Manuscript

BILDVERARBEITUNG FÜR DIE MEDIZIN 2005

Algorithmen – Systeme – Anwendungen

13. – 16. März 2005, Heidelberg

dkfz.
DEUTSCHES
KREBSFORSCHUNGSZENTRUM

gmds

IEEE

DAGM BVMI

Manuscript

Marcus Prümmer, Elmar Nöth, Joachim Hornegger, Axel Horndasch:
Mensch-Maschine Interaktion für den interventionellen Einsatz.
In: Bildverarbeitung für die Medizin 2005, Heidelberg, 2005, 485-489

Mensch-Maschine Interaktion für den interventionellen Einsatz

Marcus Prümmer[1], Elmar Nöth[1,2], Joachim Hornegger[1], Axel Horndasch[1,2]

[1] Lehrstuhl für Mustererkennung (Informatik 5),
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Martensstr. 3, 91058 Erlangen, Germany
Tel.: +49-9131-8527775, Fax: +49-9131-303811
{pruemmer,noeth,hornegger,horndasch}@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de
[2] Sympalog Voice Solutions GmbH
Karl-Zucker-Str. 10, 91052 Erlangen, Germany
Tel.: +49-61661-0, Fax: +49-61661-20
{noeth,horndasch}@sympalog.de
http://www.sympalog.de

Zusammenfassung. In diesem Beitrag beschreiben wir die Möglichkeiten der Steuerung von Geräten mittels natürlicher Sprache am Beispiel eines sprachgesteuerten 3D-Gefäßanalysesystems. Das System versteht ganze Sätze und erkennt selbständig, ob eine Äußerung an das System gerichtet ist oder an eine andere Person. Die Sprachsteuerung wurde am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg in Zusammenarbeit mit der Firma Sympalog Voice Solutions GmbH für ein Gerät zur Stenosenvermessung der Firma Siemens Medical Solutions (Leonardo Workstation) entwickelt und erfolgreich einer klinischen Erprobung unterzogen.

1 Problemstellung

Wenn die Durchblutung des Gehirns durch Gefäßengstellen beeinträchtigt wird, kann ein Stent (maschenförmige Gefäßwandnachbildung) das verengte Gefäß von innen her offen halten. Zur Risikoeinschätzung und für die richtige Wahl des Stents ist eine 3D-Darstellung und quantitative Auswertung der rekonstruierten Gefäße (Abb.1) von großer Bedeutung. Die Visualisierung der 3D-rekonstruierten Gefäße erfolgt während des klinischen Workflows und wird weitgehend per Joystick gesteuert, da andere Eingabegeräte aus hygienetechnischen Gründen am OP-Tisch nicht verwendbar sind. Für die Auswertung der rekonstruierten Gefäße muss der Arzt den OP-Tisch - und somit auch den sterilen Bereich - verlassen, da sich die Workstation in einem abgetrennten Bereich befindet. Daher is es wünschenswert, per Sprachsteuerung die bildverarbeitenden Algorithmen zur Visualisierung und Quantifizierung direkt vom OP-Tisch aus zu bedienen.

2 Stand der Forschung

Die Mensch-Maschine-Interaktion (MMI) ist bisher im interventionellen Umfeld auf herkömmliche Eingabegeräte wie Maus und Joystick beschränkt. Eine

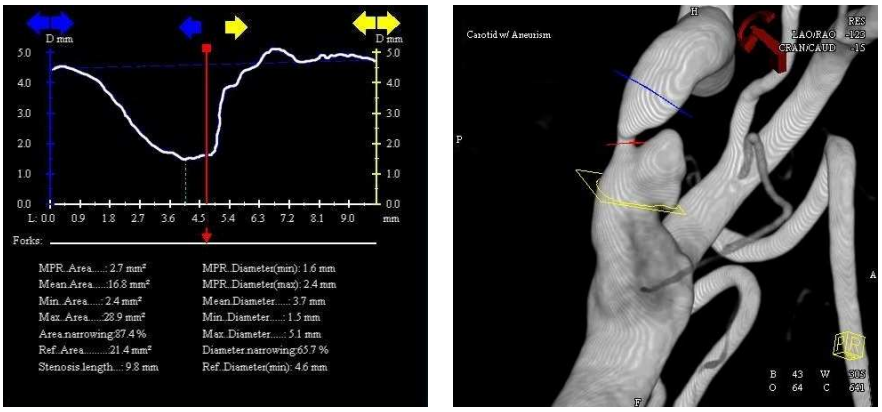

Abb. 1. Gefäßprofil einer Stenose (links) und der quantifizierte Gefäßabschnitt (rechts).

