(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
24. April 2003 (24.04.2003)

PCT

(10) Internationale Veröffentlichungsnummer
WO 03/034403 A1

(51) Internationale Patentklassifikation[7]: G10L 15/06, G06T 15/70, G10L 13/04

(21) Internationales Aktenzeichen: PCT/EP02/11016

(22) Internationales Anmeldedatum: 1. Oktober 2002 (01.10.2002)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
01124642.8 15. Oktober 2001 (15.10.2001) EP

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **SIEMENS AKTIENGESELLSCHAFT** [DE/DE]; Wittelsbacherplatz 2, 80333 München (DE).

(72) Erfinder; und
(75) Erfinder/Anmelder *(nur für US)*: **LUKAS, Klaus** [DE/DE]; Niemöllerallee 6, 81739 München (DE).

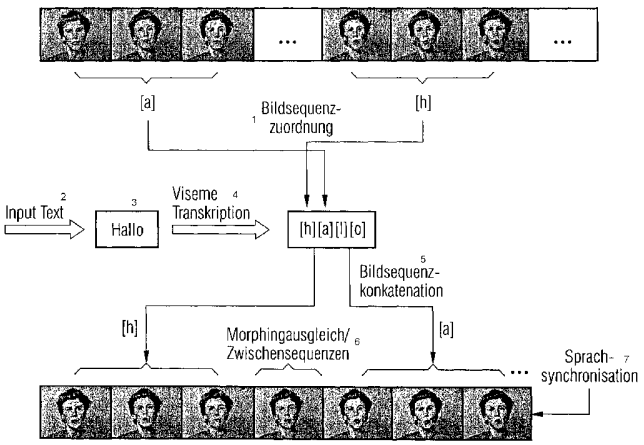
(74) Gemeinsamer Vertreter: **SIEMENS AKTIENGESELLSCHAFT**; Postfach 22 16 34, 80506 München (DE).

(81) Bestimmungsstaat *(national)*: US.

[Fortsetzung auf der nächsten Seite]

(54) Title: METHOD FOR IMAGE-ASSISTED SPEECH OUTPUT

(54) Bezeichnung: VERFAHREN ZUR BILDUNTERSTÜTZTEN SPRACHAUSGABE



(57) **Abstract:** The invention relates to a method for image-assisted speech output of a text, converted into a speech signal sequence, whereby a continuous moving face image is output synchronously with the speech, whereby short image sequences, previously recorded, of a natural person are synchronously allocated to sections of the text to be outputted on the pronunciation of predetermined speech elements or samples and the continuous moving image is produced from said image sequences.

(57) **Zusammenfassung:** Verfahren zur bildunterstützten Sprachausgabe von in eine Sprachsignalfolge gewandeltem Text, bei dem ein kontinuierliches Bewegtbild eines Gesichtes synchron zur Sprache ausgegeben wird, wobei vorab aufgenommene kurze Bildfolgen des Gesichtes einer natürlichen Person bei der Aussprache vorbestimmter Sprachelemente bzw. -muster Textabschnitten des auszugebenden Textes synchron zugeordnet werden und aus den Bildfolgen das kontinuierliche Bewegtbild zusammengesetzt wird.

WO 03/034403 A1

(84) Bestimmungsstaaten *(regional)*: europäisches Patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR).

Erklärungen gemäß Regel 4.17:

— *hinsichtlich der Berechtigung des Anmelders, ein Patent zu beantragen und zu erhalten (Regel 4.17 Ziffer ii) für die folgenden Bestimmungsstaaten europäisches Patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR)*

— *Erfindererklärung (Regel 4.17 Ziffer iv) nur für US*

Veröffentlicht:

— *mit internationalem Recherchenbericht*

Zur Erklärung der Zweibuchstaben-Codes und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

Beschreibung

Verfahren zur bildunterstützten Sprachausgabe

Die Erfindung betrifft ein Verfahren zur bildunterstützten Sprachausgabe nach dem Oberbegriff des Anspruchs 1.

Der Einsatz multimodaler Benutzungsoberflächen gewinnt zunehmend an Bedeutung. Synchronisierte Benutzerinteraktionen mit sprachlichen und visuellen Komponenten erhöhen den Benutzerkomfort und erlauben erweiterte Gestaltungsmöglichkeiten der Benutzerschnittstelle. Ein wichtiger Aspekt ist hierbei die kombinierte verbale und visuelle Ausgabe von dynamisch erzeugten Texten, die aus verschiedenen Kommunikationskanälen wie z.B. Internet-Inhalten, E-Mails oder Datenbank-Suchergebnissen, resultieren und dem Benutzer dargestellt werden sollen.

