have to validate the overall benefit: Does the workflow improve the preoperative planning of TJRs without introducing additional manual effort? This validation will be validated in the near future on the example of at least one of the major joints (ideally all). The direct feedback of surgeons is essential to validate this pipeline, this can be achieved by surveys or user studies. The details of such a validation are still open questions.

Another minor aspect that needs to be validated is the reconstruction of 3D anatomy from 2D image data, like X-ray images that is still the most common form of image data in orthopaedics. This will make the methods developed here available to a much broader audience.

Although not all problems are solved in its entirety, the single aspects provide robust modules that now need to be plugged together to reach our initial goal: providing patient-specific decision support system that provides easy access to functional prediction as it is currently not available.

REFERENCES

[1] H. Kiefer, "Current trends in total hip arthroplasty in Europe and experiences with the bicontact hip system," *Treatment of Osteoarthritic Change in the Hip*, pp. 205–210, 2007.

[2] C. Dick, J. Georgii, R. Burgkart, and R. Westermann, "Stress tensor field visualization for implant planning in orthopedics," *IEEE Transactions on Visualization and Computer Graphics (Proceedings of IEEE Visualization 2009)*, vol. 15, no. 6, pp. 1399–1406, 2009.

[3] T. Fuchs, M. Kachelriess, and W. A. Kalender, "Technical advances in multi-slice spiral ct." *Eur J Radiol*, vol. 36, no. 2, pp. 69–73, Nov 2000.

[4] M. O. Heller, J. H. Schröder, G. Matziolis, A. Sharenkov, W. R. Taylor, C. Perka, and G. N. Duda, "Musculoskeletal load analysis. a biomechanical explanation for clinical results – and more?" *Der Orthopäde*, vol. 36, pp. 188–194, 2007.

[5] E. Schileo, F. Taddei, A. Malandrino, L. Cristofolini, and M. Viceconti, "Subject-specific finite element models can accurately predict strain levels in long bones." *J Biomech*, vol. 40, no. 13, pp. 2982–2989, 2007.

[6] R. Bryan, P. B. Nair, and M. Taylor, "Use of a statistical model of the whole femur in a large scale, multi-model study of femoral neck fracture risk." *Journal of biomechanics*, vol. 42, no. 13, pp. 2171–6, Sep. 2009.

[7] M. Fadda, D. Bertelli, S. Martelli, and M. Marcacci, "Computer assisted planning for total knee arthroplasty," in *CVRMed-MRCAS'97*, ser. Lecture Notes in Computer Science, J. Troccaz, E. Grimson, and R. Mösges, Eds., vol. 1205. Berlin/Heidelberg: Springer-Verlag, 1997, pp. 617–628.

[8] R. Ellis, C. Tso, and J. Rudan, "A surgical planning and guidance system for high tibial osteotomy," *Journal of Computer Aided Surgery*, vol. 1496, 1999.

[9] W. Müller, U. Bockholt, G. Voss, A. Lahmer, and M. Börner, "Planning system for computer assisted total knee replacement." *Studies In Health Technology And Informatics*, vol. 70, pp. 214–219, 2000.

[10] H. Handels, J. Ehrhardt, W. Plötz, and S. Pöppl, "Virtual planning of hip operations and individual adaption of endoprostheses in orthopaedic surgery," *International Journal of Medical Informatics*, vol. 58-59, no. 1, pp. 21–28, Sep. 2000.

[11] M. Viceconti, D. Testi, M. Simeoni, and C. Zannoni, "An automated method to position prosthetic components within multiple anatomical spaces." *Computer Methods and Programs in Biomedicine*, vol. 70, no. 2, pp. 121–127, 2003.

[12] E. Valstar, "Towards computer-assisted surgery in shoulder joint replacement," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 56, no. 5-6, pp. 326–337, 2002.

[13] F. Galloway, M. Kahnt, H. Seim, P. B. Nair, P. Worsley, and M. Taylor, "A large scale finite element study of an osseointegrated cementless tibial tray," in *23th Annual Symposium International Society for Technology in Arthroplasty*, 2010.

[14] P. R. Krekel, P. W. de Bruin, E. R. Valstar, F. H. Post, P. M. Rozing, and C. P. Botha, "Evaluation of bone impingement prediction in pre-operative planning for shoulder arthroplasty," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 223, no. 7, pp. 813–822, 2009.

[15] P. Fürnstahl, "Computer-assisted planning for orthopedic surgery," Ph.D. dissertation, ETH ZURICH, 2010.

[16] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, Jr., "Quo vadis, atlas-based segmentation?" in *The Handbook of Medical Image Analysis - Volume III: Registration Models*, ser. Topics in Biomedical Engineering International Book Series, J. S. Suri, D. L. Wilson, and S. Laxminarayan, Eds. Boston, MA: Springer US, 2005, ch. 11, pp. 435–486.

[17] C. Xu, D. Pham, and J. Prince, "Image segmentation using deformable models," *Handbook of Medical Imaging*, vol. 2, pp. 129–174, 2000.

[18] D. Jayadevappa, S. Kumar, and D. Murty, "Medical Image Segmentation Algorithms using Deformable Models: A Review," *IETE Technical Review*, vol. 28, no. 3, pp. 248–255, 2011.

[19] J. Ehrhardt, H. Handels, W. Plötz, and S. J. Pöppl, "Atlas-based recognition of anatomical structures and landmarks and the automatic computation of orthopedic parameters." *Methods Inf Med*, vol. 43, no. 4, pp. 391–397, 2004.

[20] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[21] T. Heimann, *Statistical Shape Models for 3D Medical Image Segmentation*. VDM Verlag, 2009.

[22] J. Fripp, S. Crozier, S. K. Warfield, and S. Ourselin, "Automatic segmentation of the bone and extraction of the bone-cartilage interface from magnetic resonance images of the knee." *Phys Med Biol*, vol. 52, no. 6, pp. 1617–1631, 2007.

[23] H. Lamecker, M. Seebaß, H.-C. Hege, and P. Deuflhard, "A 3D statistical shape model of the pelvic bone for segmentation," in *Proceedings*

of SPIE - Volume 5370 Medical Imaging 2004: Image Processing, J. Fitzpatrick and M. Sonka, Eds., May 2004, pp. 1341–1351.

[24] J. Pettersson, H. Knutsson, and M. Borga, "Automatic hip bone segmentation using non-rigid registration," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE Computer Society, 2006, pp. 946–949.

[25] G. Zheng, "Reconstruction of Patient-Specific 3D Bone Model from Biplanar X-Ray Images and Point Distribution Models," in *ICIP06*, 2006, pp. 1197–1200.

[26] H. Lamecker, T. H. Wenckebach, and H.-C. Hege, "Atlas-based 3{D}-shape reconstruction from x-ray images," in *Proc. Int. Conf. of Pattern Recognition (ICPR2006)*, vol. Volume I. IEEE Computer Society, 2006, pp. 371–374.

[27] N. Baka, B. L. Kaptein, M. de Bruijne, T. van Walsum, J. E. Giphart, W. J. Niessen, and B. P. F. Lelieveldt, "2D-3D shape reconstruction of the distal femur from stereo X-ray imaging using statistical shape models." *Medical image analysis*, vol. 15, no. 6, pp. 840–50, Dec. 2011.

[28] C. Fitzpatrick, D. FitzPatrick, D. Auger, and J. Lee, "A tibial-based coordinate system for three-dimensional data." *The Knee*, vol. 14, no. 2, pp. 133–7, Mar. 2007.

[29] G. E. Lewinnek, J. L. Lewis, R. Tarr, C. L. Compere, and J. R. Zimmerman, "Dislocations after total hip-replacement arthroplasties." *The Journal of bone and joint surgery. American volume*, vol. 60, no. 2, pp. 217–20, Mar. 1978.

[30] G. Wu, F. C. van der Helm, H. (DirkJan) Veeger, M. Makhsous, P. Van Roy, C. Anglin, J. Nagels, A. R. Karduna, K. McQuade, X. Wang, F. W. Werner, and B. Buchholz, "ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motionŮPart II: shoulder, elbow, wrist and hand," *Journal of Biomechanics*, vol. 38, no. 5, pp. 981–992, May 2005.

[31] I. Otomaru, M. Nakamoto, M. Takao, N. Sugano, Y. Kagiyama, H. Yoshikawa, Y. Tada, and Y. Sato, "Automated Preoperative Planning of Femoral Component for Total Hip Arthroplasty (THA) from 3D CT Images," *Medical Imaging and Augmented Reality*, vol. 5128, pp. 40–49, Aug. 2008.

[32] S. Wörz and K. Rohr, "Localization of anatomical point landmarks in 3D medical images by fitting 3D parametric intensity models," *Med Image Anal*, vol. 10, no. 1, pp. 41–58, 2006.

[33] C. Izard, B. Jedynak, and C. Stark, "Spline-based probabilistic model for anatomical landmark detection," *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2006*, pp. 849–856, 2005.

[34] M. Dikmen, Y. Zhan, and X. S. Zhou, "Joint detection and localization of multiple anatomical landmarks through learning," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 6915, Apr. 2008.

[35] E. A. Bier, "Snap-dragging in three dimensions," *ACM SIGGRAPH Computer Graphics*, vol. 24, no. 2, pp. 193–204, Mar. 1990.

[36] C. Dick, R. Burgkart, and R. Westermann, "Distance visualization for interactive 3D implant planning." *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2173–82, Dec. 2011.

[37] K. Ong and S. Kurtz, "The use of modelling to predict implant behaviour." *Medical device technology*, vol. 19, no. 5, pp. 64–6, Sep. 2008.

[38] S. Owen, "A survey of unstructured mesh generation technology," in *7th International Meshing Roundtable*, 1998, pp. 239–267.

[39] C. Wang, K. Pekkan, D. de Zélicourt, M. Horner, A. Parihar, A. Kulkarni, and A. P. Yoganathan, "Progress in the CFD modeling of flow instabilities in anatomical total cavopulmonary connections." *Annals of biomedical engineering*, vol. 35, no. 11, pp. 1840–56, Nov. 2007.

[40] V. D. Liseikin, *Grid Generation Methods*, ser. Scientific Computation. Dordrecht: Springer Netherlands, 2010.

[41] Y. Zhang, T. J. R. Hughes, and C. L. Bajaj, "An automatic 3D mesh generation method for domains with multiple materials," *Computer Methods in Applied Mechanics and Engineering*, 2009.

[42] H. Seim, D. Kainmueller, M. Heller, H. Lamecker, S. Zachow, and H.-C. Hege, "Automatic segmentation of the pelvic bones from ct data based on a statistical shape model," in *Eurographics Workshop on Visual Computing for Biomedicine (VCBM)*, Delft, Netherlands, 2008, pp. 93–100.

[43] D. Kainmüller, H. Lamecker, H. Seim, M. Zinser, and S. Zachow, "Automatic extraction of mandibular nerve and bone from cone-beam CT data," in *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, September 20 - 24 2009, accepted for publication.

[44] H. Seim, D. Kainmueller, H. Lamecker, M. Bindernagel, J. Malinowski, and S. Zachow, "Model-based Auto-Segmentation of Knee Bones and Cartilage in MRI Data," in *Proc. of Medical Image Computing and Computer-Assisted Intervention (MICCAI); Medical Image Analysis for the Clinic: A Grand Challenge*, 2010, pp. 215–223.

[45] K. C. Ho, S. K. Saevarsson, H. Ramm, R. Lieck, S. Zachow, G. B. Sharma, E. L. Rex, S. Amiri, B. C. Wu, A. Leumann, and C. Anglin, "Computed tomography analysis of knee pose and geometry before and after total knee arthroplasty," *Journal of Biomechanics (accepted)*, 2012.

[46] H. Seim, H. Lamecker, and S. Zachow, "Segmentation of bony structures with ligament attachment sites," in *Bildverarbeitung für die Medizin 2008*, ser. GI Informatik aktuell, T. et al., Ed. Springer, 2008, pp. 207–211.

[47] H. Seim, D. Kainmueller, M. Heller, S. Zachow, and H.-C. Hege, "Automatic extraction of anatomical landmarks from medical image data: An evaluation of different methods," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, Jun. 2009, pp. 538–541.

[48] H. Seim, D. Kainmüller, H. Lamecker, and S. Zachow, "A system for unsupervised extraction of orthopaedic parameters from ct data," in *GI Workshop Softwareassistenten - Computerunterstützung für die medizinische Diagnose und Therapieplanung*, ser. GI-Edition Lecture Notes in Informatics, 2009, pp. 1328–1337.

[49] J. Pons, F. Ségonne, J. Boissonnat, L. Rineau, M. Yvinec, and R. Keriven, "High-quality consistent meshing of multi-label datasets," in *Information Processing in Medical Imaging*. Springer, 2007, pp. 198–210.

[50] M. Kahnt, F. Galloway, H. Seim, H. Lamecker, M. Taylor, and S. Zachow, "Robust and Intuitive Meshing of Bone-Implant Compounds," in *Jahrestagung der Deutschen Gesellschaft für Computer- und Roboterassistierte Chirurgie e. V.({CURAC})*, 2011, pp. 71–74.

Motion compensation of ultrasonic perfusion images using MRFs and coupled segmentation

Sebastian Schäfer
Department of Simulation and Graphics
Otto-von-Guericke University Magdeburg
Magdeburg, Germany
Email: sebastian.schaefer@ovgu.de
Supervisor: Klaus Tönnies

***Abstract*—Contrast-enhanced ultrasound (CEUS) is a rapid and inexpensive medical imaging technique to assess tissue perfusion with a high temporal resolution. It is composed of a sequence with ultrasound brightness values and a contrast sequence acquired simultaneously. However, the image acquisition is disturbed by various motion influences, which must be compensated to extract valid information about perfusion for diagnostic purposes. In this work a new approach for motion compensation using markov random fields (MRF) to constrain the search of transformation parameters spatially and temporally is presented. Furthermore, the method is coupled with a segmentation as an additional similarity term to lower the influence of noise in the data. In first experiments, the concept is tested with simulated artificial data at different noise levels. It is shown that the presented method produces results with smooth and stable deformation fields yielding the same accuracy compared to a method allowing arbitrary and unconstrained transformation parameters. The new method also produces superior results with real patient data.**

I. Introduction

Dynamic medical imaging techniques are used to measure functional processes for early detection and diagnosis of diseases and pathologies. Perfusion imaging is a substantial part of dynamic medical imaging to describe and quantify the passage of fluids through blood vessels, the lymphatic system, organs or tissue. Therefore, signal acquisition is performed consecutively, depicting multiple instances of the same region of interest (ROI) over time, resulting in an additional function dimension t. The spatial domain of the region of interest can either be two-dimensional or three-dimensional leading to 2D+t or 3D+t data.

2D ultrasonography (US) is one of the most widely used medical imaging techniques because it enables immediate and inexpensive examinations with high spatial resolution. It is well suited for imaging abdominal and thoracic organs. There are no contraindications and the patient is not exposed to radiation. US is also used for perfusion imaging employing contrast agents (CA), consisting of gas-filled micro bubbles that have a high degree of echogenicity as they increase the US backscatter [1]. By acquiring 2D contrast enhanced US (CEUS) multi-frame sequences, propagation and contrast uptake after the injection of the CA can be measured to assess perfusion kinetics. This has become a suitable tool for delineating the vascular structure in normal and pathological tissue in order to detect primary tumors or metastases in various organs or to assess disease activity in Crohn's disease [2], [3]. The perfusion analysis is performed by extracting and understanding the perfusion kinetics of the blood in the tissue of interest from the acquired multi-frame data. An essential advantage of CEUS imaging is the assessment of contrast enhancement with considerably higher temporal resolution compared to other perfusion imaging techniques [4].

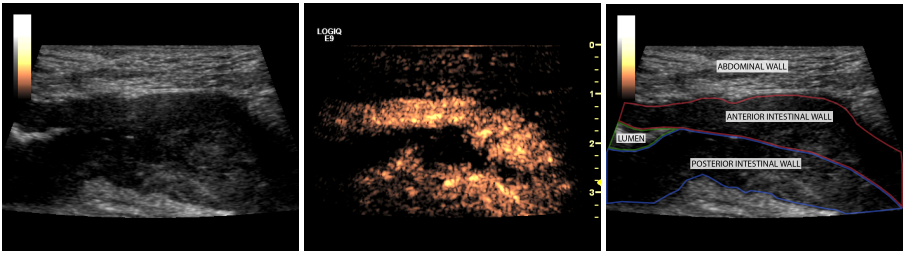
During CEUS examinations studying perfusion, the sonographer normally holds the US probe still in a particular position and orientation to image a suitable slice of tissue of interest during CA administration. However, the data acquired with this examination methodology often contains significant motion. This is due to patient movement through breathing and differently induced motion (intrinsically induced motion) are present in addition to the motion caused by tilting or moving the US probe (extrinsically induced motion), since US imaging is normally performed hand-held.

While both motion types can normally be interpreted by well trained physicians [5], in computer-assisted analysis the different image frames of a time-dependent acquisition need to be aligned in order to extract valid perfusion parameters over time.

Motion compensation in medical image analysis can be achieved by registration [6]. Manual correction of the motion by an expert for sequences with up to 1000 image frames is tedious and also prone to error because of inter- and intra-observer variability. Thus, automatic motion compensation of image sequences is required.

The automatic alignment of this stack of time-dependent image frames is complicated by the fact that the 2D US image plane can "miss" the region of interest during part of the examination, due to the three dimensional nature of the motion described above. Therefore, two types of motion effects on the images are distinguished:

- *correctable image motion*, when the image plane contains the region of interest (ROI), but this region is moved or deformed within the same plane with respect to a reference image
- *uncorrectable image motion*, when the US image plane does not contain the ROI at all, because it is moved out of the image plane



(a) (b) (c)

Fig. 1: A representative frame of a CEUS acquisition of a patient with a stenosis of the small bowel due to Crohn's disease, showing b-mode data (a) and contrast data (b), the most important regions labeled in (c).

This is taken into account by defining temporal regions with groups of frames that show the same region of interest (ROI) [7]. Through a frame similarity analysis temporal regions can be generated semi-automatically. Only frames within the same temporal region should be included in a registration and thus can provide valid pixel correspondences over time.

Research presented in this work will be targeted on the formulation of a concept and an implementation for first results to approach the specific problem of motion compensation for CEUS data. Restrictions and constraints of the data have to be included as a priori knowledge when solving the registration problem to account for deficiencies within the data such as noise or undetected out-of-plane motion. Two important issues of registration must be taken into account: the calculation of transformations to perform alignment and the determination of similarity.

To compensate for consistent global motion influence linear shifts are applied to each image in the sequence. As most of the image data contains large parts of soft tissue local deformation has to be expected as well. Most applications use free-form deformations combined with a spatial constraint to obtain a smooth deformation field. If time-dependent data with high resolution is registered, transformation can be constrained over time as well [8]. Both constraints assume that spatial and temporal continuity is given and make the approach less susceptible to outliers, because more samples are used to calculate the transformation. To determine transformation applying the above mentioned constraints the similarity between image frames has to be calculated.

The acquisition procedure of CEUS usually produces two parallel image sequences: standard b-mode[1] and the measured CA enhancement (Fig. 1a, 1b). The frames are acquired alternatingly resulting in approximately 10 frames per second for each sequence. The CA only slightly effects the signal intensity in the b-mode data sequence which therefore has a constant brightness over time for a specific organ or tissue. This can be used to assess the similarity between all frames of the sequence. However, noise hampers the correct determination of similarity. This has a large influence when free-form deformation is used. This could be enhanced if the true intensity value of the pixel would be known. Therefore, image segmentation, assigning each pixel a specific segment label, is proposed to enhance similarity determination between different image frames. The uptake signal from contrast sequence images can also be used for similarity determination. However, the uptake signal may change significantly over time, thus, valid similarity can only be calculated for images in a small temporal neighborhood. As both sequences are acquired at the same time, a single transformation can be calculated using information from b-mode and contrast images and applied to both.

To formulate an optimization scheme, incorporating all the above mentioned calculations and constraints, a markov random field (MRF) is used. A MRF is an undirected graph with nodes representing the unknowns of the system. Nodes can take labels which express a certain configuration, a transformation and a segment label in this specific case. The configuration with the highest probability leads to a low energy at the node (singleton energy). Constraints can be included through edges. Edges combine exactly two nodes evaluating the probability of both node configurations (pairwise/doubleton energy). The overall MRF energy must be minimized to obtain the global best configuration.

The research plan of this project includes evaluation of different similarity measures for motion compensation of CEUS patient data. The major challenge is to find compromises between the constraints for spatial and temporal transformation smoothness and the similarity of multiple frames of the sequence.

II. Related Work

CEUS imaging has a time dependent component resulting in motion influence stemming from different sources. It also comprises functional information in terms of blood perfusion

[1]brightness modulation: measured amplitudes are mapped onto intensity values

leading to CA enhancement. As quantification and visualization of functional signals require correct spatial correspondence of temporal samples the removal of motion influence is necessary. Yet, the literature does not offer adapted methods for the specific scenario of motion compensation of CEUS perfusion images.

However, intra-modal registration of time-dependent data has been applied in other scenarios. Shekhar et al. used mutual information as similarity measure to register time-dependent 3D US data of the left ventricle [9]. It is preceded by median filtering the image data, dropping least significant bit information and using partial volume distribution interpolation when transforming the moving image. This results in a smoother objective function and reduces the probability of the optimizer to end up in a local maxima.

To calculate and analyze the deformation of the human heart Ledesma-Carbayo et al. [10], [8] presented a combined spatio-temporal registration procedure. Similarity of the deformed 2D US frames was measured by the mean squared distance of all frames in the temporal sequence to a specified reference frame. Transformation parameters for B-Splines are found for all frames simultaneously restricting the parameters to be continuously smooth over time.

In literature MRFs are often applied to incorporate spatial information to model probabilities for the true value of the element (image restoration) or the label the element is belonging to (image segmentation) [11]. The neighborhood system is exchangeable and may contain neighbors of first, second or higher orders. The energy function is defined on the members of a neighborhood (in graph theory called cliques) and expresses the potential that the members are belonging to the same group. Still, the energy function is a strictly local description assuring conditional probabilities on the specified neighborhood only and independence between all other samples (Markov Property). As the information about segments enhances the finding of the true transformations to obtain registration and correct registration improves the generation of valid image segments, this processes have been coupled to be solved at the same time. Mahapatra et al. [12] recently proposed their approach combining both tasks, but it is only applied to two images at a time (fixed and moving image).

Wyatt et al. [13] introduce a combined segmentation and registration framework based upon MRF. The authors introduced two different schemes for parameter estimation: a simultaneous estimation scheme and a joint estimation scheme. The first one estimates all parameters (both for the registration and the segmentation) in one optimization step. The joint scheme divides the estimations into two steps being optimized independently until convergence of both optimization processes. Iterated Conditional Modes (ICM) [14] is used to find minima with quick convergence. However, this algorithm needs good initialization, otherwise local optima are found. The spacial probability is calculated for each class combination and stored in a joint class histogram (co-occurrence matrix) which is used as similarity measure (similar to the mutual information measure). As the number of segmentation classes is usually smaller than the number of intensity levels, the measure is much faster compared to standard mutual information. Finally, convergence of the registration measure can be used as termination criterion. Adaption to non-rigid transformations is planned but not explained in detail within the scope of the paper.

