ヒューマンインタフェース 80－6
音声言語情報処理 23－6
（1998. 10. 16）

音声と画像とテキストの相互検索パラダイム

— Multi-modal Example Based Approach —

岡 隆一

新情報処理開発機構 つくば研究センタ

あらまし 音声、静止画、動画、テキストの４つのメディア間で相互に検索を行なうための方法論を議論する。相互検索システムはデータベースの自己組織化アルゴリズムと検索エンジンからなる。データベースの自己組織化アルゴリズムは multi-modal の example data を用いて行なうことが有望である。ここでは、２種類の自己組織化アルゴリズムによるデータベース作成とそれに基づく相互検索の例を示す。

A Paradigm for Mutual Retrieval Between Speech, Image and Text

— Multi-modal Example Based Approach —

Ryuichi Oka

Tsukuba Research Center
Real World Computing Partnership

Abstract A methodology for mutual retrieval among speech, still image, motion image and text databases is discussed. Each retrieval method consists of a self-organizing algorithm and a retrieval engine. The self-organizing algorithm based on multi-modal example databases seems promissing. Two self-organizing algorithms are introduced to obtain organized databases.

1 まえがき

実世界の情報の多くは、音声と画像（動画を含む）とテキストの形で表現されている。具体的な例としては、放送された TV の動画像と音声信号、会話や講演の録音音声、ビデオカメラによる動画像、手話を撮影した動画像、日常の対話におけるジェスチャを撮影した動画像、新聞や雑誌や書籍にあるテキストや写真の静止画像、インターネット上のテキストや静止画、動画、音声信号、などがあげられる。

これらの情報はそれ自体大量であるとともに、日々量的に増大している。通常これらの大部分はその場で使われるだけである。一方、これらの情報の一部は加工された形で、一部は生のままで記録される。記録されたものは再利用されることが期待されているものである。再利用の１つの例として、テキスト、音声、動画像、静止画像、の間で、一つのものから他のものを取り出すという検索がある。例えば、今放送されているニュースの動画像を入力して、類似した動画像を過去の記録から見い出し、そのときに流れていた音声を取り

出して聞いてみたいという検索の要求である。また、何人かで時事問題を（音声で）議論しているとき、討論の音声を認識して関連する過去の新聞記事をその場にある計算機の端末に表示するという検索の要求などの例もある。

このような機能を実現するには、インデックスのついていない、テキスト、音声、動画像、静止画像、からなる実世界情報のそれぞれについて、検索しやすい形で組織化することと、異なるメディアの間での繋がり（共有パラメータ）の情報を利用することが必要になってくる。

マルチメディア実世界情報間の共有パラメータは、時間や状況を共有するデータの収集によって基本的にはうることができる。単一メディア内の類似度に基づく検索がなされるとすると、その結果にリンクする共有パラメータを通じて異なるメディアの検索が可能になる。このような方式は**マルチメディア事例ベースの検索方式** [1] とよばれる。

ここで、マルチメディアの検索についてのこれまでの研究に若干触れておこう。文献 [2] は印象語による画像データベースの検索方式を提案しており、文献 [3] には、画像の特徴からのキーワード定義に基づく画像検索の提案があり、文献 [4] では、カットのコーディングによる映像シーンの識別法を提案している。また、国際会議の論文を集めた「Intelligent Multimedia Information Retrieval」[5] がよく知られているが、この文献中での章立ては、

Section 1: Content based Retrieval of Imagery

Section 2: Content-based Retrieval of Graphics and Audio

Section 3: Content based Retrieval of Video

Section 4: Speech and Language Processing for Video Retrieval

Section 5: Architectures and Tools

Section 6: Intelligent Hypermedia

Section 7: Emperical Evaluation

となっている。ここでは、音声や画像を対象としてそれに含まれる content の抽出とそれを用いた種々の検索方式と実験結果が数多く報告されている。また、1998 年の 11 月に、1st International Conference on Advances Multimedia Content Processing (AMCP '98) が大阪で開かれる予定となっている。ここでも、マルチメディアデータの content による検索方式の発表が主となっている。検索を content の情報で行なう研究は今後も盛んになると思われる。