Für die visuelle Komponente ist der Einsatz von Avataren üblich, d. h. künstlich generierten Charakteren, die nur bedingt ein natürliches Aussehen vorweisen. In der Regel werden künstliche Kopfstrukturen über Gittermodelle erzeugt und mit menschlichen Texturen versehen oder Gesichtspunkte menschlicher Köpfe auf künstlich generierte Körper übertragen. Diese Darstellung ergibt jedoch nur begrenzt eine natürliche Darstellung und erzeugt einen roboterhaften Eindruck.

Eine sprachsynchrone Darstellung natürlicher Menschen zur visuellen Ausgabe von beliebigen Texten ist derzeit nicht bekannt.

Auf der Seite der Sprachausgabe bestehen im wesentlichen zwei Grundmethoden zur Text-To-Speech-Transformation, die formantbasierten Methoden sowie die konkatenierte Sprachsynthese. Die formant-basierte Methode erzeugt mittels Formant-Algorithmen künstliche Sprache, die vorteilhafterweise nur geringe Resourcenanforderungen stellt, aber von der Sprachqualität

beim derzeitigen Stand der Technik nur für kurze Textwiedergaben als geeignet erscheint.

Die konkatenierte Sprachsynthese basiert auf der Zerlegung von vorhandenem natürlichem Sprachmaterial in kleine Abschnitte, wie z. B. Phoneme, und der Zusammensetzung dieser Phoneme im gegebenen Textzusammenhang. Diese Form der Spracherzeugung erreicht einen hohen Grad an Natürlichkeit, benötigt allerdings mehr Ressourcen. Somit ist auf sprachlicher Seite die Natürlichkeit der Ausgabe durchaus bereits gegeben, auf der visuellen Seite bietet der aktuelle Stand der Technik jedoch keine adäquate Qualität.

Der Erfindung liegt daher die Aufgabe zugrunde, ein verbessertes Verfahren der gattungsgemäßen Art zur visuell unterstützten Darstellung von arbiträren Texten anzugeben, um eine gesamtheitlich lebensechte Ausgabe in Sprach- und Visualisierungsform zu erhalten.

Diese Aufgabe wird durch ein Verfahren mit den Merkmalen des Anspruchs 1 gelöst.

Die Erfindung schließt den wesentlichen Gedanken einer grundlegenden Abkehr von der bisherigen Herangehensweise an eine bildunterstützte Sprachausgabe - nämlich der Generierung von Avataren - ein.

Die fliessende visuelle Ausgabe von Bilddaten zu vorgegebenen Textdaten wird stattdessen durch die Konkatenierung von kurzen Abschnitten an Bilddaten erreicht. Durch die Konkatenationsmethode können beliebige Texte in lebensnaher Qualität visualisiert werden. Die verwendeten kurzen Bildsequenzen entsprechen den Visemen (Mundbewegungen) und deren Übergängen. Mittels einer Transkription des Eingangstextes in eine Viseme-Darstellung können die zugeordneten kurzen Bildabschnitte zusammengefügt und mit Übergängen zwischen den einzelnen Visemen zu einem fliessenden Ablauf gebracht werden.

Die Übergänge können gemäß alternativen Fortbildungen des Erfindungsgedankens entweder durch eigene kurze Bildsequenzen oder durch Morphingalgorithmen erzeugt werden, um einen harmonischen Verlauf zu gewährleisten.

Parallel zur Viseme-Transkription erfolgt die Phoneme-Transkription für die Sprachausgabe. Mittels Synchronisationsmechanismen (z.B. Tagging des Bildmaterials) erfolgt eine Synchronisation der Mundbewegungen mit der entsprechenden synthetischen Sprachausgabe. Die zu den Bildsequenzen zugehörigen Original-Sprachdaten können aufgrund des fehlenden Prosodie-Verhaltens vermutlich nicht verwendet werden.

Zur Aufnahme der Sequenzen werden zweckmäßigerweise statische Szenarien (z. B. fixe Hintergründe) benötigt, bei denen nur die Sprechbewegungen der Mundpartie eines Sprechers einen dynamischen Anteil darstellen. Geringfügige Kopfbewegungen o. ä. können durch Normalisierung ausgeglichen werden.