As extension Xiaohua et al. [15] combine a hidden MRF segmentation with B-spline based free-form deformations to register images non-rigidly. The actual state of a hidden MRF is not observable, wherefore an additional observable random field is introduced. This field is used to model the state of the hidden MRF for specific probability distribution parameters for each label configuration [16]. Using a 2-step probabilistic model from Marroquin et al. [17], the method of Xiaohua et al. is claimed to be more robust towards noise and initialization.

The advantage of MRF-based methods is that prior information of the scene can be specified on a non-object basis. That makes it suitable for a large class of problems. US image sequences have a low signal-to-noise ratio resulting in poor data quality and artifacts. Intensity distribution information can be used to generate more stable features for registration [18]. Calculation of both, transformation parameters and segment labels at the same time with a MRF also enables a combined calculation of parameters at different time steps. This should further increase robustness, as parameters are determined using more samples of the measured data. MRFs can be solved efficiently using graph cut-based approaches [19], [20].

III. Method

The temporal resolution of both simultaneously acquired sequences in CEUS is between 7 and 12 frames per second (typically 10 fps). I.e. registration parameters for each individual image are not arbitrary. They depend on corresponding parameters in adjacent images and neighboring parameters in the same image as well ensuring temporal and spatial continuity. To assign a segment label to each pixel in the images the influence of the neighboring pixels can be used to obtain more reliable results with the help of a regional neighborhood. Both dependencies can be modeled using a MRF formulation.

The MRF model in general consists of nodes $v_i \in \mathcal{V}$ and edges $e \in \mathcal{E}$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. An edge always connects exactly two nodes $e = f(v_i, v_j) \in \mathcal{E}$ with $v_i, v_j \in \mathcal{V}$. The goal is to find the most probable configuration of random variables X within the system. The Hammersley-Clifford theorem [14] stipulates the random variables X to be a MRF with respect to a neighborhood N if and only if the probability distribution of $P(X)$ is a Gibbs distribution:

$$P(X) = Z^{-1} \times e^{-U(X)}. \tag{1}$$

Z is a normalizing constant and $U(X)$ is the energy defined by the sum of all different clique potentials depending on the neighborhood system N. In image analysis and computer vision the problem is often regarded as an energy minimization

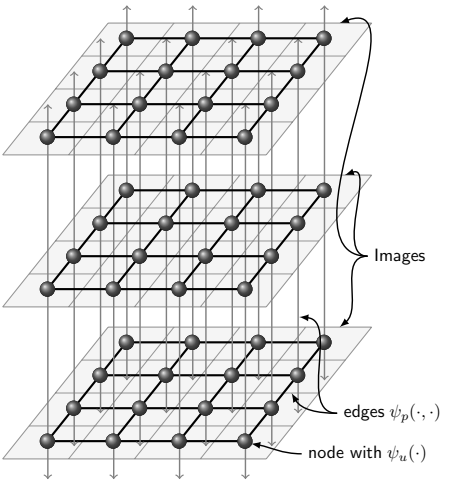


Fig. 2: The MRF model represented by a graph with nodes being assigned a unary potential function ψ_u, spatial edges within the image frames and temporal edges between image frames (both assigned a pairwise potential function ψ_p).

(according to the terminology of similar problems) using a neighborhood of pairwise interaction only. Nodes are assigned a unary potential function $\psi_u(\cdot)$ and edges are assigned a pairwise potential function $\psi_p(\cdot,\cdot)$ (Fig. 2). Then, the global energy of the MRF is defined by the sum of all unary and pairwise potential functions

$$E_{global} = \sum_{v_i \in \mathcal{V}} \psi_u(v_i) + \sum_{e=(v_i,v_j)\in\mathcal{E}} \psi_p(v_i, v_j), \tag{2}$$

and must be minimized in order to find the best solution according to the constraints made in the system by varying the labels of the nodes in the MRF.

A. MRF for registration

In the case of image registration these potential functions can be used to evaluate different transformation parameters with respect to the overall problem. As a consequence, each node v_i represents a transformation parameter in the system and must be assigned a label $l_k \in \mathcal{L}$ representing the value of the current parameter, a translation in 2D defined by a vector $l_k = (t_x, t_y)$. Edges can be inserted between two nodes exhibiting a certain dependency or constraint. Unary potential functions $\psi_u(\cdot)$ represent the contribution of the current parameter to the fitting quality in terms of a similarity definition and pairwise potential functions $\psi_p(\cdot,\cdot)$ are used to model dependencies between parameters.

Common choices for similarity calculation apart from using segment label comparison are Normalized Correlation and Mean Squared Distance. Two different scenarios are possible. The first one is to calculate similarity of each frame to a predefined fixed frame image. This assures best fitting to a fixed basis under the restriction of temporal and spatial smoothness to the neighbor images both controlled by the pairwise potential (Fig. 2). The second one is to measure the similarity to a regional neighborhood without a fixed image basis. This optimally would lead to higher transition quality (smoothness) as abrupt changes in terms of similarity are omitted. In practice, changes between temporal frames are very small. Thus, just concentrating on local similarity may not lead to an overall registered sequence. Therefore, the first scenario will be used to achieve registration to a common basis, followed by another iteration using local similarity to assure temporal smoothness not only in terms of the transformation parameters but also through the similarity measure.

B. MRF for segmentation

In the case of image segmentation potential functions are used to evaluate to which segment $s \in \mathcal{S}$ a pixel should be assigned. Therefore, segment regions must be established, either by the user or automatically by analyzing the intensity distribution function of the images. In both cases segment membership can be determined by finding the smallest value of

$$log\left(\sqrt{2\pi \cdot std}\right) + \frac{(p_i - avg)^2}{2 \cdot std} \tag{3}$$

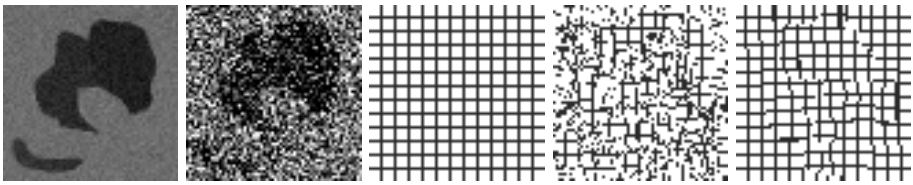
using the pixels intensity p_i, average intensity avg and standard deviation std of the predefined regions. However, this still means that noise influence will be represented in the segments. Therefore, edges in the MRF can be used to model the probability of pixels to belong to the same segment than the pixels in a defined neighborhood. Appropriate pairwise potentials have to be formulated, one of the simplest and straight forward being the Kronecker delta function:

$$\delta(i,j) = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j. \end{cases} \tag{4}$$

In this way segments are found with spatial continuity which can be defined by neighborhoods of different orders. Note that segments used here do not represent specific objects, organs or tissue. Instead, they represent a certain group of intensity values comparable to an image reconstruction scheme.

C. Combination of registration and segmentation using MRF

Both approaches (registration and segmentation) assume spatial and temporal continuity and can be combined in a single MRF. In general, the number of labels must be extended by combining all labels from registration (transformation parameter) with those from segmentation (segment label) leading to $l_k = (t_x, t_y, s)$. This includes calculation of pairwise



(a) (b) (c) (d) (e)

Fig. 3: Single frames of the simulated (artificial) dataset with SNR=3 (a) and SNR=1 (b) are shown. The grid image from (c) is deformed with parameters generated without spatial and temporal constraints (d) and with the constraints enabled in the motion compensation method (e).

potentials of these newly created set of labels using weighted euclidean distance between transformation parameters and the Kronecker delta of segment labels:

$$\psi_p(v_i, v_j) = \lambda(v_i^l, v_j^l) + \omega \cdot \mu(v_i^l, v_j^l), \quad \langle v_i, v_j \rangle \in \mathcal{E} \tag{5}$$

$$\lambda(m, n) = d\left(\begin{pmatrix} m_{t_x} \\ m_{t_y} \end{pmatrix}, \begin{pmatrix} n_{t_x} \\ n_{t_y} \end{pmatrix}\right) / (2t_{max}) \tag{6}$$

$$\mu(s_1, s_2) = 1 - \delta(s_1, s_2) \tag{7}$$

v_i^l and v_j^l being the current labels of the respective nodes. $d(\cdot, \cdot)$ is the euclidean distance of two vectors and t_{max} is the maximal translation parameter in one direction. ω is a weighting parameter between the transformation and segmentation pairwise energy.

The MRF can be efficiently solved using the graph-cut based α-expansion algorithm [21]. A faster method has been proposed by Komodakis et al., which is based on α-expansion and guarantees the same output [20]. It uses the duality principle of linear programming to generate a more generalized view to the problem. Both algorithms reach the optimal solution for a two-label problem. For multiple labels it is shown that deviation of the energy from that of the optimal solution is bounded [22].

A particular challenge is the memory allocation for the MRF data as it may contain over 50 million nodes. For large datasets the MRF has to be divided in several parts which are solved independently. The different parts have to be overlapped with neighboring parts to guarantee temporal smoothness over the borders of the single parts.

IV. First results

This chapter covers first results, which have been achieved using artificial data and patient datasets. An artificial dataset with a spatial resolution of 64 $\times$ 64 pixels and 5 time frames has been produced. Continuous transformations over space and time are applied using B-Splines with known parameters. Gaussian noise is added to the data to simulate uncertainty in the data. Two different signal-to-noise (SNR) levels are used[2], a SNR of 3 (Fig. 3a) and 1 (Fig. 3b).

TABLE I: Deviation in pixel of transformation parameters from landmark locations in x and y-direction using constrained and unconstrained registration for the artificial image with a SNR of 3.

	without constraints			**with** constraints		
Landmark	x	y	x,y	x	y	x,y
L1	0.2	0.0	0.1	0.6	0.0	0.3
L2	0.8	0.4	0.6	0.8	0.6	0.7
L3	0.6	0.4	0.5	0.4	0.2	0.3
average			**0.40**			**0.43**

TABLE II: Deviation in pixel of transformation parameters from landmark locations in x and y-direction using constrained and unconstrained registration for the artificial image with a SNR of 1.

	without constraints			**with** constraints		
Landmark	x	y	x,y	x	y	x,y
L1	1.8	0.6	1.2	1.2	0.6	0.9
L2	2.0	1.2	1.6	1.0	1.0	1.0
L3	1.6	0.6	1.1	1.4	0.2	0.8
average			**1.30**			**0.90**

The proposed method using similarity calculation in b-mode data to a predefined fixed frame is applied. Transformation parameters are compared at three given landmark positions in the artificial dataset to measure the accuracy of the system. This test was run twice for each SNR level, with and without spatial and temporal constraints. Accuracy results are shown in Tab. I and II for both SNR levels.

Results without spatial and temporal constraints used, perform equal to those calculated with the constraints for a SNR of 3. This is an expected result, as the artificial data with a SNR of 3 is simple registration task, because edges are

[2]The SNR has been determined by the ratio betwenn the average intensity and the standard deviation in the images.

TABLE III: Standard deviation within regions of interest in b-mode data and perfusion curve smoothness within contrast data is measured for three patient datasets before registration, after classic pairwise registration [18] and registration with the presented MRF-based approach. Improvement compared to values before registration are indicated in percent.

dataset	no. 1	no. 2	no. 3	overall
Standard deviation in b-mode data				
before registration	20.8	21.9	27.7	
classic pairwise registration	19.7 (6.1 %)	18.1 (19.2 %)	25.4 (6.7 %)	10.6 %
new MRF-based registration	18.8 (8.9 %)	16.1 (29.0 %)	23.4 (16.4 %)	18.1 %
perfusion curve smoothness in contrast mode data				
before registration	3.4	2.7	3.6	
classic pairwise registration	3.3 (0.7 %)	2.7 (0.8 %)	3.5 (3.0 %)	1.5 %
new MRF-based registration	3.3 (1.5 %)	2.6 (5.9 %)	3.5 (2.3 %)	3.2 %

clearly visible despite the noise. However, considering the result of the overall deformation field, the constrained registration approach produces considerably smoother parameters (compare Fig. 3d and 3e). This is important as a deformation field with smooth transitions can be explained physiologically. Also, the deformation field is unaltered in areas where only noise is present. This is an important observation, because if only unreliable information about motion can be retrieved, the local transformation should rather be zero than arbitrary. Considering the results generated for the artificial dataset with a SNR of 1, the spatially and temporally constrained approach outperforms the un-constrained approach, although the accuracy is worse compared to those of the dataset with less noise. This indicates that constraining the transformation parameters and using coupled segmentation stabilizes the search for true transformation parameters if a high noise level is present.

The second evaluation is targeted on three different patient datasets. The method was used without coupled segmentation, as this feature has to be adapted to work with the large patient datasets with over 500 temporal frames. First, the variation of intensities is measured within three different regions of interest in the b-mode sequence. These regions are defined for each registered temporal region [7]. They are chosen such that they represent the main areas of interest (e.g. Fig. 1c). The standard deviation is used as quality measure for evaluation. The variation of pixel intensities has to drop as a consequence of registration. Second, the perfusion time curves from the contrast sequence within the same regions of interest have been computed. The smoothness, being a physiological explanation of validity of perfusion, is measured in terms of the averaged absolute differences (MAD) between neighboring time points. This is a first indicator of improved contrast signal correspondence over time, although this signal is still disturbed by noise and speckle artifacts and thus contains enough potential to distort the time course of contrast enhancement. The MAD should decrease with improved registration quality.

These experiments are performed before registration and after registration with the proposed method. Example results are shown in Fig. 4a, 4b and 4c. Additionally, a previously published, classic registration approach is tested as well [18]. Results of this experiment are shown in Tab. III. They indicate that the newly presented approach using MRF registration is able to produce less variation in terms of intensity compared to a standard registration scheme incorporating pairwise frame registration without temporal dependency [18]. The MRF-based approach achieves an overall improvement for the three patient datasets of 18.1 % compared to 10.6 % if the pairwise image registration is used. The curve smoothness also leads to superior results, 3.2 % compared to 1.5 %.

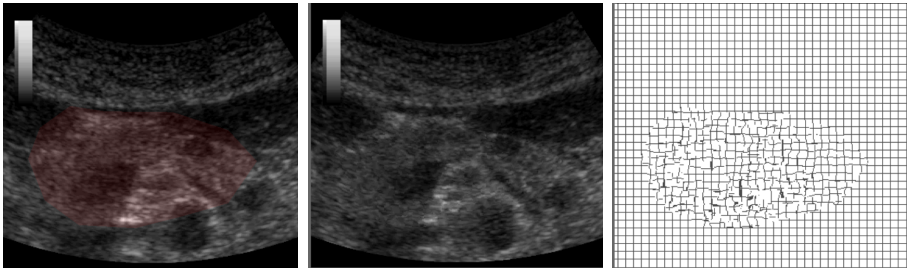
V. Conclusion

An approach for motion compensation of ultrasonic perfusion image sequences has been presented in this work. It takes into account the conditions of the specific image acquisition and includes them as a priori knowledge in a MRF formulation. This a priori knowledge in form of spatial and temporal smoothness of the deformation field of the images in the sequence is combined with another a priori term, the segment labeling which is supposed to take similar labels at corresponding locations over time. The results presented in this work show that both preconditions lead to more robustness against the noise influence. Within this well constrained system, a solution is obtained using a graph-cut based state-of-the-art optimization technique.

Open tasks in this project are the adaption of the combined method to work with large patient datasets and the performance investigation of the different features (spatial and temporal constraints and the coupled segmentation) to analyze their contribution. Additionally, a special focus will be put on the weighting between transformation constraints and segmentation label constraints and how robust the results are against small changes of these parameters. A new evaluation is planned as well, where manually defined landmarks in patient datasets are used to assess the quality of the calculated transformations.

Acknowledgment

The author would like to thank Kim Nylund and Odd Helge Gilja from the University of Bergen, Norway for providing the medical data.



(a) (b) (c)

Fig. 4: A single frame of a patient dataset with a ROI to be registered marked in red (a), the same frame motion corrected with a linear registration and a free-form deformation approach (b) together with the deformation grid image (c).

REFERENCES

[1] M. Postema and O. H. Gilja, "Contrast-enhanced and targeted ultrasound," *World Journal of Gastroenterology*, vol. 17, no. 1, p. 28, 2011.

[2] D. Klein, M. Jenett, H. Gassel, J. Sandstede, and D. Hahn, "Quantitative dynamic contrast-enhanced sonography of hepatic tumors," *European Radiology*, vol. 14, no. 6, pp. 1082–1091, 2004.

[3] K. Nylund, S. Ødegaard, T. Hausken, G. Folvik, G. A. Lied, I. Viola, H. Hauser, and O. H. Gilja, "Sonography of the small intestine," *World Journal of Gastroenterology*, vol. 15, no. 11, p. 1319, 2009.

[4] M. Claudon, D. Cosgrove, T. Albrecht, L. Bolondi, M. Bosio, F. Calliada, J. Correas, K. Darge, C. Dietrich, M. D'Onofrio *et al.*, "Guidelines and good clinical practice recommendations for contrast enhanced ultrasound (CEUS)," *Ultraschall in der Medizin*, vol. 29, no. 1, pp. 28–44, 2008.

[5] G. Renault, F. Tranquart, V. Perlbarg, A. Bleuzen, A. Herment, and F. Frouin, "A posteriori respiratory gating in contrast ultrasound for assessment of hepatic perfusion." *Physics in Medicine and Biology*, vol. 50, no. 19, pp. 4465–80, 2005.

[6] J. V. Hajnal, D. J. Hawkes, and D. L. G. Hill, *Medical image registration*, 2001.

[7] S. Schäfer, P. Angelelli, K. Nylund, O. H. Gilja, and K. Tönnies, "Registration of ultrasonography sequences based on temporal regions," in *Proc. of 7th Intl. Symp. on Image and Signal Processing and Analysis (ISPA 2011)*, Dubrovnik, Croatia, 2011, pp. 749–759.

[8] M. Ledesma-Carbayo, J. Kybic, M. Desco, A. Santos, and M. Unser, "Cardiac motion analysis from ultrasound sequences using non-rigid registration," in *Medical Image Computing and Computer-assisted Intervention (MICCAI)*, vol. 32, no. 4, 2010, pp. 889–896.

[9] R. Shekhar and V. Zagrodsky, "Mutual information-based rigid and nonrigid registration of ultrasound volumes," *IEEE Transactions on Medical Imaging*, vol. 21, no. 1, pp. 9–22, 2002.

[10] M. Ledesma-Carbayo, J. Kybic, M. Desco, A. Santos, M. Suhling, P. Hunziker, and M. Unser, "Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation." *IEEE Transactions on Medical Imaging*, vol. 24, no. 9, pp. 1113–26, 2005.

[11] S. Li, "Markov random field models in computer vision," in *Computer Vision ECCV*, vol. 801. Springer, 1994, pp. 361–370.

[12] D. Mahapatra and Y. Sun, "Integrating segmentation information for improved mrf-based elastic image registration," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 170–83, Jan 2012.

[13] P. Wyatt and J. Noble, "MAP MRF joint segmentation and registration of medical images," in *Medical Image Computing and Computer-assisted Intervention (MICCAI)*. Springer, 2002, pp. 580–587.

[14] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 48, no. 3, pp. 259–302, 1986.

[15] C. Xiaohua, M. Brady, and D. Rueckert, "Simultaneous segmentation and registration for medical image," in *Medical Image Computing and Computer-assisted Intervention (MICCAI)*. Springer, 2004, pp. 663–670.

[16] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm." *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[17] J. L. Marroquin, E. Santana, and S. Botello, "Hidden markov measure field models for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1380–1387, Nov. 2003.

[18] S. Schäfer, K. Nylund, O. H. Gilja, and K. Tönnies, "Motion compensation of ultrasonic perfusion images," in *Medical Imaging 2012: Ultrasonic Imaging, Tomography, and Therapy*, vol. 8320, no. 1. SPIE, 2012.

[19] Y. Boykov, V. Lee, H. Rusinek, and R. Bansal, "Segmentation of dynamic nd data sets via graph cuts using markov models," in *Medical Image Computing and Computer assisted Intervention (MICCAI)*. Springer, 2001, pp. 1058–1066.

[20] N. Komodakis, G. Tziritas, and N. Paragios, "Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies," *Computer Vision and Image Understanding*, vol. 112, no. 1, pp. 14–29, Oct 2008.

[21] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1–18, 2001.

[22] N. Komodakis and G. Tziritas, "Approximate labeling via graph cuts based on linear programming." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1436–53, Aug. 2007.

Be-greifbare Magische Linsen auf & über Tabletops

Martin Spindler
Institut für Simulation und Grafik
Fakultät für Informatik
Otto-von-Guericke Universität, Magdeburg
spindler@ovgu.de

***Abstract*—Aktuelle technische Entwicklungen am Smartphone- und Tablet-Markt zeigen, dass die Verschmelzung von Eingabe und Ausgabe im selben Gerät zu einer direkteren und als natürlicher empfundenen Interaktion führt. Während sich ein Großteil bisheriger Forschung der Entwicklung von Touch-Techniken auf derartigen berührungsempfindlichen Displays widmet, löse ich mich im Rahmen meines Dissertationsvorhabens von den Beschränkungen einer interaktiven Oberfläche und erweitere den Interaktionsraum auf den physischen dreidimensionalen (3D) Raum oberhalb eines Tabletops. Mit lagebewussten, in der Hand gehaltenen Papierdisplays stelle ich einen vielversprechenden Ansatz vor, der diesen Raum nutzt. Durch die zusätzliche Verwendung der Position und Orientierung dieser Papierdisplays (Linsen) mit unterschiedlicher Form und Größe können Nutzer eine sehr direkte, greifbare Interaktion mit verschiedenen Informations-räumen erleben. Der gleichzeitige Einsatz multipler Linsen unterstützt kollaboratives Arbeiten explizit. Dieser Artikel stellt neben grundlegenden Interaktionskonzepten auf und mit den Linsen auch konkrete Anwendungsfälle vor, diskutiert die Vor- und Nachteile aktiver und passiver Displays und erläutert den technischen Aufbau des Systems.**

Tangible Interaction, PaperLens, mobile Displays, Multitouch, Stifteingabe, Gesten, 3D-Interaktion, Kollaboration

I. EINLEITUNG

Mit der Markteroberung von leistungsfähigen Tablets und Smartphones hat sich binnen weniger Jahre ein neues Interface-Paradigma (Post-PC) durchgesetzt, das die Art und Weise stark verändert hat, wie wir mit Computern interagieren. Ein augenscheinliches Merkmal solcher Post-PC-Geräte ist das Fehlen von zusätzlicher Eingabe-Hardware, wie beispielsweise Tastatur und Maus. Stattdessen präsentieren sich diese Geräte nahezu vollständig als Displays, die wir oft als ständige Begleitung mit uns führen - sei es für die tägliche Arbeit oder für die Freizeit. Die Interaktion bleibt hierbei jedoch meist *auf* die Oberfläche der Displays beschränkt, d.h. visuelle Elemente werden durch direktes Zeigen ausgewählt oder manipuliert, beispielsweise mit den Fingern oder einem Stift. Im Gegensatz dazu spielt die Interaktion *mit* den Geräten, etwa durch die Ausnutzung der räumlichen Lage und Orientierung der Geräte, eher eine untergeordnete Rolle. Aktuelle Trends zeigen darüber hinaus, dass die Kombination von mobilen Displays untereinander oder sogar mit größeren stationären Displays (wie HD-Fernseher oder Tabletops) aufregende neue Möglichkeiten bietet - nicht nur im Sinne eines vergrößerten Darstellungsraums, sondern vor allem auch als erweiterter Interaktionsraum. Das Ziel meines Dissertationsvorhabens ist, solche zusätzlichen Freiheitsgrade besser nutzbar zu machen, so dass Nutzer direkter und natürlicher mit komplexen Informationsräumen interagieren können. Aufgrund der Neuheit des Ansatzes, war es zunächst nötig, ein technisches Framework aufzubauen, das eine praktische Umsetzung und Evaluierung der konzipierten Interaktionstechniken ermöglicht.