さて、本稿で議論するマルチメディア間の相互検索はかならずしも content のみによるものではない。content はマルチメディア情報において、言語的な表現に直接結び付く側面にしかすぎない。マルチメディアの相互検索では、まず 同一メディアにおける検索が行なわれ、その後それと異なるメディアの検索へと繋がるものである。したがって、同一メディア内における検索が良好に行なわれることが前提である。それには、大量の生データが検索可能となるように自動的に組織化される必要がある。また、この組織化が異なるメディアとの相互作用によってなされると後の相互検索がより良好に行なわれる。なぜなら、2 つのメディアにおいてそれぞれの中に定義される "位相" が、異なるデータ間でも保存されることによって、メディア間をスムーズに繋ぐことができるためである。この位相を利用すると相互検索は、1 つのメディアで類似しているものが、対応する他のメディアの中でも類似しているということを利用して行なわれる。この考え方に基づくと、content もその "位相" をどのようにもっているかによって新たに評価される。例えば、音声の認識により言語記号としてとり出された音声波形の content データ集合はその要素の間でどのような近さ関係をもっているかによって評価され、その結果として相互検索に利用できるかどうかが判断される。

相互検索の中で重要な役割を果たす、データの近傍関係が良好に表現する "位相" は、従来の content data の中にも考えることができ、一般にはより広い概念である。マルチメディアの生データを自己組織化するアルゴリズムの役割は bottom up 的にその生データに "位相" を与えることにある。

2 検索の型

前節の議論を踏まえると、マルチメディア間の相互検索を考えるには "検索の型" というものを定義しておくことが便利である。"検索の型" としては次の 4 つのものがありえる。

- 信号 - 信号 検索
- 信号 -Content 検索

- Content- 信号 検索
- Content-Content 検索

第1の信号 - 信号検索とは、例えば、音声波形区間から類似した音声波形区間を検索したり、静止画から類似静止画を検索することであり、同一メディア内の検索といえるものである。第2の信号 - Content 検索とは、例えば、音声波形区間からそこで話されている意味に関連したテキストを検索したり、画像からその印象語を検索したりすることである。この検索には、信号の“認識・理解”とみなせる部分もある。第3の Content- 信号 検索 とは、例えば、テキストを与えて、それに関連した音声波形区間を検索したり、テキストを与えてそれに関連した画像を検索することである。このとき、音声波形や画像にはテキストラベルがつけられていないことが前提になる。もし、検索される音声や画像にテキストラベルがつけられていれば、それは実質的にはテキスト - テキスト検索というものになるからである。第4の Content-Content 検索とは、従来もっとも通常に行なわれているもので、テキストからテキストの検索や keyword からテキストの検索、テキストからその要約文の検索などが含まれる。ここでの content と data との違いについては明確にし難いが、狭い意味での content は、単語、単語集合、テキストという記号で表現されるとしてよい。

通常、信号は数値やパターンで表され、また狭い意味の content は単語やテキストで表せることから、信号と content との間での検索はメディア間の検索ということができる。その意味では文字認識や音声認識も信号による content の“検索”といってもよいが、ここでは検索されるべきものの“一意性”が強調される（例えば認識率を問題とする）ということで、前節や次節で述べられている“位相”の積極的な利用という側面が抜け落ちている。

3 メディア間検索の意味

いま、画像をメディア A とテキストをメディア B としよう（図1を参照）。query の画像 A_x を入力して、この query に最も近いメディア A の画像 A_y 、あるいは query に近いメディア A の画像群 A_Y を決定する。そのとき、メディア A の画像とメディア B のテキストとの間に対応がとれていれば、A_y に対応する B_y を検索結果とすることができる。この結果、ユーザは画像 A_x によってテキストの情報 B_y を得ることができ、A_x について、メディア A にはない質的に異なるメディア B の情報をうることができる。これがメディア間検索の意味であるが、検索結果は単に B_y ではなく、A_Y を加工したテキストによることもある。このとき、検索が良好に働くためには、メディア A の要素の近さ関係とメディア B での対応する要素の近さ関係とが保存されていることである（位相の保存）。