Für den Einsatz des beschriebenen Verfahrens gibt es vielfältige Anwendungsfälle. Beispiele sind das Vorlesen von E-Mails oder SMS mit verschiedenen, zielgruppenangepassten Sprechern (Charakteren), die visuell unterstützte Sprachausgabe nach Abfrage datenbasierter Informationsdienste oder die Ausgabe von Termin- und Adressdaten aus lokalen Organizerdatenbasen eines PDAs. Besondere Vorteile läßt der Einsatz des Verfahrens für visuell unterstützte Sprachausgaben in Ausbildungs-Kontexten erwarten, und hier speziell bei Angeboten, die sich an Kinder oder Jugendliche einerseits oder ältere Menschen oder Hörbehinderte andererseits wenden. Insbesondere die Akzeptanz von Sprachsyntheseanwendungen durch die letztgenannte Zielgruppe dürfte bei Anwendung des Verfahrens stark ansteigen, denn Personen dieser Zielgruppe haben eine tief verwurzelte Abneigung gegen die bisher verwendeten Avatare.

Vorteile und Zweckmäßigkeiten der Erfindung ergeben sich im

übrigen aus den abhängigen Ansprüchen sowie der nachfolgenden Beschreibung eines Ausführungsbeispiels.

Eine schematische Darstellung des Verfahrens anhand eines Ausführungsbeispieles wird in der einzigen Figur gezeigt.

Aus einem beim Sprechen eines vorgegebenen Textes mit einer Vielzahl von Phonemkonstellationen aufgenommenen Sprecher-Bewegtbild werden kurze Bildsequenzen einzelner Viseme gebildet; in der Abbildung beispielhaft für die Viseme [a] und [h]. Hierbei werden gegebenenfalls Normalisierungen hinsichtlich der Bildqualität durchgeführt. Entsprechende Verfahren sind dem Fachmann von Techniken der Trick-Nachbearbeitung bei Trickfilmen und Spielfilmen an sich bekannt.

Ein Input-Text wird in die Viseme-Darstellung transkribiert. Den einzelnen Visemen werden die entsprechenden Bildsequenzen zugeordnet und miteinander konkateniert. Die Glättung der Bildsequenz-Übergänge erfolgt entweder mit nach einem vorbestimmten (ebenfalls an sich bekannten) Morphing-Bildern oder mit Glättungs-Sequenzen, die ebenfalls aus den vorab aufgenommenen Sprecherdarstellungen ermittelt wurden. Zum zeitgleichen Abspielen des Bildmaterials mit künstlich erzeugten Sprachdaten wird eine Synchronisation durchgeführt.

Die Ausführung der Erfindung ist selbstverständlich nicht auf dieses Beispiel und die oben genannten Anwendungsfelder sowie hervorgehobenen Aspekte beschränkt, sondern für beliebige Texte in beliebigen Sprachen und eine Vielzahl weiterer Anwendungen ebenso möglich.

Patentansprüche

1. Verfahren zur bildunterstützten Sprachausgabe von in eine Sprachsignalfolge gewandeltem Text, bei dem ein kontinuierliches Bewegtbild eines Gesichtes synchron zur Sprache ausgegeben wird,
d a d u r c h g e k e n n z e i c h n e t, daß
vorab aufgenommene kurze Bildfolgen des Gesichtes einer natürlichen Person bei der Aussprache vorbestimmter Sprachelemente bzw. -muster Textabschnitten des auszugebenden Textes synchron zugeordnet werden und aus den Bildfolgen das kontinuierliche Bewegtbild zusammengesetzt wird.

2. Verfahren nach Anspruch 1,
d a d u r c h g e k e n n z e i c h n e t, daß
durch einen Morphingalgorithmus subjektiv als fließend wahrgenommene Übergänge zwischen den einzelnen kurzen Bildfolgen gebildet werden.

3. Verfahren nach Anspruch 1 oder 2,
d a d u r c h g e k e n n z e i c h n e t, daß
durch Einfügung von vorab aufgenommenen Glättungs-Einzelbildern oder kurzen Glättungs-Bildfolgen subjektiv als fließend wahrgenommene Übergänge zwischen den einzelnen Bildfolgen gebildet werden.