Mit dem PaperLens System [10] habe ich im Rahmen meines Promotionsvorhabens solch ein System entwickelt. Es macht das Konzept der lagebewussten greifbaren Displays (Tangible Magic Lens) für eine Tabletop-Umgebung verfügbar. Mehrere Nutzer können gleichzeitig mit diesem System interagieren - einfach nur indem sie ein Display greifen und es dann durch den dreidimensionalen (3D) Raum auf und über dem Tisch bewegen. Im Folgenden soll dieses System näher vorgestellt werden. Dazu werden zunächst verwandte Arbeiten umrissen und die wesentlichen Eigenschaften zweier grundverschiedener Typen von lagebewussten Displays erörtert. Anschließend werden einige der von mir entwickelten Interaktionskonzepte skizziert. Danach wird der technische Aufbau des Systems beschrieben und verschiedene Anwendungsszenarien diskutiert. Gesammelte Erfahrungen, aktuelle Entwicklungen und ein Fazit runden den Beitrag ab.

II. VERWANDTE ARBEITEN

In diesem Abschnitt wird ein grober Überblick über vorangegangene Arbeiten zu lagebewussten mobilen Displays gegeben. Das Ziel ist hierbei, bisherige Interaktionstechniken zu beleuchten und grundlegende technische Ansätze für passive (projizierte) Displays zu skizzieren.

A. Lagebewusste Displays

Die Vereinigung der digitalen Welt mit der physischen (analogen) Welt ist die Vision des Ubiquitous Computing, wie sie von Weiser bereits 1991 vorgeschlagen wurde [32]. Dieses Konzept wurde von Ishii und Ullmer mit den „Tangible User Interfaces“ (TUI) konsequent fortgeführt [25], indem die Interaktion mit digitalen Informationen auch durch die physische Manipulation von Alltagsgegenständen ermöglicht wurde. Inspiriert durch die Idee der „See-Through Interfaces“ [20], wo kontext-abhängige (lokale) Sichten und Werkzeuge zwischen Nutzer und Anwendung auf einem Desktop-Monitor platziert werden, können diese Alltagsgegenstände auch lagebewusste physische Displays (z.B. Mobiltelefone) sein. Diese fungieren als greifbare Magische Linsen in eine virtuelle Welt. Einer der ersten Vertreter solcher mobilen Displays war der Chameleon-Prototyp von Fitzmaurice [21], der einen lagebewussten Palmtop Computer für das Explorieren von virtuellen Informationsräumen vorstellte, die an beliebige Objekte in einer Büroumgebung geknüpft waren.

Im Gegensatz zum Ansatz von Fitzmaurice macht das metaDESK – Projekt von Ullmer and Ishii [31] nur von einem Referenz-Objekt Gebrauch: einem Tabletop, der kontextuelle

Informationen (z.B. die Karte des MIT-Campus) zeigt. Hier können Nutzer frei durch polygonale 3D-Modelle navigieren (z.B. verschiedene Gebäude auf dem Campus), indem sie ein an einem Greifarm befestigtes LCD-Display durch den Raum über dem Tabletop bewegen. In anderen Arbeiten stellten sowohl Hirota & Saeki [23] als auch Konieczny et al. [28] technische Ansätze für das Slicing von 3D Volumendaten mittels lagebewusster, handgehaltener Displays vor. Sie verwendeten dabei den (horizontalen) Fußboden als Bezugsfläche, nicht jedoch einen Tabletop.

Die Peephole Displays von Yee [33] kombinieren die Navigation eines virtuellen 2D-Arbeitsraums mit digitaler Stifteingabe. Der 2D-Arbeitsraum ist hierbei zylinderförmig um den Nutzer herum angeordnet und kann durch das Bewegen einen PDAs, der als Guckloch in diesen Arbeitsraum dient, exploriert werden. Dabei wird der Abstand zwischen Nutzer und PDA explizit für die Interaktion verwendet, beispielsweise für Zoom-Aufgaben im Kontext von typischen Desktopanwendungen, wie einer Terminverwaltung, einem Webbrowser und geografischen Karten.

B. Tracking von lagebewussten Displays

In der Literatur sind unterschiedliche Ansätze für das Tracking von lagebewussten Displays zu finden. Visuelle Marker wurden von Bandyopadhyay, Raskar und Fuchs [19] verwendet. Ein Hough-Transform-basierter Ansatz für das markerlose optische Tracking einer Bedienplatte wurde von Zhang et al. [34] vorgestellt. Markerbasiertes Infrarot (IR) Tracking wurde unter anderem von Holman et al. [24] eingesetzt.

C. Passive Displaylösungen

Passive (projizierte) Displays wurden unter anderem in PaperWindows von Holman et al. [24] präsentiert, die traditionelle GUI Elemente auf in der Hand gehaltene Papiere projizieren. Ein System für faltbare bewegliche Displays wurde von Lee, Hudson und Tse 2008 [30] vorgestellt. Beide Systeme nutzen einen ähnlichen technischen Ansatz, den auch ich in abgewandelter Form für mein System verwende, d.h. ein Deckenprojektor projiziert dynamische Bildinhalte auf im Raum getrackte handgehaltene Projektionsmedien. Im Gegensatz dazu ist das SecondLight – System von Izadi et al. [26] technisch anspruchsvoller. Es basiert auf elektronisch umschaltbaren Diffusern, die nahtlos den Grad ihrer Transparenz verändern können (je nach angelegter Spannung). Damit ist es in einer Tabletop-Umgebung möglich, von unterhalb der Tischplatte Bildinhalte sowohl auf die Tischplatte als auch auf Tangible Magic Lenses darüber zu projizieren. Aufgrund technischer Einschränkungen kann dabei der Tisch jedoch nicht viel größer als 40 cm in der Diagonale sein.

D. Fazit für das Dissertationsvorhaben

Der zentrale Beitrag meines Dissertationsvorhabens ist die Übertragung und konsequente Weiterentwicklung der Interaktionskonzepte für lagebewusste Displays *auf eine Tabletop-Umgebung*. Hierbei ist das Ziel, den physischen 3D-Raum über einer (interaktiven) Tischplatte für die Interaktion direkt nutzbar zu machen. Ein wesentlicher Beitrag meiner Forschung ist daher die Entwicklung eines vereinheitlichenden konzeptionellen sowie auch technischen Frameworks, welches die vielfältigen Interaktionsfreiheitsgrade, die sich dadurch ergeben, zusammenfasst. Dazu zählen neben der Lage der mobilen Displays, auch Stift- und Touch-Eingabe sowie die Kopflage der Nutzer. Solch ein systematisches Framework gab es bisher in dieser Form noch nicht. Es ermöglicht es erst, komplexere Anwendungen zu entwickeln und zu evaluieren und dabei das Zusammenspiel der verschiedenen Interaktionsmodalitäten zu untersuchen.

III. Passive und aktive Displays

Zunächst einmal möchte ich zwischen zwei grundlegenden Arten von lagebewussten Displays unterscheiden: *passive* und *aktive Displays*. Im Folgenden werden ihre wichtigsten Eigenschaften gegenübergestellt. Diese sind prägend sowohl für die technische Realisierung als auch für die Interaktionskonzepte.

A. Passive Displays

Passive Displays sind aus Papier, Pappe, Acrylglas, Porzellan, Holz oder anderen Materialien gefertigt. Dies schließt auch Alltagsgegenstände, wie Kaffeetassen, Spielkarten oder den Küchentisch ein. Die Displayfunktionalität wird hierbei durch die Projektion der relevanten Information auf die Oberfläche der passiven Displays mittels Beamer sichergestellt. Ein wesentlicher Vorteil von passiven Displays ist ihre Flexibilität hinsichtlich der Formfaktoren. Sie können sehr dünn und leicht sein (z.B. durch Verwendung von Pappe), weisen keine störenden Displayrahmen auf, sind billig und sind typischerweise einfach herzustellen. Außerdem können sie Bildinhalte auf der Vorder- und Rückseite darstellen, erlauben beliebige Formen (z.B. Scheiben) und können ihre Form und Größe verändern, wie beispielsweise von Khalilbeigi et al. mit einem ausrollbaren Papier-Display [27] demonstriert. Darüber hinaus können passive Displays leicht in die dritte Dimension erweitert werden (z.B. als Zylinder oder Würfel [16]).

Als Kehrseite weisen passive Displays durch die Abhängigkeit vom konkreten Projektionsvolumen eine begrenzte Mobilität auf. Dies liegt daran, dass sie nur in technisch komplexen Umgebungen funktionieren, die meist stationär sind. Solche Umgebungen sind sowohl für die präzise Lagebestimmung der passiven Displays als auch für die Projektion notwendig. Passive Displays leiden oft auch an einer geringen Bildauflösung und einer wahrnehmbaren Verschiebung von Bild- und Objektraum. Darüber hinaus können Verdeckungen (Schatten) ein Problem darstellen.

B. Aktive Displays

Die Verwendung von aktiven Displays, z.B. Smartphones und Tablets, löst viele dieser Nachteile. Sie weisen eine hohe Displayqualität auf (z.B. das Retina-Display des iPads) und benötigen kein kompliziertes Projektionssetup. Dies bedeutet wiederum, dass das Tracking der Geräte nur für die Interaktion benötigt wird und somit weniger akkurat sein kann, da nicht millimetergenau auf das Display projiziert werden muss. Dies kann ein entscheidender Vorteil sein, da es den Einsatz von Tracking-Technologie erlaubt, die weniger aufdringlich ist (z.B. Verzicht auf Marker). Ein anderer nicht zu unterschätzender Vorteil von aktiven Displays ist die Unterstützung von präzisen Touch-Eingaben direkt auf den Geräten. Darüber hinaus weisen aktive Displays oft zusätzliche Sensorik auf, wie Beschleunigungssensoren oder einen Kompass, die sich für die Interaktion nutzbar machen lassen.

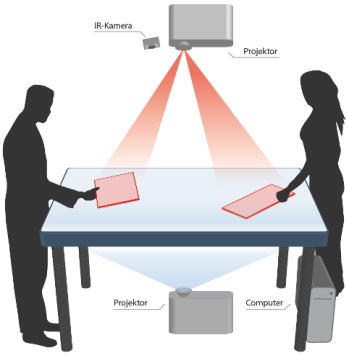


Bild 1: Prinzipieller Aufbau des Tangible Magic Lens - Systems. Der Tabletop bildet das Zentrum des Systems und bietet globale Sichten in verschiedene virtuelle Informationsräume. Ein oder mehrere mobile Displays zeigen lokale (persönliche) Sichten in diese virtuellen Informationsräume, die eine physische Ausdehnung einnehmen können, z.B. über und unter dem Tisch.

Trotz all dieser Vorteile sind aktive Displays häufig unflexibel in Bezug auf die Formfaktoren. Sie sind dicker und schwerer, besitzen störende Displayrahmen, sind weniger variabel in der Form und besitzen meist nur ein Display auf der Vorderseite. Obwohl sich dies mit dem technischen Fortschritt in Zukunft ändern könnte, ist eine nahtlose Integration von Alltagsgegenständen (wie Küchentellern oder Papierdokumenten) in die digitale Welt kaum möglich, wenn nur aktive Displays berücksichtigt werden.

C. *Fazit für das Dissertationsvorhaben*

Trotz einiger Vorteile von aktiven Displays, habe ich mich beim Design des technischen Systems für eine passive Displaylösung entschieden. Verglichen mit einem aktiven Displayansatz (z.B. iPad) waren hierfür zwar anfangs schwierigere technische Hürden zu meistern, dafür bietet die Verwendung von projektiven Displays eine erheblich größere Freiheit im Design der Formfaktoren – ein entscheidender Vorteil für einen Forschungsprototypen.

IV. INTERAKTIONSKONZEPTE

In diesem Abschnitt wird der konzeptionelle Aufbau des Tangible Magic Lens Systems vorgestellt und einige von mir entwickelten Interaktionstechniken grob skizziert, die die räumliche Lage und Orientierung von handgehaltenen Displays in einer Tabletop-Umgebung ausnutzen.

A. *Prinzipieller Aufbau*

Der prinzipielle Aufbau des Tangible Magic Lens Systems ist in Bild 1 dargestellt. Ein interaktiver Tisch bildet das Zentrum dieses Systems, das von einem oder mehreren Nutzern gleichzeitig verwendet werden kann. Das große Tischdisplay zeigt dabei globale Sichten in verschiedene virtuelle Informationsräume. Diese virtuellen Informationsräume können räumlich auf das physische Volumen über und unter dem Tisch ausgedehnt sein, innerhalb dessen der Tabletop als Sichtfenster platziert ist. Damit können sich Teile des Informationsraums oberhalb oder auch unterhalb der Tischoberfläche befinden. Ein oder mehrere mobile Displays können von jedem Nutzer in die Hand genommen bzw. auf dem Tisch abgelegt werden. Sie bieten lokale (persönliche) Sichten auf den jeweiligen virtuellen Informationsraum. Diese Sichten können sehr einfach durch das Bewegen, Halten und Rotieren der mobilen Displays im Raum manipuliert werden. Auf diese Weise werden Interaktionstechniken möglich, die die Exploration und Manipulation von großen, komplexen Informationsräumen natürlicher und intuitiver gestalten.

B. *Interaktionsvokabular*

Die Kombination aus einem stationären horizontalen Display (Tabletop) und mehreren in der Hand gehaltenen lagebewussten Displays bietet etliche Vorteile. Im Sinne einer Multi-Display-Umgebung ermöglicht sie die gleichzeitige Verwendung von globalen und lokalen (persönlichen) Ansichten, was kollaboratives Arbeiten explizit unterstützt. Darüber hinaus können durch die Bereitstellung von sehr direkten, be-greifbaren Interaktionstechniken (*engl.: Tangible Interaction*) verschiedene Informationsräume natürlicher erlebt werden, so dass Nutzer ihre Aufmerksamkeit stärker auf die Lösung ihrer Aufgaben lenken können. Neben der herkömmlichen Interaktion *auf* den mobilen Displays, z.B. über Touch- und Stift-Eingaben, werde ich daher im Folgenden hauptsächlich Techniken adressieren, die direkt *mit* den mobilen Displays arbeiten, d.h. ihre physische Lage im 3D-Raum auf und über dem Tisch ausnutzen. Dabei stehen sechs zusätzliche Freiheitsgrade (6DOF) zur Verfügung. Dies sind die Position und Orientierung der Displays im Bezug zu den drei Koordinatenachsen des Interaktionsraums, wobei der Interaktionsraum entweder *absolut* im Raum verankert sein kann (z.B. im Mittelpunkt der Tischplatte) oder *relativ* zu einer beliebigen Position und Orientierung im physischen Raum steht. Letzteres kann frei durch die Nutzer definiert werden, beispielsweise durch das Drücken und Halten einer Schaltfläche auf dem Display, wenn es sich an der gewünschten Position im Raum befindet. Die sich hieraus ergebenden Möglichkeiten sind vielfältig und gehen über die reine Translation der Linsen hinaus. Eine im Rahmen des Promotionsvorhabens entwickelte Klassifikation verschiedener Interaktionsmuster für mobile Displays [6] umfasst u.a. die in Bild 2 dargestellten Kategorien *Translation*, *Einfrieren*, *Rotation* und *Gesten*. Diese Interaktionsmuster sind größtenteils unabhängig voneinander und können somit leicht unterschiedlichen Interaktions-Teilaufgaben zugewiesen werden, was im Abschnitt Anwendungsszenarien noch ausführlicher betrachtet wird.

1) *Translation*

Die Position bzw. ihre Veränderung über die Zeit ist sehr gut für die Interaktion nutzbar. Entsprechend der drei Raumachsen stehen drei Freiheitsgrade zur Verfügung. Für unser System unterscheide ich zwischen *horizontaler* (entlang der X- und Y-Achse) und *vertikaler Translation* (entlang der Z-Achse), siehe Bild 2(a).

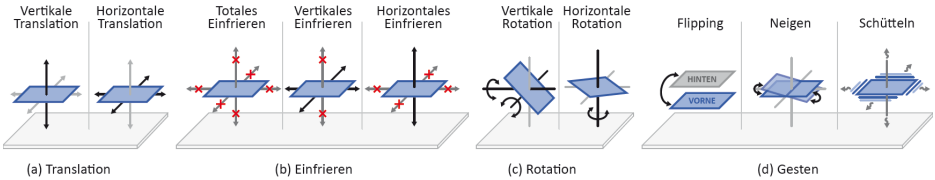


Bild 2: Auszug aus dem im Rahmen der Promotion entwickelten Interaktionsvokabular für Tangible Magic Lenses. Die Bilder zeigen die Interaktion *mit* den Displays (räumliche Lage und Orientierung sowie Gesten als Eingabe). Die Interaktion *auf* den Displays, wie z.B. durch Stift- und Touch-Eingabe, vervollständigt diese Kategorisierung, siehe auch [6].

2) Einfrieren

Gelegentlich möchten Nutzer ein mobiles Display bewegen, ohne die Absicht zu verfolgen, mit dem System zu interagieren. Dies kann beispielsweise notwendig sein, wenn ein Nutzer eine bestimmte Ansicht auf den Tisch legen möchte, um sie dann mit einem Stift zu annotieren [11]. Zu diesem Zweck müssen alle bzw. ausgewählte Raumachsen für die Interaktion ausgeschaltet bzw. *eingefroren* werden, siehe Bild 2(b). Ich unterscheide drei Fälle: *totales Einfrieren*, *vertikales Einfrieren* (Z-Achse blockiert) und *horizontales Einfrieren* (X- und Y-Achse blockiert).

3) Rotation

Die lokale Orientierung bzw. ihre Veränderung über die Zeit ist ein anderer wichtiger Interaktionsfreiheitsgrad. Ich unterscheide zwischen *horizontaler* (um die Z-Achse) und *vertikaler Rotation* (um die X- oder Y-Achse), siehe Bild 2(c).

4) Gesten

Neben der direkten Nutzung der Lageinformation können auch komplexere Bewegungsmuster (Gesten) als Eingabe für das System verwendet werden, siehe Bild 2(d). Für verschiedene Anwendungen nutze ich Flipping-Gesten (Umdrehen des Displays von der Vorder- auf die Rückseite und umgekehrt), Schüttelgesten und Neigegesten (leichtes kurzes Neigen des Displays nach links/rechts bzw. oben/unten).

V. Technischer Aufbau

Bei der technischen Umsetzung des Tangible Magic Lens Systems habe ich mich hauptsächlich auf den passiven Display-Ansatz konzentriert. Das System ist wie folgt aufgebaut (siehe Bild 1): Ein rückprojizierter digitaler Tisch bildet den Mittelpunkt des Systems. Darüber befinden sich mehrere, an die Zimmerdecke montierte, Infrarot (IR) -Kameras und ein auf den Tisch ausgerichteter Deckenprojektor. Als mobile Displays werden passive Displays verwendet. Das sind zumeist aus Pappe und Papier gefertigte Projektionsmedien, die von den Nutzern in die Hand genommen und frei im Raum auf und über dem Tisch bewegt werden können. Zu den technischen Problemen, die hierbei gelöst werden müssen, gehören das Tracking der Displays, das Projizieren von dynamischen Bildinhalten, das Erkennen von Bewegungsgesten, die Unterstützung von Touch- und Stifteingaben und Anwendungsfunktionalität. Viele dieser Aufgaben können unabhängig voneinander gelöst werden, so dass ich mich für eine verteilte Rechnerarchitektur entschieden haben. Für die Rechnerkommunikation habe ich offene Protokolle verwendet, z.B. VRPN für das Streaming von Gerätezuständen und XML-RPC für das Aufrufen von entfernten Funktionen.

A. Tracking von Papierdisplays

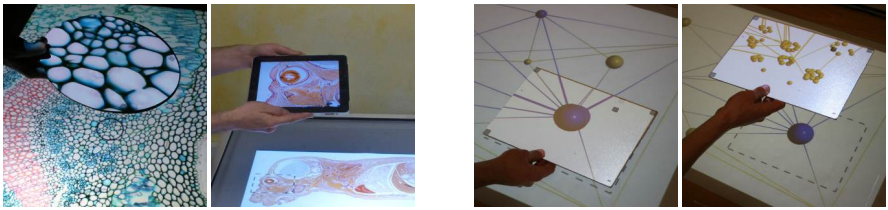
Eine exakte Bestimmung der Position und Orientierung der Papierdisplays in Echtzeit ist unerlässlich für das Gesamtsystem. Ich habe mich für ein optisches Verfahren mit momentan zehn IR-Kameras (Optitrack FLEX:V100R2) entschieden. Hierfür wurden jeweils sechs IR-reflektierende Marker (ca. 4 mm groß) auf die Kanten der Papierdisplays geklebt, die von den Benutzern kaum wahrgenommen werden. Durch die Verwendung von unterschiedlichen Marker-Kombinationen können die Papierlinsen leicht vom System auseinander gehalten werden. Eine kommerziell verfügbare Trackinglösung (Tracking Tools von Natural Point) liefert dann 6DOF-Informationen mit ca. 100 Hz in ausreichender Präzision (Fehlerrate ca. 1 mm).

B. Projektion von Bildinhalten

Die Papierlinsen sind als passive Displays (Projektionsmedien) umgesetzt, d.h. dynamische Bildinhalte werden über den Deckenprojektor auf sie projiziert, so dass auch ihre Rückseite als Display genutzt werden kann. Die Verwendung preiswerter Projektionsmaterialen wie Papier und Pappe fördert eine kostengünstige Herstellung der Displays in beliebigen Größen und Formen (z.B. rechteckig oder kreisförmig). Die Auflösung schwankt je nach Abstand zum Projektor zwischen 35 bis 50 Pixel/cm. Um eine perspektivisch korrekte Projektion auf den Papierlinsen zu gewährleisten, verwende ich OpenGL, mit dessen Hilfe ich den physischen Raum über dem Tabletop nachbilde. Die OpenGL-Kamera sitzt dabei an der virtuellen Position des Deckenprojektors, und die Papierlinsen werden als texturierte Polygone repräsentiert, die der jeweiligen Form, Position und Orientierung der physischen Papierlinse nachempfunden werden. Jeder dieser Texturen ist ein FrameBufferObject (FBO) zugewiesen, in das beliebige Bildinhalte in Echtzeit gerendert werden können. Auf diese Weise wird der generische Projektionscode von dem eigentlichen Applikationscode getrennt. Dies macht es beispielsweise einfacher, die passiven Displays später durch handgehaltene aktive Displays (z.B. Tablets) zu ersetzen.