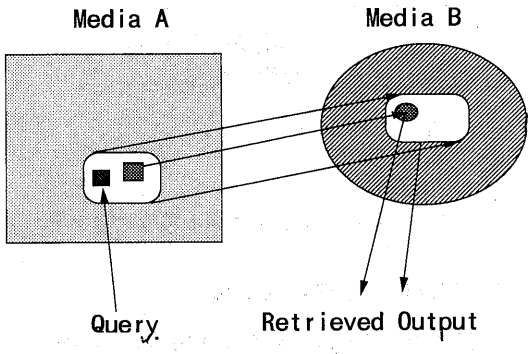

図 1: Inter-Media Retrieval

通常われわれの対象とするメディアは、音声、静止画、動画、テキストの4種のものが考えられるので、図2のような相互の検索関係とその解くべき検索課題があることになる。

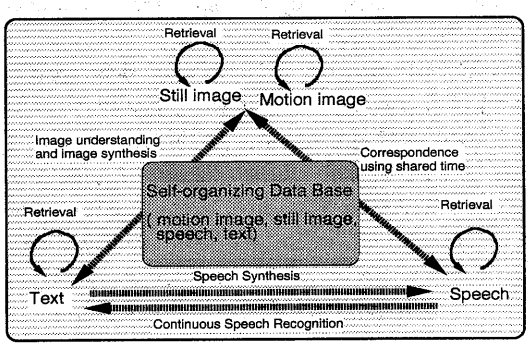

図 2: Mutual retrieval system among multi-modal informations

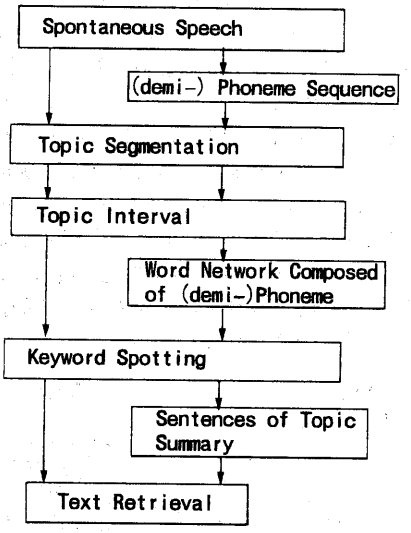

図 3: 音声によるテキスト検索の流れ

4 音声によるテキスト検索

通常の会話や講演の音声波形から、関連するテキストを検索することを考える。このとき、queryとなる音声区間はtopicを表す音声区間であり、この区間からの記号的言語情報がテキストと結びつく（図3参照）。通常の会話や講演の音声波形からのtopic区間のspotting検出については、IRIFCDPというものを用いて行なう試みがある[6]（図4参照）。topic区間抽出後はその区間における大語彙のkeyword spottingによるkeyword群あるいはシソーラス辞書などを用いてtopicの要約文をqueryとして、テキスト検索（後述）を行なうことになる。大語彙の単語スポッティングは困難な課題であるが、音声波形から得られる記号的な言語情報として、波形からframeごとに音素片記号を認識するなど単語より細かな単位の記述を用い[7]（図5参照）、この単位で単語ネットワークを記述することで大語彙の単語スポッティングを実現することも1つの方法である[8]。