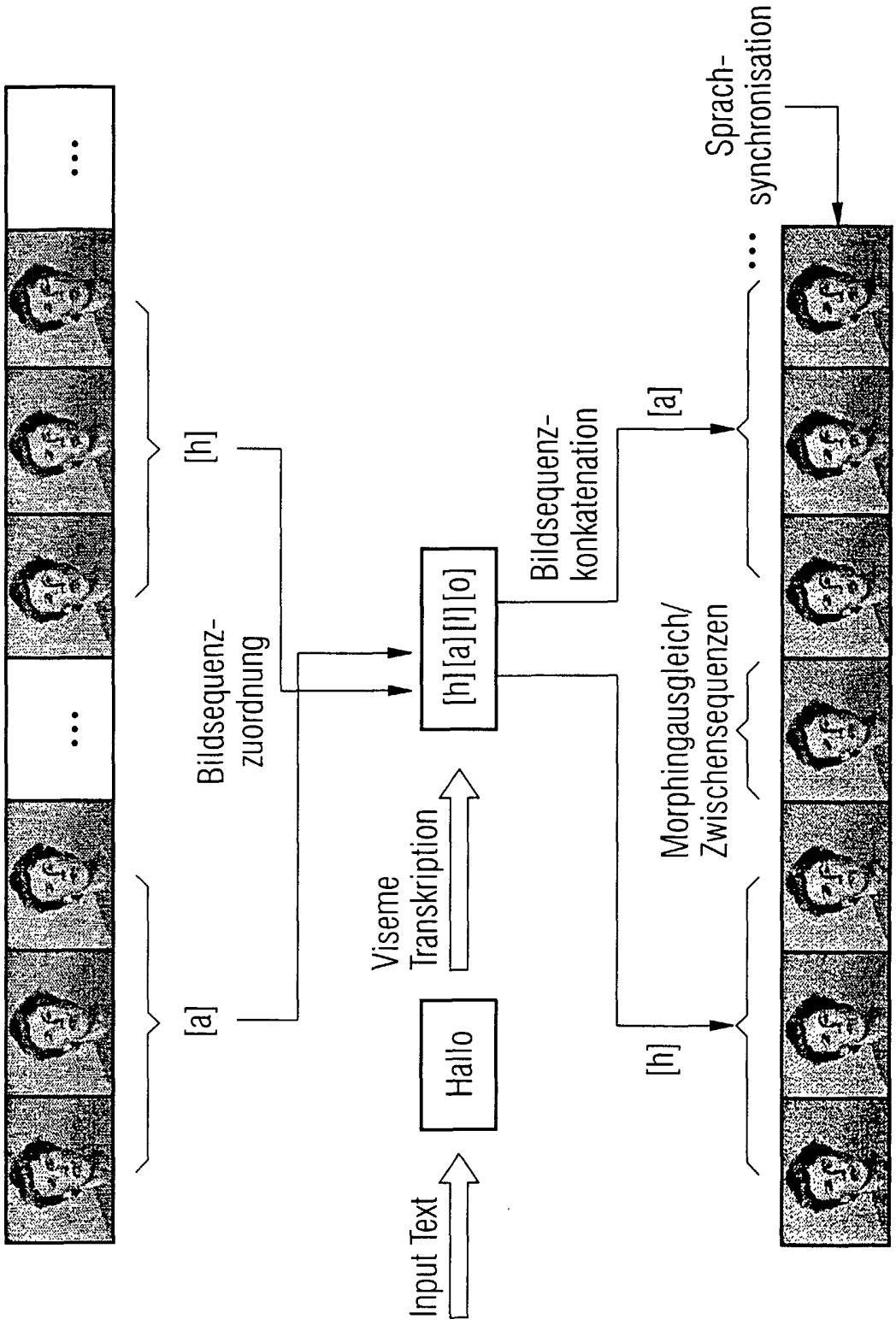
4. Verfahren nach einem der vorangehenden Ansprüche,
d a d u r c h g e k e n n z e i c h n e t, daß
die kurzen Bildfolgen oder das zusammengesetzte Bewegtbild durch einen Tagging-Algorithmus mit den Sprachsignalen synchronisiert werden.

5. Verfahren nach einem der vorangehenden Ansprüche,
d a d u r c h g e k e n n z e i c h n e t, daß
vor einem statischen Hintergrund oder mit Bluescreen-Technik und mit im wesentlichen statischer Sprecherhaltung erzeugte kurze Bildfolgen und wahlweise Glättungs-Einzelbilder bzw.