C. Gestenerkennung

Für die Unterstützung von Gesten (wie Flipping-, Schüttel-, und Neigegesten) habe ich einen rudimentären Gestenerkenner



(a) Passive Zoomlinse (b) Aktive Zoomlinse (iPad) (c/d) Node-Link-Diagram (Semantischer Zoom)

Bild 3: Zoomlinsen. Durch simples Heben und Senken des passiven (a) oder aktiven Displays (b) kann in hochaufgelöste Bilder (a/b) oder in Node-Link-Diagramme (c/d) hinein- und hinaus-gezoomt werden. Die schwarze (kreisrunde bzw. viereckige) Konturlinie auf der Tischoberfläche markiert dabei das aktuell ausgewählte Detail und erleichtert so die Orientierung.

implementiert, der die Lageinformationen der einzelnen Papierdisplays nutzt. Der Gestenerkenner sucht kontinuierlich

nach charakteristischen Bewegungsmustern und löst gegebenenfalls ein Ereignis aus. Für Schüttelgesten sind das beispielsweise schnelle unregelmäßige Hin- und Her-Bewegungen mit kleiner Ausdehnung.

D. Stift- und Touch-Eingaben

Unser System unterstützt Touch- und Stifteingaben, sowohl auf dem Tabletop als auch auf den Papierdisplays. Digitale Stifte und die Anoto-Technologie [16] erlauben die Stiftinteraktion auf allen Oberflächen. Hierfür habe ich Anoto-Papier auf die einzelnen Papierdisplays geklebt, das ein eindeutiges Punktmuster enthält, welches durch spezielle digitale Stifte mit einer eingebauten Kamera gescannt wird. Die so ermittelte 2D-Position kann via Bluetooth in Echtzeit an das System übertragen und dann an das entsprechende Papierdisplay weitergeleitet werden. Für einfache Touch-Eingaben habe ich einige Papierdisplays mit speziellen drucksensitiven Buttons ausgestattet, die mit dem System drahtlos verbunden sind (Arduino-Xbee [17]).

VI. Anwendungsszenarien

Die beschriebenen Interaktionskonzepte können für eine Vielzahl von Anwendungsgebieten eingesetzt werden, z.B. in der Wissenschaft und Bildung. Im Folgenden werde ich drei Anwendungsszenarien vorstellen, die ich im Rahmen meines Dissertationsvorhabens prototypisch umgesetzt habe.

A. Exploration von 2D-Informationsräumen

Zweidimensionale Informationsräume sind allgegenwärtig, und es gibt sie in verschiedensten Ausprägungen. Typische Beispiele sind neben hochaufgelösten Fotografien auch Geografische Informationssysteme (GIS). Das sind georeferenzierte Karten, die mit zusätzlichen Informationsebenen wie Satellitenbildern, Höhenkarten, Straßenansichten (z.B. Google Streetview) oder sogar 3D-Modellen von Häusern angereichert sind.

In einer Tabletop-Umgebung bietet es sich an, das hochaufgelöste Bild oder die georeferenzierte Karte direkt auf dem Tisch zu zeigen. Allerdings bleibt es auch hier unmöglich, beispielsweise ein Gigapixel-Bild in allen Details darzustellen. Mithilfe von mobilen Displays, die als digitales Vergrößerungsglas dienen, können Nutzer sehr einfach in beliebige Bereiche des Gigapixel-Bildes hineinzoomen (siehe Bild 3a/b). Dies geschieht, indem sie das digitale Vergrößerungsglas zuerst an die entsprechende Stelle des Bildes bewegen (*horizontale Translation*) und dann das mobile Display heben und senken (*vertikale Translation*). Konzeptionell wird dieser Vorgang durch ein Space-Scale Diagramm [22] beschrieben, das direkt auf das physische Interaktionsvolumen abgebildet wird. Auf dieselbe Weise kann auch semantisches Zoomen umgesetzt werden, z.B. für das Explorieren von geclusterten Node-Link-Diagrammen (siehe Bild 3c/d).

Die Höhe über dem Tisch kann nicht nur für das Zoomen verwendet werden, sondern beispielsweise auch für das Auswählen von verschiedenen virtuellen Informationsschichten des Geografischen Informationssystems, die über dem Tisch physisch übereinander gestapelt sind. Diese nehmen eine Schichtdicke von beispielsweise 10 cm an. In einer aktuellen Studie [3] habe ich gezeigt, dass hierbei mehr als zehn Ebenen übereinander nicht sinnvoll sind und dass vertikale Translation (minimale Schichtdicke: 1 cm) von Nutzern deutlich präziser (mit einem Faktor von vier) ausgeführt werden kann als horizontale Translation (minimale Schichtdicke: 4 cm).

B. Exploration von wissenschaftlichen, volumetrischen Daten

Ein naheliegendes Anwendungsgebiet für Tangible Magic Lenses ist die Exploration und Annotation von großen, dreidimensionalen Datensätzen. Neben geologischen und biologischen Daten können dies auch medizinische Volumendaten sein, wie sie häufig durch MRT und CT aufgenommen werden.

In einer kollaborativen Umgebung, die aus einem Tabletop und verschiedenen mobilen Displays besteht, bietet der Tisch einen räumlichen Bezug zum volumetrischen Datenraum, der sich auf, über oder sogar unter dem Tisch befinden kann. Der Tabletop dient dabei als globale Sicht und könnte beispielsweise den Umriss des Patienten oder einen bestimmten Schnitt durch das Volumen darstellen. Die mobilen Displays zeigen dagegen lokale, persönliche Sichten. Wenn z.B. eine Ärztin ein Display durch das Interaktionsvolumen bewegt, können beliebige nutzer-definierte Schnitteben in Echtzeit berechnet und dargestellt werden, siehe auch [10]. Dies ermöglicht auf sehr direkte Weise eine schnelle und

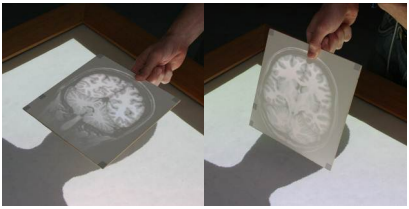


Bild 3: Volumen-Slicer. Durch beliebiges Bewegen, Neigen und Halten des Papierdisplays über dem Tisch können beliebige nutzer-definierte Schnittebenen durch einen Volumendatensatz (hier: MRT-Scan eines menschlichen Kopfes) in Echtzeit ausgewählt, berechnet und dargestellt werden.

flexible Erkundung des ganzen Datensatzes oder spezieller Strukturen darin (siehe Bild 4).

Es gibt vielfältige Einsatzszenarien, die von solch einem System profitieren würden. In einer Tumorboard-Besprechung arbeiten beispielsweise Mediziner mit verschiedener Spezialisierung zusammen, um Diagnosen zu stellen oder Therapien zu planen. Durch die Verwendung von persönlichen Displays können diese Mediziner gleichzeitig auf die Daten zugreifen. Dabei können entsprechend ihrer Spezialisierung maßgeschneiderte Sichten mit unterschiedlichem Abstraktionsgrad eingesetzt werden. Die Verwendung von hochaufgelösten aktiven Displays bietet dabei die notwendige Auflösung, um sogar noch die kleinsten Details der Patientendaten sichtbar zu machen. Zusätzlich können auf diese Weise auffällige Strukturen direkt mit den Fingern oder digitalen Stiften annotiert werden. Darüber hinaus bieten aktive Displays die Möglichkeit, beim Verlassen des Interaktionsraums die Daten mitzunehmen, um später (z.B. unterwegs) mit ihnen weiterzuarbeiten. Ein anderer Anwendungsfall sind Arzt-Patienten-Konsultationen. Hier können mobile Displays helfen, eine Diagnose zu präsentieren und die beabsichtigte Therapie zu erklären. Ähnliche Setups können auch zu Lehr- oder Trainingszwecken eingesetzt werden. Studenten können so beispielsweise interaktiv lernen, wie der menschliche Organismus aufgebaut ist.

C. *Exploration von Raum-Zeit-Würfeln*

Das Konzept des Raum-Zeit-Würfels integriert räumliche und zeitliche Aspekte in einer vereinheitlichten 3D Darstellung, siehe auch [29]. Die Analogie zwischen Raum-Zeit-Würfel und dem dreidimensionalen Interaktionsraum der Tangible Magic Lenses legt es nahe, den räumlichen Aspekt (z.B. eine geographische Karte) auf dem Tabletop darzustellen, während die Zeitdimension auf die Höhe darüber (Z-Achse) abgebildet wird.

Verschiedene mobile Displays dienen dann als greifbare Fenster in diesen Raum-Zeit-Würfel. Dessen interaktive Exploration wird dabei durch *horizontale Translation* (Navigation in der räumlichen Dimension, z.B. entlang einer Landkarte) bzw. *vertikale Translation* (Navigation durch die Zeit, z.B. die Monate eines Jahres) erreicht. Wird ein mobiles Display horizontal gehalten, zeigt es die Daten für einen bestimmten Zeitpunkt an einem Ort an (z.B. den Monat März

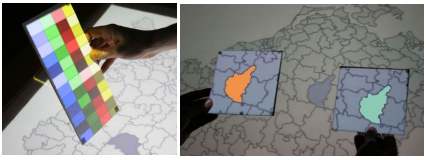


Bild 5: Raum-Zeit-Würfel. Mittels vertikalem Schneiden des Raum-Zeit-Würfels erhält der Nutzer eine Übersicht aller Zeitdaten zu einem Ort (links). Durch gleichzeitiges Heben und Senken zweier mobiler Displays, die auf denselben Ort eingerastet sind, kann der Nutzer beliebige Zeitpunkte vergleichen (rechts).

für den Bördelandkreis). Um eine Übersicht der Daten aller verfügbaren Monate für einen Ort zu erhalten, muss das mobile Display gekippt werden (*vertikale Rotation*), so dass es den Raum-Zeit-Würfel senkrecht schneidet, siehe Bild 5(links).

Bei der Exploration von Raum-Zeit-Daten müssen gewöhnlich verschiedene Orte, verschiedene Zeitpunkte oder beides in Kombination miteinander verglichen werden. Das *Einfrieren* eines mobilen Displays hilft hier den Nutzern, ihre Ziele einfacher zu erreichen. Durch *vertikales Einfrieren* kann ein bestimmter Zeitpunkt, z.B. der Monat Februar, festgelegt werden. Dies erlaubt es Nutzern, das mobile Display auf dem Tabletop abzulegen und dort mit der Exploration fortzusetzen, ohne dass sich der Monat ändert. Dies ist nützlich, wenn mehrere mobile Displays gleichzeitig verwendet werden sollen, um Attribute von verschiedenen Orten zu vergleichen oder um ein Detail für später zu sichern, zum Beispiel indem man es einfach irgendwo auf den Tisch platziert. Durch *horizontales Einfrieren* kann eine beliebige Region auf der Landkarte festgelegt werden, beispielsweise, um verschiedene Monate derselben Region zu vergleichen. Hierfür hält der Nutzer jeweils ein mobiles Display in jeder Hand, die er beide auf dieselbe Region einfriert. Durch unterschiedliches Heben und Senken der linken und rechten Hand können so verschiedene Monate ausgewählt und die Daten direkt visuell verglichen werden, siehe Bild 5(rechts).

VII. Erfahrungen und Ausblick

Während sich viele Projekte im Bereich von Tangible User Interfaces auf den Aspekt der verbesserten Eingabe über greifbare Objekte konzentrieren, besteht die Kernidee meines Dissertationsvorhabens darin, die Ausgabe (d.h. die in der Hand gehaltenen Displays) direkt manipulierbar zu gestalten. Durch das manuelle Bewegen eines oder mehrerer Displays durch den Raum sind neue Formen der Interaktion möglich, die ein echtes Be-greifen komplexer Informationsräume erlauben.

In einer aktuellen Weiterentwicklung des Systems mache ich Gebrauch von kopfgebundener Perspektive [14], die eine pseudo-stereoskopische Interaktion mit echt dreidimensionalen Objekten und Umgebungen erlaubt, siehe Bild 6. Durch die Kombination des physischen 3D-Raums mit einem virtuellen 3D-Raum unter Nutzung multipler Displays, die wiederum perspektivisch korrekte Sichten in eine 3D-Szene ermöglichen, wird ein noch reicheres Repertoire an 3D-Interaktionstechniken ermöglicht. Die im Bild 6(rechts) dargestellte direkte

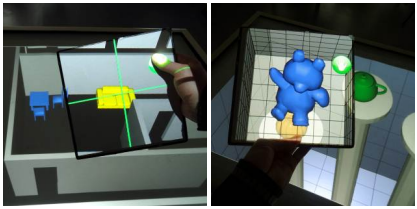


Bild 6: Ein mobiles Display (links) dient als Fenster in eine virtuelle 3D-Welt, die sich auf, über und unter dem Tabletop befindet. Durch Bewegen des Displays im Raum kann die virtuelle Welt exploriert werden. Ein Selektionsstrahl dient dabei zur Auswahl von entfernten Objekten. Diese können per Knopfdruck in die Hand bzw. auf das mobile Display geholt werden, um es sich dann von verschiedenen Seiten anzusehen (rechts). Die Verwendung von kopfgebundener Perspektive (mittels Kopftracking) erzeugt dabei einen pseudo-stereoskopischen Eindruck.

Erkundung eines mit der Linse assoziierten 3D-Objektes zeigt dieses Potential auf.

Das im Rahmen meines Dissertationsvorhabens entwickelte Tangible Magic Lens System wurde in zahlreichen Demo-Sessions und öffentlichen Präsentationen sehr vielen Besuchern präsentiert, z.B. zur Langen Nacht der Wissenschaft oder auf internationalen HCI-Konferenzen. Hierbei hat sich durchweg bestätigt, wie einfach das System zu erlernen und zu bedienen ist (gerade auch von Kindern). Da man als Nutzer des Systems nur ein Stück Pappe in der Hand hält, auf dem dynamische Bildinhalte je nach Raumposition und Anwendung gezeigt werden, wurde häufig kommentiert, wie natürlich und auf gewisse Weise magisch sich diese Form der be-greifbaren Interaktion anfühlt. Meine Erfahrungen aus Nutzerstudien [10] untermauern diese Aussagen. In diesem Zusammenhang wurde ich auch immer wieder gefragt, ob bzw. wann die von mir entwickelten Techniken für eine breitere Öffentlichkeit zugänglich wären.

Ich denke, dass in absehbarer Zukunft leichtgewichtige organische Displays (OLED) und unkomplizierte räumliche Trackinglösungen (z.B. durch Tiefenkameras) den hier vorgestellten Konzepten und Interaktionstechniken zu tatsächlich nutzbringender Anwendung verhelfen werden. In diesem Sinne arbeite ich gegenwärtig an einem kosten-günstigeren Setup [1], das mit leicht verfügbarer und erschwinglicher Hardware auskommt (z.B. iPad und Kinect). Neben der künftigen Nutzung leichter, aktiver Displays (Smartphones, Tablets) wird aber auch die Interaktion mit digital angereicherten Alltagsgegenständen, die durch Picoprojektion zu passiven Displays werden, verstärkte Anwendung finden.

Danksagung

Diese Arbeit wurde vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des ViERforES-II - Projekts (Nr. 01IM10002B) gefördert.

Eigene Publikationen

Publiziert/Angenommen

[1] Spindler, M; Dachselt, R.: Die Magische Dimension: Be-greifbare Interaktion auf und über Tabletops. i-com: Zeitschrift für Interaktive und Kooperative Medien. Wird in Ausgabe 2/2012 erscheinen.

[2] Spindler, M.; Büschel, W.; Dachselt, R.: Towards Spatially Aware Tangible Displays for the Masses. Workshop on Designing Collaborative Interactive Spaces for e-Creativity, e-Science and e-Learning at AVI '12. ACM Press, 2012.

[3] Spindler, M.; Martsch, M.; Dachselt, R.: Going Beyond the Surface: Studying Multi-Layer Interaction above the Tabletop. In Proceedings of the Conference on Human Factors in Computing Systems (CHI '12). ACM Press, 2012.

[4] Tominski, C.; Schumann, H.; Spindler, M.; Dachselt, R.: Towards Utilizing Novel Interactive Displays for Information Visualization. Dexis 2011 - Workshop on Data Exploration for Interactive Surfaces at ITS '11 (Dexis 11). ACM Press, 2011.

[5] Spindler, M.; Hauschild, M.; Dachselt, R.: Towards Making Graphical User Interfaces Tangible. In Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces 2010 (ITS' 10). ACM Press, 2010, 291-292.

[6] Spindler, M.; Tominski, C.; Schumann, H.; Dachselt, R.: Tangible Views for Information Visualization. In Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces 2010 (ITS '10). ACM Press, 2010, 157-166.

[7] Spindler, M.; Tominski, C.; Hauschild, M.; Schumann, H.; Dachselt, R.: Novel Uses for Tangible Displays above the Tabletop. In Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces 2010 (ITS '10). ACM Press, 2010, 315-315.

[8] Spindler, M.; Tominski, C.; Schumann, H.; Dachselt, R.: Towards Making InfoVis Views Tangible. In Conference USB Proceedings of IEEE Information Visualization Conference 2010 (InfoVis2010). IEEE, 2010.

[9] Spindler, M.; Dachselt, R.: Exploring Information Spaces by Using Tangible Magic Lenses in a Tabletop Environment. In Proceedings of the 28th of the International Conference Extended Abstracts on Human Factors in Computing Systems 2010 (CHI EA' 10). ACM Press, 2010, 4771-4776.

[10] Spindler, M.; Stellmach, S.; Dachselt, R.: PaperLens: Advanced Magic Lens Interaction Above the Tabletop. In Proceedings of ACM International Conference on Interactive Tabletops and Surfaces 2009 (ITS '09). ACM Press, 2009, 77-84.

[11] Spindler, M.; Dachselt, R.: Towards Pen-based Annotation Techniques for Tangible Magic Lenses Above a Tabletop. In Accompanying DVD of the ACM International Conference on Interactive Tabletops and Surfaces 2009 (ITS '09). ACM Press, 2009.

[12] Spindler, M.; Dachselt, R.: Advanced Magic Lens Interaction Above the Tabletop. TechDemo at ACM International Conference on Interactive Tabletops and Surfaces 2009 (ITS '09). 2009.

[13] Spindler, M.; Sieber, J.; Dachselt, R.: Using Spatially Aware Tangible Displays for Exploring Virtual Spaces. In Proceedings of Mensch und Computer 2009 (MuC2009). Oldenbourg Publishing, 2009, 253-262.

Eingereicht

[14] Spindler, M.; Büschel, W.; Dachselt, R.: Use Your Head: Tangible Windows for 3D Information Spaces in a Tabletop Environment. Submitted to ACM Interactive Tabletops & Surfaces (ITS) 2012.

[15] Spindler, M.; Cheung, V.; Witt, H.; Dachselt, R.: Dynamic Tangible User Interface Palettes. Submitted to ACM Interactive Tabletops & Surfaces (ITS) 2012.

Literatur

[16] Anoto Digital Pen Technology. Anoto Group AB. http://www.anoto.com

[17] Arduino. http://www.arduino.cc

[18] Akaoka, E.; Ginn, T.; Vertegaal, R.: DisplayObjects: Prototyping Functional Physical Interfaces on 3D Styrofoam, Paper or Cardboard Models. In Proc. of TEI, pages 49–56. ACM Press, 2010.

[19] Bandyopadhyay, D.; Raskar, R.; Fuchs, H.: Dynamic Shader Lamps: Painting on Movable Objects. In Proc. of ISAR'01. IEEE Computer Society, 2001, 207-216.

[20] Bier, E. A.; Stone, M. C.; Pier, K.; Buxton, W.; DeRose, T. D.: Toolglass and Magic Lenses: The See-through Interface, In Proc. of SIGGRAPH '93, ACM Press, 1993, 73-80.

[21] Fitzmaurice, G. W: 1993. Situated Information Spaces and Spatially Aware Palmtop Computers. Communications of ACM 36, 7. ACM Press, Juli 1993, 39-49.

[22] Furnas, G. W.; Bederson, B. B.: Space-Scale Diagrams: Understanding Multiscale Interfaces", In Proc. of CHI '95. ACM Press, 1995, 234-241.

[23] Hirota, K.; Saeki, Y.: Cross-section Projector: Interactive and Intuitive Presentation of 3D Volume Data using a Handheld Screen. *Proc. of 3DUI 2007*, IEEE Computer Society Press (2007), 57-63.

[24] Holman, D.; Vertegaal, R.; Altosaar, M.; Troje, N.; Johns, D.: Paper Windows: Interaction Techniques for Digital Paper. In Proc. of CHI '05. ACM Press, 2005, 591-599.

[25] Ishii, H.; Ullmer, B.: Tangible Bits: Towards Seam-less Interfaces between People, Bits and Atoms. *Proc. CHI 1997*, ACM Press (1997), 234-241.

[26] Izadi, S.; Hodges, S.; Taylor, S.; Rosenfeld, D.; Villar, N.; Butler, A.; Westhues, J.: Going Beyond the Display: A Surface Technology with an Electronically Switchable Diffuser. *Proc. of UIST '08*, ACM Press (2008), 269-278.

[27] Khalilbeigi M.; Lissermann R.; Mühlhäuser, M.; Steimle, J.: Xpaaand: Interaction Techniques for Rollable Displays. In Proc. of CHI '11, ACM Press, 2011, 2729-2732.

[28] Konieczny, J.; Shimizu, C.; Meyer, G.; Colucci, D.: A Handheld Flexible Display System. In Proc. of VIS '05. IEEE, 2005, 591- 597.

[29] Kraak, M. J.: The Space-Time Cube Revisited from a Geovisualization Perspective. In Proc. of the 21st Intern. Cartographic Conference. 2003, 1988-1996.

[30] Lee, J. C.; Hudson, S. E.; Tse, E.: Foldable Interactive Displays. In Proc. of UIST '08. ACM Press, 2008, 287-290.

[31] Ullmer, B.; Ishii H.: The metaDESK: Models and Prototypes for Tangible User Interfaces. In Proc. of UIST '97. ACM Press, 1997, 223-232.

[32] Weiser, M.: The Computer for the 21st Century. *Scientific American*. 265, 3 (1991), 66-75.

[33] Yee K.: Peephole Displays: Pen Interaction on Spatially Aware Handheld Computers. In Proc. of CHI '03. ACM Press, 2003, 1-8.

[34] Zhang, Z.; Wu, Y.; Shan, Y.; Shafer S.: Visual Panel: Virtual Mouse, Keyboard and 3D Controller with an Ordinary Piece of Paper. In Proc. of PUI '01. ACM Press, 2001, 1-8.

Verification of Software Product Lines Using Contracts

Thomas Thüm
School of Computer Science
University of Magdeburg
Magdeburg, Germany

***Abstract*—Software product lines are widely used to achieve high reuse of code artifacts for similar software products. While there are many efficient techniques to implement product lines, such as feature-oriented programming, the analysis and verification of product lines got only little attention so far. But as product lines are increasingly used in safety critical scenarios, efficient verification techniques are indispensable. We give an overview on the state-of-the-art in product-line verification, in which we classify approaches according to their strategy to scale specification and verification approaches known from single-system engineering. We propose to use contracts (i.e., preconditions and postconditions) to specify the intended behavior of a product line implemented with feature-oriented programming. Based on these contracts, we discuss different approaches to verify that all products of a product line fulfill its specification.**

***Keywords*-Software product lines, feature-oriented programming, design by contract, specification, verification**

I. Introduction

A major challenge in software engineering is to reduce the effort required to implement a certain functionality. In the last century, software engineering focused on reuse *within* one software system. For example, imperative programming encapsulates functionality in procedures to enable software reuse, whereas object-oriented programming provides more high-level reuse techniques such as class inheritance [1].