hilfreiche Ergänzung dieser MMI stellt die Sprachsteuerung dar. Um Problemstellungen hinsichtlich der Sicherheit und Zuverlässigkeit besser verständlich zu machen, wollen wir sprachgesteuerte MMI-Systeme in verschiedene Dimensionen kategorisieren:

1. **Kommando- und Kontroll-Systeme:** Sehr einfache Systeme dieser Art ordnen einer Benutzeräußerung genau einen Befehl in einer Liste zu. Der Ansatz stößt sehr schnell an Grenzen, wenn es um die Parametrierbarkeit der Befehle geht (z.B. *rotiere nach links um <n> Grad*). Um nicht alle kombinatorisch möglichen Äußerungen speichern zu müssen, kann man die erkannte Wortkette mit Hilfe einer Grammatik interpretieren und sprecherunabhängige Verfahren wie *Hidden Markov Modelle* für die Erkennung einsetzen.
2. **Dialog-Systeme:** Enthält ein Befehl an ein sprachgesteuertes System nicht genügend Information oder ist aus Sicherheitsgründen vor Ausführung des Kommandos eine Bestätigung notwendig, muss ein Dialogsystem verwendet werden. Im folgenden Beispiel interagiert ein Benutzer mit einem C-Arm: [B:] *Rotiere den C-Bogen.* [S:] *Für diesen Befehl ist eine Richtungsangabe mit Gradzahl notwenig. Um vieviel soll sich der C-Bogen in welche Richtung drehen?* [B:] *Nach links um 45 Grad.*
3. **Benutzer-/system-initiierte Interaktion:** Die Art der Interaktion und die Komplexität des Systems werden stark davon beeinflusst, ob das System nur auf Benutzeräußerungen reagieren kann, oder auch von sich aus Interaktionen initiieren kann, etwa um den Benutzer auf eine kritische Situation aufmerksam zu machen. Beispiel *intelligenter C-Bogen*: Ist ein C-Bogen an das visualisierte Gefäß gekoppelt und der Arzt wählt eine neue Ansicht, so dass der C-Arm in den Tisch oder Patienten fahren würde, gibt das System eine Warnung aus: *Bitte wählen sie eine andere Ansicht!*

4. **Push-to-talk-Syteme:** In vielen Situationen kann es sinnvoll sein, das Interaktionssystem erst durch einen Funktionsknopf zu aktivieren. Der Funktionsknopf bzw. -hebel hat den Vorteil, dass die Aktivierung *fehlerfrei* ist. Die Steuerung eines C-Bogen wäre eine typische Anwendung, da sicherheitsrelevante Aktionen ausgeführt werden können.
5. **Online-Systeme:** Systeme dieser Art müssen nicht aktiviert werden, hören also ständig zu. Damit sie trotzdem nur auf relevante Befehle reagieren, müssen alle möglichen irrelevanten Äußerungen modelliert werden. Der Vorteil der Lösung ist, dass der Arzt zwischen den für das System relevanten Befehlen beliebig mit seiner Umwelt kommunizieren kann.

3 Methoden

Am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg wurde zusammen mit der Firma Sympalog Voice Solutions GmbH eine Sprachsteuerung für ein Gerät zur Stenosenvermessung (Leonardo Workstation der Firma Siemens Medical Solutions) entwickelt und einer klinischen Erprobung unterworfen. Das System ist sprecherunabhängig, versteht ganze Sätze und erkennt selbständig, ob eine Äußerung an das System gerichtet ist oder an eine andere Person. Somit kann zusätzlich zu den bisherigen Eingabegeräten die Visualisierung der Gefäße und eine vollständige Gefäßanalyse via Sprache direkt am OP-Tisch eingestellt und durchgeführt werden. Die dialogorientierte Sprachsteuerung bietet den Vorteil, dass alle Steuerbefehle in einer Kommandoebene angeordnet und somit umständliche hierarchische Menüs vermieden werden können. Für die Visualisierung verschiedener Gewebearten oder Knochen können vordefinierte Histogramm-Einstellungen abgerufen werden. Ebenso können spezielle Ansichten wie beispielsweise *zeige mir den Patient von Oben/Rechts* direkt angewählt werden.