Glättungs-Bildfolgen verwendet werden.

6. Verfahren nach einem der vorangehenden Ansprüche, dadurch gekennzeichnet, daß Kopfbewegungen des Sprechers auf den vorab aufgenommenen kurzen Bildfolgen durch einen Normalisierungsalgorithmus ausgeglichen werden.

7. Verfahren nach einem der vorangehenden Ansprüche, dadurch gekennzeichnet, daß als kurze Bildfolgen nachträglich aus einem kontinuierlichen primären Bewegtbild, das einem zusammenhängenden Sprachfluß zugeordnet ist, isolierte Segmente verwendet werden.


Input Text
Hallo
Viseme Transkription
[h][a][l][o]
Bildsequenz-zuordnung
[a]
[h]
Bildsequenz-konkatenation
[a]
[h]
Morphingausgleich/ Zwischensequenzen
Sprach-synchronisation

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP 02/11016

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L15/06 G06T15/70 G10L13/04

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
WPI Data, IBM-TDB, COMPENDEX, EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 878 396 A (HENTON CAROLINE G) 2 March 1999 (1999-03-02)	1,3-5,7
Y	abstract column 9, line 8 -column 9, line 38 column 10, line 14 -column 10, line 40	2,6
Y	US 6 232 965 B1 (WATSON STEPHEN HILARY ET AL) 15 May 2001 (2001-05-15)	2,6
A	abstract column 3, line 60 -column 4, line 26 column 5, line 53 -column 5, line 61 column 6, line 19 -column 6, line 45	1,4
A	WO 01 45088 A (INTERACTIVE SOLUTIONS INC) 21 June 2001 (2001-06-21) page 6, line 25 -page 7, line 4	1,7

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
3 January 2003	13/01/2003

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Bourdier, R

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/EP 02/11016

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 5878396	A	02-03-1999	NONE		
US 6232965	B1	15-05-2001	AU	4411596 A	19-06-1996
			US	6097381 A	01-08-2000
			WO	9617323 A1	06-06-1996
WO 0145088	A	21-06-2001	US	6377925 B1	23-04-2002
			AU	1931401 A	25-06-2001
			WO	0145088 A1	21-06-2001

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/EP 02/11016

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 G10L15/06 G06T15/70 G10L13/04

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G10L

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
WPI Data, IBM-TDB, COMPENDEX, EPO-Internal

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	US 5 878 396 A (HENTON CAROLINE G) 2. März 1999 (1999-03-02)	1,3-5,7
Y	Zusammenfassung Spalte 9, Zeile 8 -Spalte 9, Zeile 38 Spalte 10, Zeile 14 -Spalte 10, Zeile 40 ---	2,6
Y	US 6 232 965 B1 (WATSON STEPHEN HILARY ET AL) 15. Mai 2001 (2001-05-15)	2,6
A	Zusammenfassung Spalte 3, Zeile 60 -Spalte 4, Zeile 26 Spalte 5, Zeile 53 -Spalte 5, Zeile 61 Spalte 6, Zeile 19 -Spalte 6, Zeile 45 ---	1,4
A	WO 01 45088 A (INTERACTIVE SOLUTIONS INC) 21. Juni 2001 (2001-06-21) Seite 6, Zeile 25 -Seite 7, Zeile 4 -----	1,7

☐ Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

☒ Siehe Anhang Patentfamilie

° Besondere Kategorien von angegebenen Veröffentlichungen :

"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist

"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist

"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)

"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht

"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist

"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist

"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden

"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist

"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
3. Januar 2003	13/01/2003

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Bourdier, R

INTERNATIONALER RECHERCHENBERICHT

Angaben zu Veröffentlichungen, die zur selben Patentfamilie gehören

Internationales Aktenzeichen
PCT/EP 02/11016

Im Recherchenbericht angeführtes Patentdokument		Datum der Veröffentlichung	Mitglied(er) der Patentfamilie		Datum der Veröffentlichung
US 5878396	A	02-03-1999	KEINE		
US 6232965	B1	15-05-2001	AU	4411596 A	19-06-1996
			US	6097381 A	01-08-2000
			WO	9617323 A1	06-06-1996
WO 0145088	A	21-06-2001	US	6377925 B1	23-04-2002
			AU	1931401 A	25-06-2001
			WO	0145088 A1	21-06-2001