Software reuse *across* multiple software systems has gained much attention in the last decades [2]–[6]. The idea is to develop similar software systems not from scratch, but rather define the commonalities and variabilities between them in a software product line. A *software product line* (or short product line) is a set of software systems sharing a common code base [3]–[5]. The software systems (a.k.a. software products) are distinguished in terms of features. A *feature* is a prominent or distinctive user-visible aspect, quality, or characteristic of a software system [2]. We focus on feature-oriented programming for the implementation of product lines, in which each feature is implemented in a separate module [7], [8]. Given a particular selection of features, a customized product can be generated automatically by composing the corresponding modules [3], [6].

Another major challenge in software engineering is to verify the correctness of software systems. Especially safety-critical and mission-critical software systems need to be verified in order to prove the absence of failures. *Design by contract* is a methodology to formally specify object-oriented systems in terms of method contracts [9]. A *method contract* (or short contract) is assigned to each or at least each safety-critical method consisting of a *precondition* stating what the caller of a method needs to ensure and a *postcondition* stating what the caller can rely on. Contracts can be used to formally specify the intended behavior of a software system, which in turn can be used to verify that the software system fulfills its specification.

When product lines are used to implement safety-critical software systems, we need to apply specification and verification techniques from single-system engineering to product lines. A simple strategy is to specify and verify each product separately. Unfortunately, this strategy involves redundant effort for specification as well as for verification, because products of a product line have commonalities [10]. Furthermore, the number of products is up-to exponential in the number of implemented features, and thus the strategy infeasible for large product lines [10].

Our goal is to develop efficient verification techniques for product lines implemented with feature-oriented programming. In order to verify that a product line is correct, the intended behavior need to be specified efficiently. We propose to use contracts to formally specify feature-oriented programs [11]. We assign contracts to each feature from which the specification of each product can be derived automatically [11]. Based on contracts for each feature, we discuss different approaches ranging from testing to static analysis [12] and theorem proving [13], [14], which can verify that the implementation of each feature fulfills its contracts. Each approach has strengths and weaknesses regarding soundness, completeness, and efficiency. We summarize preliminary results on the scalability of these approaches.

II. Background

We present basic concepts that are necessary to understand our remaining discussion. We give a short overview on software product lines and feature-oriented programming.

A. Software Product Lines

A software product line is a set of products defined on a set of features F. A software product P is as a subset of all features $P \subseteq F$. Theoretically, we can combine the features in F in all combinations defined by the power set 2^F.

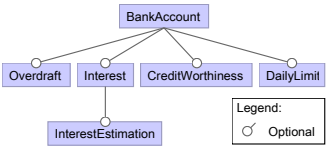


Figure 1. A feature model of a product line of bank accounts.

Practically, features may require other features or features may exclude each other [2], [3], [15]. For example, in a product line of database management systems, we may have support for different operating systems such as Windows and Unix, and a database product can either run on Windows or Unix. Hence, a software product line L is defined as a subset of all possible feature combinations: $L \subseteq 2^F$.

The formalization of a product line as a set of sets of features is often laborious. A common means to compactly describe valid feature combinations is a feature model [2], [3], [15]. A *feature model* is a hierarchy of features [2], in which each feature requires its parent feature [15]. A subfeature can be either mandatory, optional, or part of a group of alternative features [2].

In Figure 1, we give an example describing the features of a product line to manage bank accounts. The product line consists of six features. Feature *BankAccount* is part of all products and all other features are optional. Depending on the feature selection, a bank account may or may not support an overdraft limit (feature *Overdraft*), the calculation of interests (*Interest*), the estimation of credit worthiness (*CreditWorthiness*), or a maximum daily withdrawal (*DailyLimit*). Furthermore, feature *InterestEstimation* provides a calculation of the estimated interest until the end of the year. But, this feature requires feature *Interest* and is therefore modeled as a subfeature of it. Overall, the feature model defines a product line with 24 products.

B. Feature-Oriented Programming

So far, we discussed how to define valid combinations of features. Once these valid combinations have been defined, we need a product-line implementation technique that maps features to implementation units, in order to automatically generate software products for a given feature selection. While there are several implementation techniques [6], we focus on feature-oriented programming.

In feature-oriented programming, the code of each feature is encapsulated in a distinct feature module [7], [8]. A *feature module* is a set of classes and class refinements [8], in which a *class refinement* is a set of methods and fields. When composing class A with a class refinement A', all methods and fields of A' are added to A and existing methods are refined. Software products can be generated automatically by composing different combinations of feature modules.

feature module *BankAccount*

```
class Account {
  int balance = 0;
  void update(int x) {
    balance += x;
  }
}
class Application {
  Account account = new Account();
  void nextDay() {}
  void nextYear() {}
}
```

feature module *DailyLimit*

```
refines class Account {
  final static int DAILY_LIMIT = -1000;
  int withdrawal = 0;
  void update(int x) {
    original(x);
    if (x < 0) withdrawal += x;
  }
}
refines class Application {
  void nextDay() {
    original();
    account.withdrawal = 0;
  }
}
```

feature module *Interest*

```
refines class Account {
  final static int INTEREST_RATE = 2;
  int interest = 0;
  int calculateInterest() {
    return balance * INTEREST_RATE / 100 / 365;
  }
}
refines class Application {
  void nextDay() {
    original();
    account.interest +=
      account.calculateInterest();
  }
  void nextYear() {
    original();
    account.balance += account.interest;
    account.interest = 0;
  }
}
```

Figure 2. Three feature modules of the bank account product line.

In Figure 2, we illustrate three feature modules of the bank account product line. Feature module *BankAccount* provides a base implementation, which can store and update the current balance of an account. As defined in the feature model in Figure 1, we assume that this feature is part of every product, but its classes are refined by other features.

Feature module *DailyLimit* adds a field `withdrawal` to store the current withdrawal of the day. Method `update()` is refined to alter the withdrawal whenever the account balance is decreased. Keyword `original` refers to the method being subject of the refinement and is similar to `super` in object-oriented programming. Method `nextDay()` is assumed to be called every day at midnight and refined by feature module *DailyLimit* to reset the withdrawal of the day.

```
class Account {                    product {BankAccount, DailyLimit}
  final static int DAILY_LIMIT = -1000;
  int balance = 0;
  int withdrawal = 0;
  void update(int x) {
    balance += x;
    if (x < 0) withdrawal += x;
  }
}
class Application {
  Account account = new Account();
  void nextDay() { account.withdrawal = 0; }
  void nextYear() {}
}
```

Figure 3. Composition of feature modules *BankAccount* and *DailyLimit*.

Similarly, feature module *Interest* adds a field `interest` to store the cumulated interests since beginning of the year, and adds a method to calculate the interests. The stored interests are updated daily using the refinement for method `nextDay()`, and credited to the account at the end of the year using the refinement for method `nextYear()`.

Four products can be created automatically by composing these three feature modules in different combinations: $\{B\}$, $\{B, I\}$, $\{B, D\}$, and $\{B, I, D\}$. In Figure 3, we illustrate the result of composing the feature modules *BankAccount* and *DailyLimit*. Classes are merged with identically named class refinements. The resulting classes contain all fields and methods defined in the composed feature modules. Already existing methods such as method `update()` are replaced, whereas the keyword `original` is substituted by the method body of the replaced method.

III. Product-Line Strategies

When software product lines are used in safety-critical contexts, we need to verify that all products behave as intended. Consequently, we need to specify the intended behavior of all products. We found several approaches for specification and verification of product lines in the literature. In recent work, we proposed a classification and survey on product-line analysis [10]. Using our classification, we identify strategies to scale specification and verification approaches to product lines. We briefly summarize our classification and discuss how different strategies deal with variability in specification and verification.

A. Specification Strategies for Product Lines

A simple strategy to specify a software product line is to define a specification that all products need to establish, called *global specification* [10]. For example, in a product line of pacemakers, all products have to admit to the same specification stating that a heart beat is generated whenever the heart stops beating [49]. Global specifications were used for different verification techniques such as model checking [17], [40], [42], [44] and static analyses [35], [36]. But in recent work, we found that global specifications are often too restrictive, because variability in implementation does usually require variability in specifications, too [11].

When global specifications are not applicable, we can specify each product of a product line separately, called *product-based specification* [10]. Clearly, specifying the behavior for every product scales only for product lines with few products. An optimization is to specify and analyze only a subset of all products, which is applicable if only this subset is used productively. We did not find any product-based specification approach in the literature, but every specification approach for a single software system can be applied to products individually. Product-based specifications may be useful if the product specifications are largely disjunct, and thus there is a low potential to reuse specifications over several products.

Another strategy is to specify each feature and to compose these specification in a similar manner as source code, called *feature-based specification* [10]. For example, in our bank account product line, we could add a specification to feature *DailyLimit* stating that the daily withdrawal never exceeds the limit. Then, this specification applies to all products containing feature *DailyLimit*. We identified that feature-based specifications were used for model checking only [18], [19], [31], [41], [43]. The main advantage of feature-based specifications is that specifications can be reused across several products. However, specifications applying to combinations of features cannot be defined.

A *family-based specification* is a specification annotated with a propositional formulas stating for which feature combinations the specification is assumed to hold [10]. For example, in a database management system, we might want to specify that statistics over transactions are gathered whenever both features are selected. Family-based specifications generalize product-based and feature-based specifications, because each product-based and feature-based specification is a family-based specifications per definition. Family-based specifications are used for model checking only [34].

B. Verification Strategies for Product Lines

A simple strategy to verify a software product line is to generate and verify all products separately, called *product-based verification* [10]. In principle, any standard verification technique applicable to the generated products can be used for product-based verification. But, product-based verification is feasible only for product lines with few products. We found no proposal in the literature explicitly suggesting an exhaustive product-based verification without any optimizations. But, we found some approaches that actually propose product-based analyses and do not discuss how to deal with many products; these approaches apply type checking [16], model checking [17]–[19], theorem proving [20], and runtime analyses [21] to product lines.

Verification Strategy	Type Checking	Model Checking	Theorem Proving	Other Techniques
Product-based	[16]	[17]–[19]	[20]	[21]
Family-based	[22]–[29]	[19], [25], [30]–[34]		[35], [36]
Feature-product-based	[37]–[39]	[40]–[44]	[43], [45], [46]	
Feature-family-based	[47]		[48]	

Table I
OVERVIEW ON VERIFICATION APPROACHES FOR SOFTWARE PRODUCT LINES.

A more efficient strategy is to verify all products simultaneously using a *family-based verification* [10]. The idea is either to make the verification tool variability-aware [30] or to generate a metaproduct simulating the behavior of all products [25]. For example, in our bank account product line, a metaproduct can be generated by composing *all* feature modules and transforming compile-time variability into runtime variability (i.e., creating a boolean variable for each feature and using dynamic branching to simulate the behavior of all feature combinations). A family-based strategy has been applied to type checking [22]–[29], model checking [19], [25], [30]–[34], and static analyses [35], [36].

Another strategy is to verify the implementation of each feature in isolation without considering other features, called *feature-based verification* [10]. The goal of feature-based verification is to reduce the potentially exponential number of verification tasks (i.e., for every product) to a linear number of verification tasks (i.e., for every feature). But, a feature-based verification can detect only issues *within* a certain feature and does not care about issues *across* features. However, a well-known problem are *feature interactions*: several features work as expected in isolation, but lead to unexpected behavior in combination [50]. Thus, a feature-based strategy can usually not be used for verification as-is. However, it can be combined with other strategies.

In the literature, we found that product-based, family-based, and feature-based strategies are also combined to achieve synergies [10]. The most commonly proposed combination is *feature-product-based verification*, in which features are verified as far as possible in isolation and all remaining verification tasks are done for each product. A feature-product-based strategy is applied to scale type checking [37]–[39], model checking [40]–[44], and theorem proving [43], [45], [46] to product lines. Similarly, *feature-family-based verification* has been proposed using type checking [47] and theorem proving [48].

In our recent survey, we give examples for each strategy and discuss advantages and disadvantages in detail [10]. In Table I, we give an overview on all classified approaches, from which we can make some observations regarding new and underrepresented research areas. First, feature-family-based verification is a young research area for which only two approaches exist so far. Second, while there is a large number of approaches for family-based type checking and family-based model checking, we found not a single approach applying a family-based strategy to theorem proving. But, we argue that several verification techniques such as type checking, model checking, and theorem proving should be applied to product lines, because each technique has strengths and weaknesses [10]. For example, type checking is limited in the errors that can be detected [51] and model checking might not terminate or run out of memory due to the state explosion [52]. We fill this gap by proposing *family-based theorem proving* using feature-based specifications.

IV. SPECIFICATION OF FEATURE MODULES

Our goal is to verify feature modules using contracts, which naturally raises the question how to define contracts for feature modules and how to specify the intended behavior of all products. We give a short overview, how contracts can be defined for object-oriented classes. Then, we present our approach to define contracts for feature modules.

A. Contracts for Object-Oriented Classes

In 1949, Alan Turing formulated that the correctness of large methods should be verified using assertions to simplify the verification [53]. In 1969, Tony Hoare formalized assertions in terms of preconditions and postconditions using the well-known Hoare triple [54]. Two decades later, Bertrand Meyer made assertions popular in object-oriented programming as contracts and invariants [9]. *Contracts* are assigned to methods consisting of a precondition and a postcondition. The precondition is an assertion that callers of the method need to fulfill and the method can rely on. Vice versa, the postcondition must be fulfilled by the method and can be assumed by the caller. *Invariants* are assertions that must hold after each constructor call as well as before and after the execution of public methods.

We use the Java Modeling Language (JML) to define contracts, a behavioral specification language for Java with support for contracts and invariants [55]. In Figure 4, we give a JML specification of feature module *BankAccount*, which is a standard Java program. In JML, a contract is defined using keywords `requires` and `ensures`, denoting the precondition and postcondition, respectively. In our example, the precondition of method `update()` is always fulfilled and the postcondition is stating that the account balance is updated correctly. Additionally, an invariant is defined in class `Application` stating that field `account` is not null.

feature module *BankAccount*

```
class Account {
  int balance = 0;
  /*@
   @ requires true;
   @ ensures balance == \old(balance) + x;
   @*/
  void update(int x) {
    balance += x;
  }
}
class Application {
  //@ invariant account != null;
  Account account = new Account();
}
```

Figure 4. JML contracts and invariants in Java classes.

B. Contracts for Feature Modules

In recent work, we proposed and discussed five approaches to define contracts for feature modules [11]. All approaches enable feature-based specifications from which the specification of each product can be derived automatically. Contracts are composed in a similar manner as feature modules. Thus, specifications do not need to be defined for each product, because we can reuse specifications across several products. In the following, we exemplify one of these approaches, namely *explicit contract refinement*.

In Figure 5, we present contracts for class `Account` in *DailyLimit* and *Interest*. Contracts and invariants may be defined as known from object-oriented programming, except the specification of method contracts for refined methods. For example, method `calculateInterest()` is specified using a JML contract, which only holds when feature *Interest* is selected. Feature *DailyLimit* introduces a new invariant stating that the withdrawal is within the limit. Similarly, this invariant only needs to be established in all products containing feature *DailyLimit*.

The interesting case is the contract of method `update()`. In explicit contract refinement [11], a contract defined for a method refinement replaces the contract of the refined method. In our example, the contract defined in feature *DailyLimit* replaces the contract defined in feature *BankAccount*. But, the replaced contract can be reused with the keyword `original` in a similar manner as in the implementation: `original` in a precondition refers to the replaced precondition and `original` in a postcondition refers to the replaced postcondition. The refined contract extends the precondition to ensure that the daily withdrawal is within the limit and the postcondition to specify that the withdrawal is updated correctly whenever method `update()` is called.

feature module *DailyLimit*

```
refines class Account {
  final static int DAILY_LIMIT = -1000;
  //@ invariant withdrawal >= DAILY_LIMIT;
  int withdrawal = 0;
  /*@
   @ requires \original &&
   @   (withdrawal + x >= DAILY_LIMIT)
   @ ensures \original &&
   @   (x>=0 ==> withdrawal==\old(withdrawal)) &&
   @   (x<0 ==> withdrawal==\old(withdrawal)+x);
   @*/
  void update(int x) {
    original(x);
    if (x < 0) withdrawal += x;
  }
}
```

feature module *Interest*

```
refines class Account {
  final static int INTEREST_RATE = 2;
  int interest = 0;
  /*@
   @ requires true;
   @ ensures (balance >= 0 ==> \result >= 0) &&
   @   (balance <= 0 ==> \result <= 0);
   @*/
  int calculateInterest() {
    return balance * INTEREST_RATE / 100 / 365;
  }
}
```

Figure 5. Feature module specification using explicit contract refinement.

V. Verification of Feature Modules

Specifying Java programs with JML is not only beneficial for verification. Formal specification using JML can be used for documentation generation, runtime assertion checking, automatic test generation, extended static checking, and formal verification using theorem proving [56]. We and others argue that a multitude of techniques is needed to verify the correctness of programs [56], [57]. When formally proving the correctness, the program should already be tested forehand, because formal verification is too expensive for extensive bug finding [56]. Furthermore, certain properties are hard to be proved statically and should be checked at runtime, whereas not all properties should be checked at runtime to minimize the runtime overhead [57]. Hence, when verifying feature modules, a multitude of techniques is needed to efficiently detect errors as well as to prove the absence of errors. In the following, we summarize our approaches for the verification of feature modules.

A. Product-Based Runtime Assertion Checking

A popular application of contracts are runtime assertions [9], [56], [57]. The idea is to check contracts at runtime using a special compiler (e.g., JMLC). The advantage of runtime assertion checking is that contracts may be checked when testing the program, but do not cause runtime overhead in a release version compiled with a standard Java compiler.

We are currently implementing tool support for product-based runtime assertion checking in the integrated development environment FeatureIDE [58] based on an extension of the composer FeatureHouse [59]. When composing feature modules to generate a certain product, contracts are composed resulting in a standard Java program with

```
class Account {          product {BankAccount, DailyLimit}
  final static int DAILY_LIMIT = -1000;
  int balance = 0;
  //@ invariant withdrawal >= DAILY_LIMIT;
  int withdrawal = 0;
  /*@
   @ requires true &&
   @   (withdrawal + x >= DAILY_LIMIT)
   @ ensures balance == \old(balance) + x &&
   @   (x>=0 ==> withdrawal==\old(withdrawal)) &&
   @   (x<0 ==> withdrawal==\old(withdrawal)+x);
   @*/
  void update(int x) {
    balance += x;
    if (x < 0)
      withdrawal += x;
  }
}
```

Figure 6. Composition of specifications for *BankAccount* and *DailyLimit*.

JML specifications. We illustrate the result of composing class `Account` from features *BankAccount* and *DailyLimit* in Figure 6. Feature modules are composed as shown in Figure 3 and the keyword `original` in contracts is replaced similarly. The approach is product-based, because contract violations are detected at runtime for every product individually. A possible optimization is to choose a subset of all products that is likely to detect all errors [10].

B. Product-Based Extended Static Checking

The generated Java programs with JML specifications can also be used as input for static analysis tools such as extended static checkers. We pursued product-based extended static checking using ESC/JAVA2 for the detection of feature interactions [12]. We were able to detect all feature interactions in a small product line of list implementations, but the detection was not straightforward. The reason is that ESC/JAVA2 is unsound and incomplete (e.g., false positives and false negatives may occur), because loops are only unrolled a fixed number of times [60]. Hence, the tool cannot be used to prove the absence of errors, but as runtime assertion checking it is valuable for bug finding.

C. Feature-Product-Based Theorem Proving

Formal specifications in JML can be used to prove that a program behaves as intended. A verification tool translates a JML-annotated Java program into the input language of a theorem prover. One such tool is `Why/Krakatoa` [61], which supports multiple theorem provers such as COQ [62]. COQ is an interactive theorem prover meaning that user interaction is necessary to prove theorems. More precisely, the user needs to write textual commands (i.e., a proof script) that apply certain proof steps until the proof is finished.

We proposed feature-product-based theorem proving using the above mentioned tool chain and proof composition [13]. The idea is to write proof scripts for each feature and compose them together with specification and source code. Then, the composed proof scripts are checked for every product, but the user only needs to write proof scripts once per feature. Hence, this is a feature-product-based approach, in which only the feature-based part requires user interaction and the product-based part can be checked fully automatically using COQ. Using this approach, we were able to reduce the effort to write proof scripts by 88 % [13].

D. Family-Based Theorem Proving

All approaches discussed above rely on the generation of products, which can be problematically for large product lines, because the number of products is up to exponential in the number of features. When the goal is to find errors (e.g., with testing), it is usually sufficient to analyze only a subset of all products. But when the goal is to prove the absence of errors (e.g., with theorem proving), all products need to be generated to achieve completeness.

We proposed the first approach for family-based theorem proving avoiding the generation of all products [14]. The idea is to generate a metaproduct simulating all products and a metaspecification equivalent to all product specifications. We showed how to use the theorem prover KeY as-is for the verification of product lines. We evaluated our approach by means of a case study and measured that the automatic verification of the metaproduct saves 85 % calculation time compared to the individual verification of all products [14].

VI. CONCLUSION AND FUTURE WORK

Efficient strategies for specification and verifications are indispensable for safety-critical software product lines. We classified existing approaches according to product-based, feature-based, and family-based strategies. For specification, we identified global specifications as a further strategy. For verification, also combined strategies such as feature-product-based or feature-family-based verification have been proposed in the literature.

We propose to use contracts to formally specify the intended behavior of product lines implemented with feature-oriented programming. Each feature module is specified using contracts and product specifications can be generated along with source code. Based on contract composition, we discussed product-based runtime assertion checking and extended static checking for bug finding. Furthermore, we discussed feature-product-based theorem proving and family-based theorem proving for proving the absence of errors.

In future work, we intend to evaluate further approaches such as feature-family-based theorem proving and family-based testing. Furthermore, we continue building tool support for multiple specification approaches applying contracts to feature modules [58]. We also intend to formalize already presented specification approaches [11]. Finally, we are going to investigate how evolving product lines can be verified efficiently based on our previous work [63], [64].

ACKNOWLEDGMENT

I thank my supervisor Gunter Saake and Norbert Siegmund for comments on an earlier draft. I gratefully acknowledge the co-authors of previous publications, especially Christian Kästner, Ina Schaefer, Sven Apel, Don Batory, Martin Kuhlemann, and Fabian Benduhn.

REFERENCES

[1] B. Meyer, *Object-Oriented Software Construction*, 1st ed. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[2] K. C. Kang, S. G. Cohen, J. A. Hess, W. E. Novak, and A. S. Peterson, "Feature-Oriented Domain Analysis (FODA) Feasibility Study," Software Engineering Institute, Tech. Rep. CMU/SEI-90-TR-21, 1990.