3.1 Ein sprachgesteuertes Stenose-Vermessungsmodul

Da eine 3D-Selektion und Gefäßnavigation mehrere Freiheitsgrade besitzt, muss die Orientierung innerhalb des Gefäßsystems algorithmisch vorgegeben werden. Nur dadurch kann eine einfache Navigation mit *vor/zurück* oder *rechts/links* gewährleistet werden. Dafür wird zuerst eine semiautomatische Schwellwertsegmentierung berechnet, die noch durch Sprachbefehle wie *Erhöhe bitte den Schwellwert um 32* feinjustiert werden kann. Für die Selektion einer Stenose werden das Skelett des Gefäßbaumes und die Gefäßverzweigungen bestimmt. Die Selektion erfolgt mittels eines 3D-Zeigers, der permanent einem sprachgesteuerten 2D-Cursor entlang der Skelettpfade folgt. Per 2D-Cursor kann man beliebig in der Bildebene navigieren, das 3D-Gefäßbild rotieren (*rotiere das Volumen nach rechts*) und den Zoomfaktor verändern (*vergrössere/verkleinere das Volumen*). Somit lässt sich der gewünschte 3D-Gefäßabschnitt schell und unkompliziert anwählen. Der zu analysierende Gefäßabschnitt wird automatisch ausgehend von der selektierten Position zu beiden Seiten entlang des Skelettpfades quantifiziert und kann via Sprache beliebig angepasst werden. Ein mit einem

Funkmikrofon ausgestatteter Arzt ist damit in der Lage, eine Gefäßanalyse direkt am OP-Tisch via Sprache durchzuführen.

3.2 Sprachliche Mensch-Maschine-Interaktion

Im folgenden soll der im Stenose-Analyse-System verwendete Erkenner kurz charaktersiert werden: Im System wurde der Sprecherkenner *SymRec* der Firma Sympalog eingesetzt. Er basiert, wie praktisch alle im wissenschaftlichen und kommerziellen Bereich verfügbaren sprecherunabhängigen Erkenner, auf der Hidden-Markov-Technologie. Ein vergleichbarer Spracherkenner aus dem wissenschaftlichen Umfeld und aktuelle Forschungsarbeiten zu diesem Thema sind in [1,2] beschrieben. Im Sinne der Kategorisierung des letzten Kapitels handelt es sich um ein Kommando- und Kontroll-System. Alle Interaktionen sind benutzerinitiiert und es handelt sich um ein Online-System. Der Erkenner hat einen Anwendungswortschatz von 275 Wörtern (je ca. 50% deutsch und englisch). Zur Kompensation von quasi beliebigen Äußerungen außerhalb des Anwendungsbereichs wird ein komplexes Hintergrundmodell verwendet. Der Erkenner hat nur ein stochastisches Sprachmodell, d.h. der Systemzustand des Anwendungssystems *Stenose-Analyse* wird nicht ausgenutzt, um Befehle, die im aktuellen Zustand des Systems nicht sinnvoll sind, von der Erkennung auszuschließen. Sobald der Erkenner einen möglichen Befehl erkannt hat, liefert er die am wahrscheinlichsten gesprochene Wortkette an das Verstehensmodul. Dieses sucht mit Hilfe von 20 endlichen Automaten (sogenannten *Infoscannern*) in der Wortkette nach Unterketten, die gültige Befehle darstellen. In der folgenden Liste von Beispielbefehlen sind Wörter in () optional und Wörter in {} Parameter: *erhöhe/erniedrige den Schwellwert um* {*Zahl*}*, analysiere (die) Stenose, analyze (the) stenosis, rotiere (das) Volumen, rotate (the) volume, bewege (den) Zeiger nach* {*links, rechts, unten, oben*}*, (bewege die) blaue Ebene nach* {*links, rechts*}*, schneller* . Der erkannte Befehl wird an die Anwendung Stenose-Vermessung weitergegeben. Diese überprüft, ob der Befehl im aktuellen Kontext sinnvoll bzw. möglich ist (z.B. setzt der Befehl *schneller* voraus, dass vorher ein Bewegungs-/Rotationsbefehl ausgeführt wurde). Falls der Befehl möglich ist, wird er ausgeführt, falls nicht, wird er ignoriert. Es wird keine Fehlermeldung oder Bestätigung ausgegeben. Der Erkenner ist so implementiert, dass er immer in etwa in Echtzeit arbeitet, d.h. dass immer wenige Millisekunden nach Erkennung einer Sprechpause von 200 Millisekunden der erkannte Befehl an die Anwendung übergeben wird. Dies wird dadurch erreicht, dass der Erkenner bei der Suche nach der besten Wortkette mehr/weniger Alternativen zulässt, wenn er mehr/weniger CPU-Zeit zur Verfügung hat. Somit wirkt sich eine schlechtere Rechnerausstattung weniger auf die Reaktionszeit und mehr auf die Güte der Erkennung aus (weniger Alternativen bedeutet, dass gelegentlich ein gültiger Befehl verworfen wird). Eine „vernünftige“ Minimalanforderung an das System (nur für das Spracherkennungsmodul) ist: Pentium III - CPU mit 800 MHZ und 512 MB Hauptspeicher.