[3] K. Czarnecki and U. W. Eisenecker, *Generative Programming: Methods, Tools, and Applications*. New York, NY, USA: ACM/Addison-Wesley, 2000.

[4] P. Clements and L. Northrop, *Software Product Lines: Practices and Patterns*. Boston, MA, USA: Addison-Wesley, 2001.

[5] K. Pohl, G. Böckle, and F. J. van der Linden, *Software Product Line Engineering : Foundations, Principles and Techniques*. Berlin, Heidelberg, New York, London: Springer, 2005.

[6] S. Apel and C. Kästner, "An Overview of Feature-Oriented Software Development," *J. Object Technology (JOT)*, vol. 8, no. 5, pp. 49–84, 2009.

[7] C. Prehofer, "Feature-Oriented Programming: A Fresh Look at Objects," in *Proc. Europ. Conf. Object-Oriented Programming (ECOOP)*. Berlin, Heidelberg, New York, London: Springer, 1997, pp. 419–443.

[8] D. Batory, J. N. Sarvela, and A. Rauschmayer, "Scaling Step-Wise Refinement," *IEEE Trans. Software Engineering (TSE)*, vol. 30, no. 6, pp. 355–371, 2004.

[9] B. Meyer, "Applying Design by Contract," *IEEE Computer*, vol. 25, no. 10, pp. 40–51, 1992.

[10] T. Thüm, S. Apel, C. Kästner, M. Kuhlemann, I. Schaefer, and G. Saake, "Analysis Strategies for Software Product Lines," School of Computer Science, University of Magdeburg, Germany, Tech. Rep. FIN-004-2012, 2012.

[11] T. Thüm, I. Schaefer, M. Kuhlemann, S. Apel, and G. Saake, "Applying Design by Contract to Feature-Oriented Programming," in *Proc. Int'l Conf. Fundamental Approaches to Software Engineering (FASE)*. Berlin, Heidelberg, New York, London: Springer, 2012, pp. 255–269.

[12] W. Scholz, T. Thüm, S. Apel, and C. Lengauer, "Automatic Detection of Feature Interactions using the Java Modeling Language: An Experience Report," in *Proc. Int'l Workshop Feature-Oriented Software Development (FOSD)*. New York, NY, USA: ACM, 2011, pp. 7:1–7:8.

[13] T. Thüm, I. Schaefer, M. Kuhlemann, and S. Apel, "Proof Composition for Deductive Verification of Software Product Lines," in *Proc. Int'l Workshop Variability-intensive Systems Testing, Validation and Verification (VAST)*. Washington, DC, USA: IEEE, 2011, pp. 270–277.

[14] T. Thüm, I. Schaefer, S. Apel, and M. Hentschel, "Family-Based Deductive Verification of Software Product Lines," in *Proc. Int'l Conf. Generative Programming and Component Engineering (GPCE)*, 2012, to appear.

[15] D. Batory, "Feature Models, Grammars, and Propositional Formulas," in *Proc. Int'l Software Product Line Conference (SPLC)*. Berlin, Heidelberg, New York, London: Springer, 2005, pp. 7–20.

[16] S. Apel, C. Kästner, and C. Lengauer, "Feature Featherweight Java: A Calculus for Feature-Oriented Programming and Stepwise Refinement," in *Proc. Int'l Conf. Generative Programming and Component Engineering (GPCE)*. New York, NY, USA: ACM, 2008, pp. 101–112.

[17] T. Kishi and N. Noda, "Formal Verification and Software Product Lines," *Comm. ACM*, vol. 49, pp. 73–77, 2006.

[18] S. Apel, W. Scholz, C. Lengauer, and C. Kästner, "Detecting Dependences and Interactions in Feature-Oriented Design," in *Proc. Int'l Symposium Software Reliability Engineering (ISSRE)*. Washington, DC, USA: IEEE, 2010, pp. 161–170.

[19] S. Apel, H. Speidel, P. Wendler, A. von Rhein, and D. Beyer, "Detection of Feature Interactions using Feature-Aware Verification," in *Proc. Int'l Conf. Automated Software Engineering (ASE)*. Washington, DC, USA: IEEE, 2011, pp. 372–375.

[20] D. Bruns, V. Klebanov, and I. Schaefer, "Verification of Software Product Lines with Delta-Oriented Slicing," in *Proc. Int'l Conf. Formal Verification of Object-Oriented Software (FoVeOOS)*. Berlin, Heidelberg, New York, London: Springer, 2011, pp. 61–75.

[21] V. V. Rubanov and E. A. Shatokhin, "Runtime Verification of Linux Kernel Modules Based on Call Interception," in *Proc. Int'l Conf. Software Testing, Verification and Validation (ICST)*. Washington, DC, USA: IEEE, 2011, pp. 180–189.

[22] L. Aversano, M. D. Penta, and I. D. Baxter, "Handling Preprocessor-Conditioned Declarations," in *Proc. Int'l Workshop Source Code Analysis and Manipulation (SCAM)*. Washington, DC, USA: IEEE, 2002, pp. 83–92.

[23] K. Czarnecki and K. Pietroszek, "Verifying Feature-Based Model Templates Against Well-Formedness OCL Constraints," in *Proc. Int'l Conf. Generative Programming and Component Engineering (GPCE)*. New York, NY, USA: ACM, 2006, pp. 211–220.

[24] S. Thaker, D. Batory, D. Kitchin, and W. Cook, "Safe Composition of Product Lines," in *Proc. Int'l Conf. Generative Programming and Component Engineering (GPCE)*. New York, NY, USA: ACM, 2007, pp. 95–104.

[25] H. Post and C. Sinz, "Configuration Lifting: Software Verification meets Software Configuration," in *Proc. Int'l Conf. Automated Software Engineering (ASE)*. Washington, DC, USA: IEEE, 2008, pp. 347–350.

[26] M. Kuhlemann, D. Batory, and C. Kästner, "Safe Composition of Non-Monotonic Features," in *Proc. Int'l Conf. Generative Programming and Component Engineering (GPCE)*. New York, NY, USA: ACM, 2009, pp. 177–186.

[27] F. Heidenreich, "Towards Systematic Ensuring Well-Formedness of Software Product Lines," in *Proc. Int'l Workshop Feature-Oriented Software Development (FOSD)*. New York, NY, USA: ACM, 2009, pp. 69–74.

[28] S. Apel, C. Kästner, A. Größlinger, and C. Lengauer, "Type Safety for Feature-Oriented Product Lines," *Automated Software Engineering*, vol. 17, no. 3, pp. 251–300, 2010.

[29] C. Kästner, S. Apel, T. Thüm, and G. Saake, "Type Checking Annotation-Based Product Lines," *Trans. Software Engineering and Methodology (TOSEM)*, vol. 21, no. 3, 2012, to appear.

[30] A. Gruler, M. Leucker, and K. Scheidemann, "Modeling and Model Checking Software Product Lines," in *Proc. IFIP Int'l Conf. Formal Methods for Open Object-based Distributed Systems (FMOODS)*. Berlin, Heidelberg, New York, London: Springer, 2008, pp. 113–131.

[31] K. Lauenroth, K. Pohl, and S. Toehning, "Model Checking of Domain Artifacts in Product Line Engineering," in *Proc. Int'l*

Conf. Automated Software Engineering (ASE). Washington, DC, USA: IEEE, 2009, pp. 269–280.

[32] A. Classen, P. Heymans, P.-Y. Schobbens, A. Legay, and J.-F. Raskin, "Model Checking Lots of Systems: Efficient Verification of Temporal Properties in Software Product Lines," in *Proc. Int'l Conf. Software Engineering (ICSE)*. New York, NY, USA: ACM, 2010, pp. 335–344.

[33] I. Schaefer, D. Gurov, and S. Soleimanifard, "Compositional Algorithmic Verification of Software Product Lines," in *Proc. Int'l Symposium on Formal Methods for Components and Objects (FMCO)*. Berlin, Heidelberg, New York, London: Springer, 2010, pp. 184–203.

[34] A. Classen, P. Heymans, P.-Y. Schobbens, and A. Legay, "Symbolic Model Checking of Software Product Lines," in *Proc. Int'l Conf. Software Engineering (ICSE)*. New York, NY, USA: ACM, 2011, pp. 321–330.

[35] C. Brabrand, M. Ribeiro, T. Tolêdo, and P. Borba, "Intraprocedural Dataflow Analysis for Software Product Lines," in *Proc. Int'l Conf. Aspect-Oriented Software Development (AOSD)*. New York, NY, USA: ACM, 2012, pp. 13–24.

[36] E. Bodden, "Position Paper: Static Flow-Sensitive & Context-Sensitive Information-Flow Analysis for Software Product Lines," in *Proc. Workshop Programming Languages and Analysis for Security (PLAS)*, 2012, to appear.

[37] S. Apel and D. Hutchins, "A Calculus for Uniform Feature Composition," *Trans. Programming Languages and Systems (TOPLAS)*, vol. 32, pp. 19:1–19:33, 2010.

[38] L. Bettini, F. Damiani, and I. Schaefer, "Implementing Software Product Lines Using Traits," in *Proc. ACM Symposium on Applied Computing (SAC)*. New York, NY, USA: ACM, 2010, pp. 2096–2102.

[39] I. Schaefer, L. Bettini, and F. Damiani, "Compositional Type-Checking for Delta-Oriented Programming," in *Proc. Int'l Conf. Aspect-Oriented Software Development (AOSD)*. New York, NY, USA: ACM, 2011, pp. 43–56.

[40] K. Fisler and S. Krishnamurthi, "Modular Verification of Collaboration-based Software Designs," in *Proc. Europ. Software Engineering Conf./Foundations of Software Engineering (ESEC/FSE)*. New York, NY, USA: ACM, 2001, pp. 152–163.

[41] H. C. Li, S. Krishnamurthi, and K. Fisler, "Interfaces for Modular Feature Verification," in *Proc. Int'l Conf. Automated Software Engineering (ASE)*. Washington, DC, USA: IEEE, 2002, pp. 195–204.

[42] ——, "Modular Verification of Open Features Using Three-Valued Model Checking," *Automated Software Engineering*, vol. 12, no. 3, pp. 349–382, 2005.

[43] M. Poppleton, "Towards Feature-Oriented Specification and Development with Event-B," in *Proc. Int'l Working Conf. Requirements Engineering: Foundation for Software Quality (REFSQ)*. Berlin, Heidelberg, New York, London: Springer, 2007, pp. 367–381.

[44] J. Liu, S. Basu, and R. Lutz, "Compositional Model Checking of Software Product Lines using Variation Point Obligations," *Automated Software Engineering*, vol. 18, no. 1, pp. 39–76, 2011.

[45] D. Batory and E. Börger, "Modularizing Theorems for Software Product Lines: The Jbook Case Study," *J. Universal Computer Science (J.UCS)*, vol. 14, no. 12, pp. 2059–2082, 2008.

[46] B. Delaware, W. Cook, and D. Batory, "Product Lines of Theorems," in *Proc. Conf. Object-Oriented Programming, Systems, Languages and Applications (OOPSLA)*. New York, NY, USA: ACM, 2011, pp. 595–608.

[47] B. Delaware, W. R. Cook, and D. Batory, "Fitting the Pieces Together: A Machine-Checked Model of Safe Composition," in *Proc. Europ. Software Engineering Conf./Foundations of Software Engineering (ESEC/FSE)*. New York, NY, USA: ACM, 2009, pp. 243–252.

[48] R. Hähnle and I. Schaefer, "A Liskov Principle for Delta-oriented Programming," in *Proc. Int'l Conf. Formal Verification of Object-Oriented Software (FoVeOOS)*. Karlsruhe, Germany: Technical Report 2011-26, Department of Informatics, Karlsruhe Institute of Technology, 2011, pp. 190–207.

[49] J. Liu, J. Dehlinger, and R. Lutz, "Safety Analysis of Software Product Lines using State-based Modeling," *J. Systems and Software (JSS)*, vol. 80, no. 11, pp. 1879–1892, 2007.

[50] M. Calder, M. Kolberg, E. H. Magill, and S. Reiff-Marganiec, "Feature Interaction: A Critical Review and Considered Forecast," *Computer Networks*, vol. 41, no. 1, pp. 115–141, 2003.

[51] B. C. Pierce, *Types and Programming Languages*. Cambridge, Massachusetts, USA: MIT Press, 2002.

[52] E. M. Clarke, O. Grumberg, and D. A. Peled, *Model Checking*. Cambridge, Massachussetts: The MIT Press, 1999.

[53] A. Turing, "Checking a Large Routine," in *Conference on High Speed Automatic Calculating Machines*, 1949, pp. 67–69.

[54] C. A. R. Hoare, "An Axiomatic Basis for Computer Programming," *Comm. ACM*, vol. 12, no. 10, pp. 576–580, 1969.

[55] G. T. Leavens, A. L. Baker, and C. Ruby, "Preliminary Design of JML: A Behavioral Interface Specification Language for Java," *SIGSOFT Softw. Eng. Notes*, vol. 31, no. 3, pp. 1–38, 2006.

[56] L. Burdy, Y. Cheon, D. R. Cok, M. D. Ernst, J. R. Kiniry, G. T. Leavens, K. R. M. Leino, and E. Poll, "An Overview of JML Tools and Applications," *Int'l J. Software Tools for Technology Transfer (STTT)*, vol. 7, no. 3, pp. 212–232, 2005.

[57] M. Barnett, M. Fähndrich, K. R. M. Leino, P. Müller, W. Schulte, and H. Venter, "Specification and Verification: The Spec# Experience," *Comm. ACM*, vol. 54, pp. 81–91, 2011.

[58] T. Thüm, C. Kästner, F. Benduhn, J. Meinicke, G. Saake, and T. Leich, "FeatureIDE: An Extensible Framework for Feature-Oriented Software Development," *Science of Computer Programming (SCP)*, 2012, to appear; accepted 2012-06-07.

[59] S. Apel, C. Kästner, and C. Lengauer, "FeatureHouse: Language-Independent, Automated Software Composition," in *Proc. Int'l Conf. Software Engineering (ICSE)*. Washington, DC, USA: IEEE, 2009, pp. 221–231.

[60] P. Chalin, J. R. Kiniry, G. T. Leavens, and E. Poll, "Beyond Assertions: Advanced Specification and Verification with JML and ESC/Java2," in *Proc. Int'l Symposium on Formal Methods for Components and Objects (FMCO)*. Berlin, Heidelberg, New York, London: Springer, 2005, pp. 342–363.

[61] J.-C. Filliâtre and C. Marché, "The Why/Krakatoa/Caduceus Platform for Deductive Program Verification," in *Computer Aided Verification*. Berlin, Heidelberg, New York, London: Springer, 2007, pp. 173–177.

[62] Coq Development Team, *The Coq Proof Assistant Reference Manual*, LogiCal Project, 2010, version 8.3.

[63] T. Thüm, D. Batory, and C. Kästner, "Reasoning about Edits to Feature Models," in *Proc. Int'l Conf. Software Engineering (ICSE)*. Washington, DC, USA: IEEE, 2009, pp. 254–264.

[64] T. Thüm, C. Kästner, S. Erdweg, and N. Siegmund, "Abstract Features in Feature Modeling," in *Proc. Int'l Software Product Line Conference (SPLC)*. Washington, DC, USA: IEEE, 2011, pp. 191–200.

Bewertungsmodelle für Datenpersistenz in Business-Data-Warehouse-Systemen

Thorsten Winsemann
Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg
D-39106 Magdeburg
thorsten.winsemann@t-online.de
Betreuer: Prof. Dr. rer. nat. habil. Gunter Saake

Abstract — **Persistente Datenhaltung über mehrere Schichten innerhalb eines Business-Data-Warehouse-Systems ist notwendig, um den dort vorhandenen, sehr großen Datenbestand nutzen zu können. Die Pflege und Wartung solcher meist redundanten Daten ist jedoch sehr komplex und erfordert einen hohen Aufwand an Zeit und Ressourcen. Es stellt sich die Frage, welche Daten aus welchem Grund bzw. für welchen Zweck persistent abgelegt werden müssen – und wie sich dies effizient entscheiden lässt. Das vorliegende Papier stellt ein Promotionsprojekt vor, welches diese Fragestellungen behandelt. Neben der Problemstellung werden Ziele und Nutzen sowie das Konzept der Arbeit dargestellt. Bisherige Ergebnisse werden ebenso erläutert wie noch ausstehende Arbeiten.**

I. Einleitung

In Business-Data-Warehouse-Systemen (BDW) werden Daten für analytische Anwendungen aufbereitet. Um schnelle Zugriffe zu ermöglichen, werden die großen Datenmengen häufig in Verdichtungsebenen aggregiert. Ankündigungen [1] versprechen auf In-Memory Datenbanken (IMDB) basierende Anwendungen, die auf größte Datenbestände performant zugreifen können – ohne zusätzliche Verdichtungsebenen. Es stellt sich die Frage: Wie viel Persistenz, d.h. nicht-flüchtige Datenspeicherung, ist in solchen Systemen überhaupt noch notwendig? Ist es möglich, jede Art von Analyseanfrage direkt auf dem Rohdatenbestand abzusetzen, welcher „on-the-fly" transformiert wird? Oder gibt es dennoch gewichtige Gründe der Datenspeicherung? Diese Fragen sind jedoch unabhängig vom genutzten Datenbanksystem (DBMS). Wir untersuchen die Fragestellungen, welche Daten in BDW gespeichert werden müssen und wie Notwendigkeit der Datenpersistenz bewertet werden kann.

Kapitel II skizziert die Problemstellung von Persistenz in Data-Warehouse-Systemen (DWS) und geht auf die Besonderheiten von BDW ein. Es werden Ziele, Herausforderungen und Nutzen der Arbeit beschrieben. In Kapitel III erläutern wir Stand der Forschung der angrenzenden Themenbereiche sowie Unterschiede zur Related Work. Kapitel IV beschreibt das Konzept, das Vorgehen bei der Umsetzung und Validierung der Arbeit und skizziert den Arbeitsfortschritt anhand abgeschlossener und noch offener Arbeitspakete. Aktuelle Ergebnisse werden in Kapitel V ausführlich dargestellt. Kapitel VI beendet mit einem Fazit des Forschungsplans und gibt einen Ausblick auf anstehende Aufgaben.

II. Die Problemstellung

A. Datenpersistenz in Data Warehouses

DWS beinhalten häufig sehr große Datenmengen [2], die persistent gespeichert werden. Persistenz bedeutet dauerhafte Speicherung und ist zu unterscheiden von volatiler, bei der die Daten verloren gehen, wenn das DBMS heruntergefahren wird oder abstürzt. Der Aufbau und Betrieb solcher DWS erfordert hohe Anforderungen an die Datenbereitstellung, insbesondere hinsichtlich Performanz, Datengranularität, -flexibilität und -aktualität. Hieraus ergeben sich „natürliche" Konfliktpotentiale; die Anforderungen der Anwendung und die Schaffung der hierfür notwendigen Voraussetzungen stehen sich oftmals konträr gegenüber. Die Anforderung nach hoher Berichtsperformanz, also einer schnellen Datenabfrage, wird zum Beispiel durch den Aufbau zusätzlicher Verdichtungsebenen erfüllt. Durch diese redundante Datenhaltung wiederum wird nicht nur Speicherplatz belegt, sondern zusätzlicher Zeit- und Ressourcenaufwand zur Aggregatsaktualisierung und Datenkonsistenzsicherung notwendig. Gleichzeitig wird die zeitnahe Verfügbarkeit der Daten limitiert. Der definierte Datenbestand schränkt außerdem die Flexibilität der Datenanalyse ein.

Konfliktpotentiale in diesem Umfeld sind beispielsweise:

- Bedarf an detaillierten Informationen erfordert große Datenmengen infolge feiner Granularität.
- Bedarf an historischen Informationen erfordert große Datenmengen infolge vieler „alter" Datensätze.
- Hohe Zugriffsperformanz erfordert Datenredundanz und Konsistenzsicherung durch den Aufbau speziell vordefinierter materialisierter Sichten.
- Flexible Datenauswertung wird durch diese speziell vordefinierten Datenbestände beeinträchtigt.
- Integrierte Informationen erfordern komplexe Transformationsprozesse.
- Auswertung aktuell(st)er Daten wird durch komplexe Transformation eingeschränkt.

Die skizzierten Konflikte verdeutlichen, dass Datenpersistenz große Auswirkungen auf Aufwände des Betriebs von DW-

Systemen hat. Die Vermeidung solcher Persistenz verringert diese Aufwände.

B. Exkurs: Besonderheiten von Business Data Warehouses

DWS finden mittlerweile vielfältige Einsatzmöglichkeiten – zum Beispiel bei Regierungsinformationssystemen [3], für Wetterdaten [4] oder zur Quelltext-Analyse [5]. Unsere Untersuchungen beschränken sich auf BDW [6], in denen entscheidungsunterstützende Informationen für alle Unternehmensebenen und alle Geschäftsbereiche zur Verfügung gestellt werden. Darüber hinaus stellen BDW eine wichtige Datenbasis für eine Vielzahl von Anwendungen dar, wie zum Beispiel Business Intelligence, Customer Relationship Management (CRM) und Planung. Innerhalb einer umfassenden Systemlandschaft sind BDW die „Single Source of Truth“ (vgl. [1], [7]) für alle analyserelevanten Daten des Unternehmens. Das heißt, sie ermöglichen eine allgemein gültige Sicht auf einen zentralen, harmonisierten, validen und konsistenten Datenbestand. Ein BDW integriert sehr große Datenbestände aus einer Vielzahl unterschiedlicher Quellsysteme des Konzerns – oftmals weltweit, so dass Daten verschiedener Zeitzonen zusammengeführt werden müssen. Dies erfordert eine fortlaufende Datenverfügbarkeit bei gleichzeitigem Datenladen und -zugriff. Zudem gibt es weitere Anforderungen an den Datenbestand: Ad-hoc-Berichte, „near-real-time“ Verfügbarkeit und Anwendungen, wie beispielsweise CRM, mit einem Bedarf an detaillierten, historischen Daten. Ein sich ändernder Informationsbedarf muss schnell und flexibel gedeckt werden können. Zudem wird ein umfassendes Berechtigungskonzept zur Sicherung sensibler Daten vorausgesetzt. Hieraus ergeben sich verschiedene, für ein BDW spezifische Gründe von Persistenz.

C. Ziele, Herausforderungen, Nutzen

Zielstellung der Arbeit ist die Ermittlung von Möglichkeiten, Datenpersistenz in BDW zu begründen und zu bewerten. Erst durch höhere Verarbeitungsleistung von IMDB tritt diese Fragestellung in den Vordergrund, denn geringere Leistungsfähigkeit herkömmlicher DBMS führt zu einem hohen Bedarf an aggregiert gespeicherten Daten. Unsere Arbeit wird jedoch grundsätzlich DBMS-unabhängig sein, die aktuelle Entwicklung von IMDB und deren Verwendung bei BDW aber besonders betrachten. Es soll ein allgemeines Modell zur Entscheidungsunterstützung bei folgenden Fragestellungen entwickelt werden:

- Welche Daten sind zu speichern?
- In welchem Format sind die Daten zu speichern?
- Welche Datenpersistenzen werden nicht (mehr) benötigt?

Herausforderungen ergeben sich dadurch, dass solche Entscheidungen nicht nur auf Grundlage von kostenbasierten Modellen getroffen werden können. Ein simples „Kosten versus Nutzen“ der (redundanten) Daten ist nicht ausreichend. Zusätzlich müssen subjektive Präferenzen in Betracht gezogen werden, welche naturgemäß unscharf und schwer zu quantifizieren sind. Hierunter fällt beispielsweise die Frage, wie viel zusätzlicher Wartungsaufwand (für redundante Daten) in Kauf genommen wird, um ein (wie viel) Mehr an Berichtsperformanz zu erreichen. Diese Faktoren sind zudem in einen allgemeinen Kontext einzubinden, der eine Bewertung quantifizierbarer und nicht-quantifizierbarer Größen umfasst.
Einsatzmöglichkeiten ergeben sich sowohl beim Design, als auch beim Redesign von BDW. Die Entscheidung, ob Daten (in welchem Format) gespeichert werden, basiert auf einer objektiveren Bewertung. Hieraus ergibt sich folgender Nutzen:

- Unnötige Datenpersistenz wird vermieden, wodurch sich Aufwände bei Systemerstellung und -erweiterung sowie beim laufenden Betrieb verringern.
- Die Datenbasis wird fundierter, begründeter, transparenter.
- Hardware-Ressourcen werden eingespart.

III. Stand der Forschung & Related Work

Die Fragestellung der behandelten Thematik umfasst ein breites Spektrum angrenzender Bereiche; vier Kategorien sind hervorzuheben:

- DW-Architektur und -Design,
- ETL, Datentransformation, Informationsintegration,
- Wartung, Versorgung und Bewertung materialisierter Sichten und
- In-Memory Datenbanken.

Im Bereich *DW-Architektur und -Design* geben [8] und [9] einen Überblick und Vergleich der verschiedenen Ansätze. Es gibt eine Vielzahl an Literatur zum Thema Referenzarchitektur von DWS. In diesen Architekturen werden drei Hauptbereiche für die drei Arten der Datenverarbeitung definiert: Datenbeschaffung, Datenbearbeitung und Datenbereitstellung. Hierbei variieren Detailgrad (zwischen drei und fünf) und Benennung der Bereiche – die Bedeutung jedoch bleibt gleich. [6] erwartet Daten in einem festgelegten Format und unterscheidet zwischen Schichten mit Rohdaten, detailliert angereicherten Daten und aggregiert angereicherten Daten. [10] definiert eine generische DW-Architektur, in der das DW transformierte und integrierte Daten enthält. Als eine Variante ist das DW aufgeteilt in ein zentrales „Enterprise Data Warehouse“; aus dem mehrere „Business Area Warehouses“ versorgt werden. [11], [12], [13], [14] verfolgen 3-Schichten-Ansätze mit Staging Area, Basisdatenbank und Data-Mart-Bereich – mit geringen Unterschieden in der Benennung. Datenkonsistenz behandelt allerdings keiner der Beiträge näher. [15] modelliert ebenfalls in drei Schichten: „Atomic DW“ mit aktuellen, detaillierten Daten, „departmental/data-mart“ mit geringfügig aggregierten Daten und „individual“ mit hochaggregierten Daten. Zusätzlich werden historische, detaillierte Daten vorgehalten. Im Gegensatz dazu definiert [16] ein DW als Sammlung von prozessbasierter Data-Marts, welche direkt aus der Staging Area gefüllt werden. [17] beschreibt mögliche DW-Architekturen mit einer bis drei Schichten. [7] hingegen stellt eine „Layered, Scalable Architecture” mit fünf Ebenen vor (vgl. Kap. V.A) und geht hierbei insbesondere auf die Fragestellung ein, welche Art der Datenverarbeitung und -transformation auf welcher Ebene erfolgen sollte. Obwohl die Systemarchitektur Datenpersistenz direkt beeinflusst, geht bisher keine der genannten Publikationen hierauf tiefer ein. Aufgrund der Verarbeitungsleistung ist Datenspeicherung auf jeder Schicht

oder nach jeder Transformation implizit.

Im Bereich *ETL* (Datenextraktion, -transformation, -laden) definieren und optimieren beispielsweise [18], [19] und [20] konzeptuelle und logische Workflow-Modelle. Hierbei werden allerdings Problemstellungen nicht genügend erläutert, wie zum Beispiel Datenbereinigung und Anomalien beim Laden des DW. Unsere Arbeit beschäftigt sich nicht mit der Extraktion von Daten aus Quellsystemen, wohingegen sich die erwähnten Arbeiten auf batch-orientierte Extraktion beschränken, welche mit dem Befüllen des ersten DW-Datenziels als abgeschlossen betrachtet wird. Transformations- und Ladeprozesse benötigen allerdings viel Zeit für die Datenverarbeitung über mehrere Schichten und sind wichtige Faktoren innerhalb der Gesamtaufwände heutiger BDW-Installationen. [21] propagiert ein System, welches transaktionale und analytische Datenverarbeitung kombiniert – ohne herkömmliches ETL, sondern mit Datentransformation „on-the-fly“. [22], [23] und [24] beschäftigen sich mit *Datentransformation* in OLAP-Datenbanken und beschreiben Aggregationsoperatoren, Summierbarkeit und formale Rahmenbedingungen für Aggregation. Solche Erkenntnisse können als Indikatoren der Entscheidungsfindung bei unserer Arbeit verwendet werden. [25] deckt den Bereich der *Informationsintegration* ab. Die genannten Arbeiten beschäftigen sich allerdings weder grundlegend mit der Komplexität der Datenverarbeitung im DW hinsichtlich Zeit- und Ressourcenverbrauch, noch werden die Auswirkungen und Notwendigkeit von Datenpersistenz eingehend berücksichtigt.

Datenpersistenz in DWS ist eng verbunden mit *materialisierten Sichten*, deren *Wartung* sowie *inkrementeller Datenversorgung*. [26], [27] und [28] forschen im Bereich inkrementeller Wartung von materialisierten Sichten innerhalb einzelner DBMS. Updates werden als Paare von Lösch- und Einfügeoperationen definiert, wobei Datenhistorie allerdings unberücksichtigt bleibt. Diese ist aber ausdrücklich ein Existenzgrund des DW. Im Hinblick auf Arbeiten über Sichtenpflege in DWS (z.B. [29], [30], [31], [32], [33]) müssen wir [34] beipflichten: die den Arbeiten zugrunde liegenden Beschreibungen von Data Warehousing sind praxisfern hinsichtlich DW-Ladezyklen, Zugriffsfähigkeit von Datenquellen und Bedeutung historischer Daten. [34], [35] und [36] sind im Bereich der inkrementellen Sichtenwartung in DW tätig. [34] weist auf die Ähnlichkeit von inkrementeller Datenversorgung eines DW und inkrementeller Pflege materialisierter Sichten hin, da sich beides um “incremental updates of physically integrated data” handelt. Solche Techniken sind wichtig im Data Warehousing und werden somit auch in unserer Arbeit berücksichtigt. Allerdings lassen auch diese Autoren die Persistenzgründe unberücksichtigt und beschränken ihre Betrachtungen auf das erste Datenziel im DW. Ein kostenbasiertes *Bewertung*smodell zur Identifikation optimaler materialisierter Sichten wird in [37] beschrieben. Es kann als Grundlage unserer Arbeit dienen, muss allerdings um subjektive und unscharfe Einflussfaktoren erweitert werden.

Bei den *In-Memory Datenbanken* werden ausschließlich solche mit voller ACID-Unterstützung (atomicity, consistency, isolation, durability) betrachtet. Hierunter fallen vorrangig das MonetDB-Projekt [38] sowie die Arbeiten von [39], [40] und [1]. Zudem gibt es bereits kommerziell angebotene IMDB: TimesTen von Oracle [41], IBMs solidDB [42] und SAP HANA [43]. Da erst die Verarbeitungsleistung dieser Systeme eine vertiefendende Untersuchung von Persistenzgründen initiiert hat, beeinflussen Fortschritte in diesem Bereich unserer Arbeit unmittelbar. Heutige DBMS in kommerziell genutzten DWS bieten Werkzeuge, die kostenbasierte Optimierungsmodelle für den Datenzugriff anbieten und Vorschläge zur Erstellung von Indexen und materialisierten Sichten erstellen. Unserer Kenntnis nach wird darüber hinaus die Frage nach Entscheidungsunterstützung bei Persistenzmodellen nicht diskutiert. Diese Frage wird allerdings infolge von IMDB-basierten BDW an Bedeutung gewinnen.

IV. DAS KONZEPT

A. Vorgehen zur Zielerreichung

Ziel der Arbeit ist die Entwicklung eines Bewertungsmodells zur Entscheidungsunterstützung bei der Fragestellung, ob und in welchem Format Daten in BDW persistiert werden sollen. Ein solches Modell kombiniert verschiedene Kriterien oder Faktoren (s. Kap. V.D) und führt zu einem Ergebnis, welches entscheidungsunterstützend oder auch -bestimmend sein kann. Abhängig von der „Ausgangssituation“ (z.B.: „Soll eine existierende oder eine zu erstellende Persistenz bewertet werden?“) ist die Anzahl der Kriterien variabel. Es ist offen, ob dieses Modell einstufig, mehrstufig oder kombiniert angewendet werden kann. Einstufig heißt, dass das Modell alle erforderlichen Kriterien in einem Schritt verarbeitet. Mehrstufig hingegen bedeutet eine sukzessive Vorab-Abfrage von Ausschlusskriterien. Eine kombinierte Anwendung ist denkbar, da bei einer Bewertung nicht (immer) alle Kriterien zu prüfen und in das Modell einzubringen sind.

Neben Literaturrecherche zu den relevanten Themen werden Persistenzgründe ermittelt, beschrieben und gegliedert. Die für eine Bewertung nötigen Kenngrößen sind zu definieren und gewichten und in dem Modell zusammengefasst. Das Modell wird theoretisch erstellt und rechnerunterstützt umgesetzt.

B. Umsetzung und Arbeitspakete

Die bei der Umsetzung der Arbeit anfallenden Bereiche und die daraus resultierenden Arbeitspakete (AP) werden zunächst kurz skizziert und grob quantifiziert. In Kapitel V stellen wir sowohl die Ergebnisse der fertigen als auch die Teilergebnisse und Aufgaben der offenen Arbeitspakete detailliert dar.

AP1. Bei der Untersuchung von Persistenz in BDW fällt ein Zusammenhang zwischen Architektur und Datenspeicherung auf. Wir vergleichen Referenzarchitekturen mit einer Architektur klar-definierter Schichten, in der zugeordnet ist, wo welche Daten in welchem Format zu speichern sind (s. Kap. V.A). AP1 ist ein geringer Teil der Gesamtarbeit und abgeschlossen.

AP2. Wir bestimmen den Persistenzgrund als eine wichtige Basis zur Beurteilung der Datenspeicherung in BDW und definieren und gliedern diese Gründe (s. Kap. V.B). Aufbauend auf eine Einteilung der Gründe in „verpflichtend bzw. essentiell notwendig“ entwickeln wir ein Diagramm zur

Entscheidungsunterstützung (s. Kap. V.C). AP2 stellt einen mittleren Umfang der Gesamtarbeit dar und ist abgeschlossen.
AP3. Als Ergebnis dieser Kategorisierung ist festzustellen, dass sich die Frage nach Datenpersistenz bereits aus der Notwendigkeit des Speichergrundes beantworten lassen kann (z.B. durch Gesetze). Schwieriger ist die Bewertung bei unscharfen Begriffen (z.B. „komplex"), bei denen neben objektiven, mess- oder schätzbare Kennzahlen auch subjektive Präferenzen mit einfließen müssen. Wir definieren objektive und subjektive Kenngrößen und fassen sie gewichtet in einem Bewertungsmodell zur Gesamtbeurteilung zusammen. Als Grundlage für dieses Modell, welches rechnerbasiert arbeiten soll, bieten sich verschiedene Methoden der Multikriteriellen Entscheidungsanalyse (MCDA) an (s. Kap. V.D). Die für das Modell benötigten objektiven Kenngrößen sind in DWS zu bestimmen, subjektive durch Befragung von Administratoren produktiver BDW zu ermitteln. Begleitend soll durch nicht-repräsentative Umfrage ermittelt werden, welche Aufwände für Persistenzpflege generell anfallen. AP3 ist der Hauptteil der Gesamtarbeit und gegenwärtig in Arbeit.
AP4. Abschließend soll ein Ausblick auf besondere Konzepte von IMDB gegeben und untersucht werden, ob IMDB technische Möglichkeiten zur Vermeidung von Persistenz bieten. Können zum Beispiel besondere Zeitstempel-Verfahren (vgl. [1], [40]) anstatt Datenspeicherung aufgrund „konstanter Datenbasis" oder „En-bloc-Datenversorgung" (vgl. Kap. V.B) verwendet werden? AP4 stellt einen mittleren Umfang der Gesamtarbeit dar und ist offen.

C. *Validierung der Arbeit*

Die Validierung der Arbeit erfolgt anhand der einzelnen Arbeitspakete. Bei der Bearbeitung der Pakete fließt über die hauptberufliche Zusammenarbeit des Autors mit DW-Architekten und -Anwendern sowie DW-Entwicklern kontinuierlich praxisrelevantes Feedback in die Arbeit ein. Regelmäßige Rücksprachen mit den Betreuern werden ebenso verfolgt wie Teilnahmen an Doktorandentreffen, Diskussionen mit DW-Entwicklern und -Anwendern und Veröffentlichung von Forschungsergebnissen.
Die in AP1 vorgestellte Schichtenarchitektur wird bereits in der Praxis eingesetzt [7], einen Vergleich mit herkömmlichen Referenzarchitekturen stellen wir in [44], [45], [46] vor. Ergebnisse zu AP2 sind in [47], [48], [49] veröffentlicht. Das in AP3 zu entwickelnde Modell wird zunächst anhand fiktiver Werte definiert; anschließend erfolgen sukzessive Tests mit realitätsnahen und schließlich mit realen Werten. Ein Einsatz des Modells unter realen Bedingungen eines produktiven BDW – möglichst unter Einbeziehung von DW-Designern, -Nutzern und -Verantwortlichen – bilden den Abschluss der Validierung. Eine konkrete Testdefinition ist noch offen. Ergebnisse von AP4 können durch prototypische Systemimplementierungen validiert werden.

V. Beurteilung von Datenpersistenz

A. *Prolog: Eine Schichtenarchitektur für BDW*

Datenpersistenz in einem DW ist eng verbunden mit dessen Architektur, die zweckgebundene Bereiche definiert, in denen die Datenverarbeitung sukzessive erfolgt. Eine allgemeine Referenzarchitektur beinhaltet drei Bereiche, welche die drei Arten der Datenverarbeitung darstellen: Datenbeschaffung in der „Staging Area"; Datenbearbeitung in der Basisdatenbank, Datenbereitstellung im Data-Mart-Bereich. In diesem eher groben Modell ist Datenspeicherung in jedem Bereich implizit [50]. Die in [7] vorgestellte Schichtenarchitektur entwickelt diesen Ansatz hinsichtlich der bereits erwähnten Anforderungen an ein BDW weiter. Die Schichten werden zweckbestimmter; jede der fünf Schichten repräsentiert einen Bereich, in dem der Wert der Daten hinsichtlich ihrer Verwendung gesteigert wird. Die Datenspeicherung in jeder Schicht ist hier nicht zwangsläufig. Obwohl die Grenzen fließend sind [51], können die fünf Schichten den drei Arten der Datenverarbeitung zugeordnet werden. In [44], [45], [46] geben wir eine detaillierte Beschreibung der einzelnen Schichten und vergleichen die verschiedenen Architekturansätze ausführlich.

B. *Gründe der Datenpersistenz*

Wir definieren, dass der Datenverwendungszweck, das heißt der Grund der Speicherung, das erste Kriterium ist, um den Bedarf nach persistenten Daten beurteilen zu können. Zwei Gründe von Datenpersistenz im BDW werden hauptsächlich genannt: Speicherung der bereits transformierten Daten als Datenbasis und Speicherung redundanter, aggregierter Daten im Data-Mart-Bereich zur Performanzverbesserung. Darüber hinaus gibt es jedoch weitere Gründe, welche in der Literatur bisher nicht erläutert wurden. Wir gruppieren diese Gründe in Datenbeschaffung, Datenbearbeitung, Datenverwaltung, Datenverfügbarkeit sowie Gesetze und beschreiben sie im Folgenden kurz (s. [48] für eine ausführliche Beschreibung mit Beispielen).

Quellsystem-Entkopplung: Zur Entlastung des Quellsystems werden Daten direkt nach ihrer erfolgreichen Extraktion im Eingangsbereich des DW gespeichert; hierbei werden die Daten nicht oder nur in geringem Maße verändert (z.B. werden Herkunftsmerkmal oder Zeitstempel angefügt).
Datenverfügbarkeit: Oftmals sind Daten nicht mehr oder nur in einem veränderten Zustand verfügbar; hierzu zählen Daten aus dem Internet, aus Dateien oder Altsystemen. Zudem können Netzwerkprobleme zu eingeschränkter Datenverfügbarkeit führen.
Aufwendige Datenwiederherstellung: Sind Daten nicht mehr im Quellsystem verfügbar (z.B. weil sie archiviert sind), ist eine Wiederherstellung aufwendig, so dass sie im DW – oft in aggregierter Form – gespeichert werden.
Data-Lineage: Daten in Berichten oder Analysen sind häufig Ergebnis mehrstufiger Transformationsprozesse. Um eine Rückverfolgung zu den Ursprungsdaten zu erleichtern oder zu ermöglichen, etwa zur Validierung, können gespeicherte Zwischenergebnisse erforderlich sein (vgl. [52]).
Veränderte Transformationsregeln: Regeln können geändert werden. Besitzen die Daten kein Zeitmerkmal und werden die Transformationen nicht „historisiert"; so ist eine identische Transformation nicht mehr möglich.
Abhängige Transformationen: Hierunter verstehen wir solche, deren Durchführung den Zugriff auf weitere Daten erfordert; zum Beispiel erfordert die Verteilung eines Bonus' auf die

einzelnen Mitarbeiter die Gesamtanzahl der Mitarbeiter. Diese notwendigen Daten werden im DW gespeichert, um das korrekte Prozessieren der Transformation zu gewährleisten.

Komplexe Transformationen: Aufgrund ihrer Komplexität sind einige Transformationen sehr zeit- und ressourcenaufwendig, so dass die Daten gespeichert werden, um ein wiederholtes Transformieren zu vermeiden.

Komplex-abweichende Daten: Zu integrierende Daten können in Syntax und Semantik sehr von der im BDW üblichen abweichen; eine Transformation erfolgt schrittweise mit gespeicherten Zwischenergebnissen.

Konstante Datenbasis: Einige, auf Daten des DW aufbauende Applikationen (z.B. Planung) erfordern eine konstante Datenbasis, welche sich während der Benutzung nicht ändern darf und deswegen separat gespeichert wird.

„En-bloc Datenversorgung“: Üblicherweise fließen neue Daten, aus verschiedenen, gegebenenfalls weltweiten Quellen, zeitlich verteilt in ein BDW. Nachdem diese syntaktisch und semantisch integriert wurden, werden sie zwischengespeichert und erst zu einem bestimmten Zeitpunkt in die Datenbasis des DW gespielt. Hierdurch wird ein zeitlich definierter, konstanter und in sich plausibler Datenbestand gewährleistet.

Komplexe Berechtigungen: Anstatt der Definition komplexer Benutzerberechtigungen (z.B. auf Dimension oder Feldinhalt), werden bestimmte Data-Marts mit den Daten erstellt und die Berechtigungen auf dem Data-Mart vergeben.

„Single Version of Truth“ (SvoT): Transformierte Daten werden nach unternehmensweit gültigen Definitionen, aber ohne spezielle Geschäftslogik gespeichert. Hierdurch wird ein einheitlicher, vergleichbarer Datenbestand geschaffen [7].

„Corporate Data Memory“ (CDM): Alle ins BDW extrahierten Daten werden ohne oder nur mit minimaler Veränderung (z.B. durch Anfügen eines Herkunftsmerkmals) gespeichert, um eine größtmögliche Autarkie und Flexibilität von Datenquellen zu ermöglichen. So können Datenbestände (wieder-)hergestellt werden, ohne auf die Quellsysteme zuzugreifen, in denen die Daten möglicherweise nicht mehr zum Zugriff bereitstehen (vgl. [7]).

Informationsgewähr: BDW haben zu gewährleisten, dass die Daten den Benutzern in einem bestimmten Zeitraum (oftmals sogar 24 Stunden pro Tag) zur Verfügung stehen und für die Anwendungen genutzt werden können. Hierfür werden besonders kritische Datenbestände zusätzlich gespeichert.

Datenzugriffsgeschwindigkeit: Redundante Speicherung von Daten in Verdichtungsebenen zur Performanzverbesserung beim Datenzugriff stellt einen der häufigsten Gründe für die Einführung weiterer Persistenzen dar.

Corporate Governance: Daten werden gemäß den Compliance-Vorgaben des jeweiligen Unternehmens (Corporate Governance) gespeichert; zum Beispiel können so aufgrund bestimmter Daten getroffene Entscheidungen des Managements auch im Nachhinein beurteilt werden.

Gesetze und Bestimmungen: Datenspeicherung wird auch durch Gesetze und Bestimmungen begründet, beispielsweise im Finanzbereich (Handelsgesetzbuch u.a., [53]) und bei der Produkthaftung [54].

Subjektive Sicherheit: Letztlich kann das subjektive Bedürfnis an Sicherheit ein Grund für Datenspeicherung sein.

Persistenz beinhaltet häufig redundante Datenhaltung, da sowohl Quell- als auch transformierte Zieldaten gespeichert werden; ausschließliches Speichern der Zieldaten bedeutet in aller Regel Datenverlust. Hieraus entstehen hohe Anforderungen, nicht nur an die Hardware (Speicherplatz etc.), sondern auch an die Datenpflege, um etwa die Datenbestände konsistent zu halten (vgl. [55]).

C. Notwendigkeit der Datenpersistenz

Eine Entscheidung für Datenpersistenz kann nicht ausschließlich nach einem kostenbasierten Vergleich von „Plattenplatz und Kosten des Updates versus Geschwindigkeitsgewinn der Analyse” getroffen werden. Zunächst ist der Grund der Datenspeicherung (s. vorheriger Abschnitt) zu berücksichtigen. Wir führen die Notwendigkeit der Gründe als Entscheidungskriterium ein: die Speicherung der Daten ist nur unterstützend, essentiell oder verpflichtend.