4 Ergebnisse

Die Sprachsteuerung wurde von den in die Erprobung eingebundenen Ärzten insgesamt sehr positiv beurteilt. Besonders für die Bedienung von Geräten im sterilen Umfeld sehen sie durch die Sprachbedienung eine spürbare Erleichterung im beruflichen Alltag. Als möglicherweise problematisch für die Akzeptanz eines solchen Systems wird lediglich die Tatsache beurteilt, dass das System das Tragen eines Headsets verlangt. Obwohl sich hier für bestimmte Einsatzszenarien sehr komfortable Lösungen finden lassen (etwa die Integration des Mikrophons in den Mundschutz) wird in Zweifel gezogen, ob jeder Arzt bereit ist, für den Komfort- und Effizienzgewinn bei der Bedienung des Systems die hierfür notwendigen Handgriffe in Kauf zu nehmen. Ohne besondere technische Maßnahmen sinkt die Erkennungsqualität mit zunehmendem Abstand zwischen Mund und Mikrophon drastisch. Verantwortlich hierfür sind der abnehmende Signal-Rausch-Abstand des Signals, Störgeräusche und besonders die auftretenden Echo- bzw. Hall-Effekte.

5 Diskussion

Um zukünftig ein freies Sprechen mit Sprachsteuerungen ohne Headset zu ermöglichen, arbeiten wir zur Zeit intensiv an der Verwendung von Mikrophon-Arrays als Eingabemedium für die Spracherkennung und der damit verbundenen Optimierung des Spracherkenners. Mikrophon-Arrays fokussieren auf den Sprecher und kompensieren dadurch bis zu einem gewissen Grade störende Geräuschquellen, soweit sich diese nicht in Richtung des Sprechers befinden. Ohne weitere Maßnahmen sind jedoch bereits Entfernungen von mehr als 0,5 Meter problematisch. Durch ein neu entwickeltes Verfahren zur Anpassung des Spracherkenners an solche Bedingungen ist es uns bereits gelungen, auch bei einem Abstand von 1,5 Meter und mehr den überwiegenden Teil der durch die verbleibenden Störungen verursachten Erkennungsfehler zu kompensieren ohne hierfür neue, entsprechend gestörte Sprachdaten zur Adaption des Spracherkenners aufnehmen zu müssen. Als eine realistische Alternative zur Verwendung eines Headsets erscheint daher etwa im Fall der sprachgesteuerten 3D-Gefäßanalyse das Anbringen eines Mikrophon-Arrays im Bereich des Bildschirms, der an einem Schwenkarm oberhalb des Patienten positioniert ist. Ob die so im OP-Alltag erreichbare Erkennungsgenauigkeit ausreicht, um die Funktionalität des bestehenden Systems in vollem Umfang beizubehalten, ist derzeit noch eine offene Frage.

Literaturverzeichnis

1. G. Stemmer: Modeling Variability in Speech Recognition, Chair for Pattern Recognition, University of Erlangen-Nuremberg, 2004.
2. F. Gallwitz: Integrated Stochastic Models for Spontaneous Speech Recognition, Berlin 2002, Logos Verlag, Studien zur Mustererkennung, vol. 6, ISBN: 3-89722-907-2.