Tab. 1: Persistenzgründe, nach Notwendigkeit gruppiert

Grund/Zweck	Notwendigkeit	Gruppe
Datenverfügbarkeit	Verpflichtend	-
Veränderte Transformationsregeln	Verpflichtend	-
Abhängige Transformationen	Verpflichtend	-
Corporate Governance	Verpflichtend	-
Gesetze und Bestimmungen	Verpflichtend	-
Quellsystem-Entkopplung	Essentiell	Aufwand
Aufwendige Datenwiederherstellung	Essentiell	Aufwand
Komplexe Transformationen	Essentiell	Performanz
Konstante Datenbasis	Essentiell	Vereinfachung
Data-Lineage	Essentiell	Vereinfachung
Komplex-abweichende Daten	Essentiell	Vereinfachung
“En-bloc Datenversorgung”	Essentiell	Vereinfachung
Komplexe Berechtigungen	Essentiell	Vereinfachung
„Single Version of Truth“	Essentiell	Design
„Corporate Data Memory“	Essentiell	Design
„Informationsgewähr”	Essentiell	Sicherheit
Zugriffsgeschwindigkeit	Essentiell	Performanz

Verpflichtend zu speichern sind Daten aufgrund von Gesetzen und Bestimmungen sowie Regeln der Corporate Governance. Zudem gilt dies für Daten, welche nicht wieder hergestellt werden können, weil sie nicht mehr oder nur verändert zur Verfügung stehen oder aufgrund geänderter Transformation nicht mehr erstellt werden können. Auch Daten, die bei der Transformation anderer Daten benötigt werden, sind zu speichern, wenn eine gleichzeitige Verfügbarkeit nicht gewährleistet werden kann. *Essentielle* Datenpersistenz kann in bestimmte Gruppen unterteilt werden: Zum einen Daten, deren Wiederherstellung nur mit sehr hohem *Aufwand* (an Zeit und Ressourcen) möglich ist, wie z.B. archivierte oder komplex transformierte Daten. Hierbei ist „sehr hoch“ allerdings subjektiv und näher zu untersuchen. Eine zweite Gruppe sind Daten, die gespeichert werden, um den Betrieb

des DWS oder einzelner Anwendungen zu *vereinfachen*; hierzu zählen speziell abgelegte Plandaten oder Data-Marts mit Berechtigungen für besondere Benutzer. Drittens begründet sich Persistenz mit spezieller Konzeption (*Design*) eines BDW: SVoT und CDM zählen u.a. hierzu. *Sicherheit*, etwa zur Gewährleistung der Datenverfügbarkeit, stellt eine weitere Gruppe dar. Letztlich ist Datenspeicherung für eine hohe *Performanz* ein Grund; oftmals der, dem das größte redundante Datenvolumen zugrunde liegt. *Unterstützende* Datenspeicherung ausschließlich aufgrund eines subjektiven Sicherheitsempfindens berücksichtigen wir nicht, da sich dieses nach objektiven Kriterien nicht begründet lässt. Tab. 1 enthält eine komplette Auflistung der nach Notwendigkeiten gruppierten Persistenzgründe.

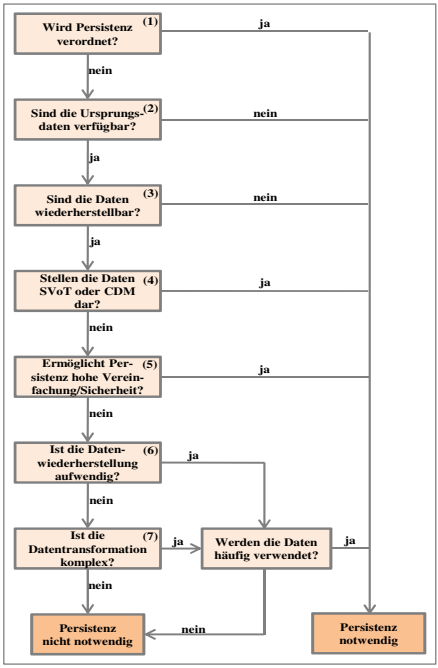


Abb. 1: Entscheidungsdiagramm „Datenpersistenz“

Wir entwickeln ein vereinfachtes Entscheidungsdiagramm für Datenpersistenz, in dem unscharfe Begriffe wie „aufwendig“; „komplex“ und „häufig“ abhängig von der Domäne spezifiziert werden müssen (s. Abb. 1). Abfragen (1) bis (3) betreffen verpflichtende Gründe, d.h. die Daten sind – auch in IMDB-basierten BDW – zu speichern. Bei den aus anderen Gründen gespeicherten Daten sind die Entscheidungsgrundlagen sehr vielfältig. Stellen die Daten eine „Single Version of Truth“ dar oder umfasst das BDW-Design ein „Corporate Data Memory“; so sind diese Daten zu speichern. Wird hingegen aufgrund komplexer Reproduktion oder Transformation gespeichert, so müssen zum Beispiel Zugriffshäufigkeit und Sicherstellung der Verfügbarkeit in Betracht gezogen werden, um entscheiden zu können. [47] und [49] erläutern diesen Themenbereich mit Hinblick auf IMDB-basierte BDW.

D. Bewertung von Datenpersistenz

Alle nicht-verpflichtend gespeicherten Daten sind Gegenstand unserer näheren Betrachtung. Das betrifft redundant gespeicherte Daten (z.B. zur Performanzverbesserung), heißt aber im Umkehrschluss nicht, dass alle redundanten Daten obsolet sind. In Abb. 1 sind dies die Abfrageschritte (4) - (7). Ist beispielsweise designbedingt ein CDM Teil des BDW, so sind die Daten auch persistent abzulegen. Komplizierter ist die Beantwortung von Fragen mit unscharfen Kriterien wie „aufwendig“; „komplex“ und „häufig“. Wir verdeutlichen den Zusammenhang anhand eines Beispiels (s. auch Abb. 2).

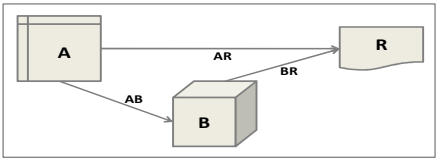


Abb. 2: Beispiel „Vergleich Performanz vs. Aufwand“

Die Fragestellung ist, ob die Daten für Bericht R direkt aus Tabelle A gelesen werden oder ob mit Cube B eine weitere Persistenz definiert wird. Beide Optionen haben Vor- und Nachteile. So ist der Zugriff auf Cube B sehr wahrscheinlich schneller als auf Tabelle A. Allerdings ist für Cube B auch zusätzlicher Wartungsaufwand nötig. Die wichtigsten Faktoren zur Bewertung des Datenzugriffs (AR und BR) sind Selektionszeit (T_σ), Transformationszeit (T_τ) und Datenzugriffshäufigkeit (F_Q). Bleiben weitere Faktoren unberücksichtigt (wie z. B. Zeit für Indexerstellung), so gelten folgende Formeln:

- $AR = \sum_{i=1}^{n} (T_{\sigma(AR)} + T_{\tau(AR)})$; $BR = \sum_{i=1}^{n} (T_{\sigma(BR)} + T_{\tau(BR)})$

Die Datenversorgung von Tabelle A zu Cube B (AB) ist gekennzeichnet durch die Ladehäufigkeit (F_Δ), die Update-Zeit (T_U) und die Reorganisationszeit und -häufigkeit (T_R, F_R):

- $AB = \sum_{i=1}^{n} (T_{\sigma(AB)} + T_{\tau(AB)} + T_U) + (T_R * F_R)$

Eine einfache Gegenüberstellung der Ergebnisse ist allerdings unzureichend, da weitere wichtige Faktoren unberücksichtigt bleiben. Ist zum Beispiel eine Höchstlaufzeit des Berichts R definiert („höchstens 5s“), die beim Zugriff auf Tabelle A nicht erreicht wird, so muss Cube B erstellt werden.

Die aufgeführten Faktoren sind Beispiele für berechenbare Indikatoren. Weitere sind Datenvolumen (mit Einfluss auf Speicherplatz und Verarbeitungszeit), Zeit für Indexerstellung und für den Aufbau materialisierter Sichten. Die Indikatoren unterscheiden sich zudem je nach Anwendungsfall. Das obige Beispiel betrifft den Bereich „Performanz“ (Schritt (7) in Abb. 1). Dieser sehr häufige Anwendungsfall ist Gegenstand unserer aktuellen Forschung. Weitere Bereiche sind

„Vereinfachung/Sicherheit“ und „Aufwand“ (Schritte (5) und (6) in Abb. 1). Eine komplette Sammlung und Zuordnung der Indikatoren für diese Anwendungsfälle ist noch offen. Hinzu kommen weitere, nicht-berechenbare Indikatoren wie Anforderungen an die Daten; zum Beispiel Verfügbarkeit, Qualität, Konsistenz und Plausibilität. Auch Aufwände für Datenpflege und -validierung sind schwer zu quantifizieren.

Ein erster Ansatz für ein Bewertungsmodell ist die Klassifizierung relevanter Indikatoren (s. Abb. 3). Definierbar bedeutet, dass die Kennzahlen für das jeweilige Modell bestimmt werden können. Berechenbare Größen sind zum Beispiel Datenvolumen oder Datenänderungshäufigkeit. Datennutzungshäufigkeit ist eine messbare Größe, Zunahme des Datenvolumens eine prognostizierbare. Festlegbar sind Größen wie Berichtsverfügbarkeit (z. B. “99.9%”) oder Mindestantwortzeit (z.B. “< 5s für 95% der Berichte”). Festgelegte Indikatoren sind vorgegeben – extern (z.B. durch Gesetze) oder intern (z.B. durch Definition eines CDM).

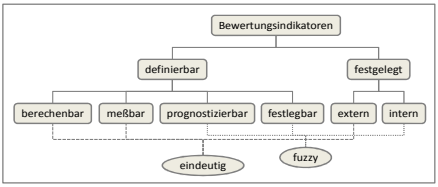


Abb. 3: Klassifizierung von Bewertungsindikatoren

Entscheidungen können allerdings nicht allein auf Basis der genannten Indikatoren getroffen werden. Zusätzlich müssen subjektive Präferenzen in Betracht gezogen werden, welche naturgemäß unscharf und schwer zu quantifizieren sind. Hierunter fällt beispielsweise die Frage, wie viel zusätzlicher Wartungsaufwand in Kauf genommen wird, um ein Mehr an Berichtsperformanz zu erreichen. Anders ausgedrückt: Wie viel mehr (an Aufwand) darf ein schnellerer Bericht „kosten“; was ist (den Entscheidungsträgern) der Performanzgewinn „wert“. Diese Faktoren sind zudem in einen allgemeinen Kontext einzubinden, welcher eine Bewertung quantifizierbarer und nicht-quantifizierbarer Größen umfasst. Bei dieser Problematik scheinen Ansätze und Methoden der Multikriteriellen Entscheidungsanalyse eine gute Grundlage für die gemeinsame Bewertung „harter“ und „weicher“ Faktoren zu bieten. [56] gibt eine einführende Übersicht der MCDA. Insbesondere Methoden wie die klassische Nutzwertanalyse (vgl. [57]) und Analytischer Hierarchieprozess [58] versprechen geeignete Anwendung (vgl. [59]).

VI. FAZIT & AUSBLICK

Persistente Datenhaltung in BDW ist notwendig, um den sehr großen Datenbestand nutzen zu können. Pflege und Wartung solcher meist redundanten Daten ist jedoch sehr komplex und erfordert einen hohen Aufwand an Zeit und Ressourcen. Das vorliegende Papier beschreibt ein Promotionsprojekt, das die Frage untersucht, welche Daten aus welchem Grund bzw. für welchen Zweck persistent abgelegt werden müssen – und wie sich dies effizient entscheiden lässt. Neben der Problemstellung werden Ziele und Nutzen sowie das Konzept der Arbeit dargestellt. Bisherige Ergebnisse und ausstehende und nächste Arbeiten werden ausführlich erläutert und im Folgenden nochmals kurz aufgeführt.

Abgeschlossen ist die ein Vergleich von BDW-Architekturen (AP1) sowie die Definition, Kategorisierung und Gruppierung möglicher Persistenzgründe in BDW und die Entwicklung eines Entscheidungsdiagramms „Datenpersistenz“ (AP2).

Aktuelles Forschungsthema ist die Definition und Gliederung objektiver Bewertungsindikatoren. Dies wird anhand des Bereichs „Performanz“ ermittelt, wie diese Indikatoren im DWS zu bestimmen sind. Gleichzeitig untersuchen wir, ob eine MCDA-Methode für das Bewertungsmodell genügt, oder ob mehrere oder eine Kombination von Methoden geeignet sind. Nachfolgend ist das Modell mit realitätsnahen Daten zu testen und mit realen Daten anzuwenden (AP3).

Schließlich werden wir untersuchen, ob und wie aktuelle Datenhaltungskonzepte in IMDB zur Vermeidung redundanter Speicherung geeignet sind (AP4).

LITERATURVERZEICHNIS

[1] H. Plattner, A. Zeier: „In-Memory Data Management“; Springer-Verlag, Berlin (2011)

[2] R. Winter, „Why are data warehouses growing so fast?“; auf: http://www.b-eye-network.com/print/7188 {01.07.2012} (2008)

[3] B.A. Devlin, P.T. Murphy: „An architecture for a business and information system“; in: IBM Systems Journal 27(1), S. 60-80 (1988)

[4] F. Finkler: „Konzeption eines Regierungsinformationssystems“; Gabler-Verlag, Wiesbaden (2008)

[5] Meteomedia: „Ein Real Time Data Warehouse für Wetterdaten“; auf: http://www.meteomedia.de/index.php?id=525 {01.07.2012} (2012)

[6] T. Frey: „Vorschlag Hypermodelling: Data Warehousing für Quelltext“; 23. GI-Workshop Grundlagen von Datenbanken 2011, Obergurgl (Österreich), 31. Mai - 3. Juni 2011, in: Proceedings of the 23rd GI-Workshop „Grundlagen von Datenbanken 2011“, S. 55-60 (2011)

[7] SAP: „PDEBW1 – Layered, Scalable Architecture (LSA) for BW“; Schulungsunterlagen, SAP AG (2009)

[8] H.J. Watson, T. Ariyachandra: „Data Warehouse Architectures: Factors in the Selection and the Success of the Architectures“; auf: www.terry.uga.edu/~hwatson/DW_Architecture_Report.pdf {01.07.2012} (2005)

[9] A. Hajmoosaei, M. Kashfi, P. Kailasam: „Comparison plan for data warehouse system architectures“; in: 3rd ICMiA Proceedings, S. 290-293 (2011)

[10] V. Poe: „Building a data warehouse for decision support“; Prentice Hall PTR, Upper Saddle River (1996)

[11] H. Muksch, W. Behme (Hrsg.): „Das Data Warehouse-Konzept“; Gabler-Verlag, Wiesbaden, 4. Auflage (2000)

[12] P. Gluchowski; P. Chamoni: „Entwicklungslinien und Architekturkonzepte des On-Line Analytical Processing“; in: Analytische Informationssysteme, Springer-Verlag, Heidelberg, 3. Auflage, S. 143-176 (2006)

[13] T. Zeh: „Referenzmodell für die Architektur von Data-Warehouse-Systemen (Referenzarchitektur)“; auf: www.tzeh.de/doc/gse-ra.ppt {01.07.2012} (2008)

[14] A. Bauer, H. Günzel (Hrsg.): „Data-Warehouse-Systeme“; dpunkt-Verlag, Heidelberg, 3. Auflage (2009)

[15] W.H. Inmon: „Building the Data Warehouse“; Wiley Inc., New York, 3. Auflage (2002)

[16] R. Kimball, M. Ross: „The Data Warehouse Toolkit“; Wiley Publishing Inc., Indianapolis, 2. Auflage (2002)

[17] M. Golfarelli, S. Rizzi: „Data Warehouse Design: Modern Principles and Methodologies“; McGraw-Hill, New York (2009)

[18] P. Vassiliadis, A. Simitsis, S. Skiadopoulos: „Conceptual Modeling for ETL Processes“; in: DOLAP'02 Proceedings, S. 14-20 (2002)

[19] A. Simitsis: „Modeling and managing ETL processes“; in: VLDB'03 PhD Workshop (2003)

[20] A. Simitsis, P. Vassiliadis, T. Sellis: „Optimizing ETL Processes in Data Warehouses“; in: ICDE'05 Proceedings, S. 564-575 (2005)

[21] H. Plattner, A. Bog, J. Schaffner, J. Krueger, A. Zeier. "ETL-less zero-redundancy system and method for reporting OLTP data", U.S. Patent Application Publication (2009)

[22] J. Gray, S. Chaudhuri, A. Bosworth et al.: „Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals“; in: Data Mining and Knowledge Discovery 1(1), S. 29-53 (1997)

[23] H.-J. Lenz, A. Shoshani: „Summarizability in OLAP and Statistical Data Bases“; in: SSDBM'97 Proceedings, S. 132-143 (1997)

[24] H.-J. Lenz, B. Thalheim: „A Formal Framework of Aggregation for the OLAP-OLTP Model“; in: Journal of Universal Computer Science 15(1), S. 273-303 (1997)

[25] U. Leser, F. Naumann: „Informationsintegration“; dpunkt-Verlag, Heidelberg (2007)

[26] T. Griffin, L. Libkin: „Incremental Maintenance of Views with Duplicates“; in: SIGMOD'95 Proceedings, S. 328-339 (1995)

[27] T. Palpanas, R. Siddle, R. Cochrane et al.: „Incremental Maintenance for Non-Distributive Aggregate Functions“; in: VLDB'02 Proceedings, S. 802-813 (2002)

[28] H. Gupta, I.S. Mumick: „Incremental Maintenance of Aggregate and Outerjoin Expressions“; in: Information Systems (31/6), S. 435-464 (2006)

[29] Y. Zhuge, H. Garcia-Molina, J. Hammer et al.: „View Maintenance in a Warehousing Environment“; in: SIGMOD'97 Proceedings, S. 417-427 (1995)

[30] A. Gupta, H.V. Jagadish, I.S. Mumick: „Data Integration using Self-Maintainable Views“; in: EDBT'96 Proceedings, S. 140-144 (1996)

[31] D. Quass: „Maintenance Expressions for Views with Aggregation“; in: VIEWS'96 Proceedings, S. 110-118 (1996)

[32] D. Agrawal, A. El Abbadi, A. Singh et al.: „Efficient View Maintenance at Data Warehouses“; in: SIGMOD'97 Proceedings, S. 417-427 (1997)

[33] K.Y. Lee, J.H. Son, M.H. Kim: „Efficient Incremental View Maintenance in Data Warehouses“; in: CIKM'01 Proceedings, S. 349-357 (2001)

[34] T. Jörg, S. Deßloch: „Towards Generating ETL Processes for Incremental Loading“; in: IDEAS'08 Proceedings, S. 101-110 (2008)

[35] T. Jörg, S. Deßloch: „Near Real-Time Data Warehousing Using State-of-the-Art ETL Tools“; in: BIRTE'09 Proceedings, S. 100-117 (2009)

[36] A. Behrend, T. Jörg: „Optimized Incremental ETL Jobs for Maintaining Data Warehouses“; in: IDEAS'10 Proceedings, S. 216-224 (2010)

[37] T.L. Achs: „Optimierung der materialisierten Sichten in einem Datawarehouse auf der Grundlage der aus einem ERP-System übernommenen operative Daten“; Dissertation, Wirtschaftsuniversität Wien (2004)

[38] MonetDB B.V.: „MonetDB – Column Store Features“; auf: http://www.monetdb.org/Home/Features {01.07.2012} (2011)

[39] H. Plattner, „A common database approach for OLTP and OLAP using an in-memory column database“; in: Proceedings of the 35th SIGMOD international conference on management of data, ACM, S. 1–2 (2009)

[40] A. Kemper, T. Neumann: „HyPer: Hybrid OLTP&OLAP High PERformance Database System“; auf: www3.in.tum.de/research/projects/HyPer/HyperTechReport.pdf {01.07.2012} (2010)

[41] Oracle Corporation: „Extreme Performance Using Oracle TimesTen In-Memory Database“; auf: www.oracle.com/technetwork/database/timesten/overview/wp-timesten-tech-132016.pdf {01.07.2012} (2009)

[42] IBM Corporation: IBM solidDB™; auf: www.ibm.com/software/data/soliddb {01.07.2012} (2010)

[43] SAP AG: „SAP® In-Memory Appliance (SAP HANA™); auf www.sap.com/platform/in-memory-computing/in-memory-appliance/index.epx {01.07.2012} (2011)

[44] T. Winsemann, V. Köppen, G. Saake: „Advantages of a Layered Architecture for Enterprise Data Warehouse Systems“; 2nd International Conference on Complex Systems Design & Management, Paris (Frankreich), 7. - 9. Dezember 2011; auf: www.csdm2011.csdm.fr/-Posters,138-.html {01.07.2012} (2011)

[45] T. Winsemann, V. Köppen, A. Lübcke, G. Saake: „A Layered Architecture Approach for Large-Scale Data Warehouse Systems“; 4th International United Information Systems Conference 2012, Jalta (Ukraine), 1. - 3. Juni 2012, akzeptiert zur Veröffentlichung in: LNBIP, Springer-Verlag, Heidelberg (2012)

[46] T. Winsemann, V. Köppen, G. Saake: „A Layered Architecture for Enterprise Data Warehouse Systems“; 10th International Workshop on System/Software Architectures, Danzig (Polen), 25. Juni 2012; in: CAiSE 2012 Workshops, LNBIP 112, S. 192-199, Springer-Verlag, Heidelberg (2012)

[47] T. Winsemann, V. Köppen, „Kriterien für Datenpersistenz bei Enterprise Data Warehouse Systemen auf In-Memory Datenbanken“; 23. GI-Workshop Grundlagen von Datenbanken 2011, Obergurgl (Österreich), 31. Mai - 3. Juni 2011, in: Proceedings of the 23rd GI-Workshop „Grundlagen von Datenbanken 2011“, S. 97-102 (2011)

[48] T. Winsemann, V. Köppen, „Persistence in Enterprise Data Warehouses“; Technical Report FIN-002-2012, Universität Magdeburg (2012)

[49] T. Winsemann, V. Köppen: „Persistence in Data Warehousing“; 6th International Conference on Research Challenges in Information Science, Valencia (Spanien), 16. - 18. Mai 2012; in: RCIS 2012 Conference Proceedings, S. 445-446, IEEE (2012)

[50] B.A. Devlin: „Business Integrated Insight (BI²)“; auf: www.9sight.com/bi2_white_paper.pdf {01.07.2012} (2009)

[51] P. Lehmann: „Meta-Datenmanagement in Data-Warehouse-Systemen“; Dissertation, Universität Magdeburg (2001)

[52] Y. Cui, J. Widom: „Lineage Tracing for General Data Warehouse Transformations“; in: The VLDB Journal 12(1), S. 41-58 (2003)

[53] §§239,257 HGB (Stand: 01.03.2011); §25a KWG (Stand: 01.03.2011); §147 AO (Stand: 08.12.2010)

[54] §13 ProdHaftG (Stand: 19.07.2002)

[55] W. Lehner: „Datenbanktechnologie für Data-Warehouse-Systeme“; dpunkt-Verlag, Heidelberg (2003)

[56] C. Schneeweiß: „Planung 1: Systemanalytische und entscheidungstheoretische Grundlagen“; Springer-Verlag, Berlin (1991)

[57] D. Delić: „Ein multiattributives Entscheidungsmodell zur Erfolgsbewertung nicht-kommerzieller Webportale“; Dissertation, Freie Universität Berlin (2008)

[58] T. Saaty: „The Analytic Hierarchy Process for Decisions in a Complex World“; McGraw-Hill, New York (1980)

[59] B. Berendt, V. Köppen: „Improving Ranking by Respecting the Multidimensionality and Uncertainty of User Preferences“; in: "Intelligent Information Access", S. 39-56; Springer-Verlag, Berlin (2010)