5 音声と動画像の相互検索

TV放送データは検索の利用度が高い。TV放送データは音声部分と動画部分があり、それらは時刻を共有している。音声をqueryとして音声

frame	1st	2nd	3rd	4th	5th	6th
.	.	.	.	.	.	.
1420	QP	QT	N'P	QK	N'F	N'H
1421	QP	QT	QK	N'F	HH	N'P
1422	QP	QT	QK	N'F	HH	
1423	QP	QT	QK	HH		
1424	PP	TT	QP	QT	QK	
1425	PP	TT				
1426	PP	TT	PO			
1427	PO	TO	PP	TT		
1428	FO	PO	TO	HO	PP	
1429	PO	OO	TO	HO		
1430	PO	OO	HO	TO		
1431	OO	PO	AN'	TO		
1432	OO	AN'	PO	TO		
1433	OO	ON'	AN'	PO		
1434	ON'	OO	AN'	PO		
1435	ON'	OO	NN'	OP	OK	
1436	ON'	OK	NN'	OO	OP	
1437	ON'	OK	NN'	OO	OP	
1438	ON'	OK	NN'	OP		
1439	OK	ON'	NN'	OP		
1440	ON'	QK	OK	N'F	QP	
1441	ON'	QK	N'F	OK	QP	
1442	QK	N'F	ON'	OK	QP	
1443	QK	N'F	QP	ET	ON'	OK
1444	QK	QP	QT	N'F	OK	
1445	QK	QP	QT	N'F		
1446	QK	QP	QT			
1447	PP	KK	QK	TT	QP	
1448	PP	KK	TT	PE		
1449	PP	KK	PE	TT	PY	
1450	PY	KK	KE	PE		
1451	YO	PY	KYO	KE		
1452	YO	IO	KYO	PY	KE	
1453	YO	IO	KYO	PY	KE	
1454	YO	IO	KYO	PY		
1455	YO	IO	KYO			
1456	YO	IO	KYO	OO		
1457	OO	YO	IO	KYO		
1458	OO	YO	KYO	OU		
1459	OO	YO	OU	OG	KYO	
1460	OO	OG	OU	YO	OR	
1461	OO	OU	OG	OR	YO	
1462	OO	OU	OG	OR		
1463	OO	OU	OR	OG		
1464	OG	OO	OP	OT	OB	OK
1465	OG	OT	OP	OO	OB	OK
1466	OG	OT	OP	OO	OB	OK
1467	OG	OT	OP	OO	BB	QK
1468	QP	QT	BB	OG	OH	OO

図 5: 波形からの音素片系列出力。連続音声中の「東京」部分のフレーム毎の音素片 (best six candidates)。

を検索し、検索された音声区間の動画区間を出力とすれば、音声による動画の検索が可能となる。逆も同様に考えると、音声 - 動画像の相互検索が実現できる[9]。ここでのkey technologyは大量の音声と動画像をコンパクトに組織化しかつqueryによってそれらを取り出すところにある。時系列データの組織化の方式としてIncremental Path Method（IPM）[10]、endless時系列からのネットワーク型データベースからのスポッティング検索方式としてContinuous Automaton [11, 12]を用いた、音声 - 動画像相互検索システムの構成を図6に示す[9]。ここで音声と動画像はいずれもVQ code列への変換が行なわれる。IPMとは、時系列データの（相互結合型）ネットワークデータベースを、図7のような単位ネットを図8のルールで自動的に組み合わせて構成するもので

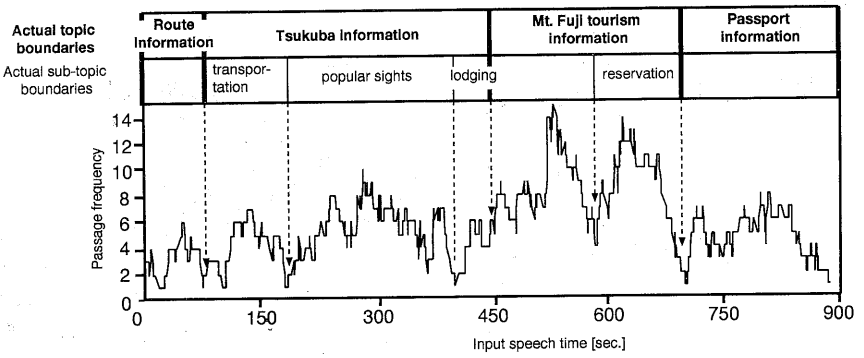

図 4: 話題分節の実験例。IRIFCDP 出力の running histogram の dip が話題の切れ目候補を与える。

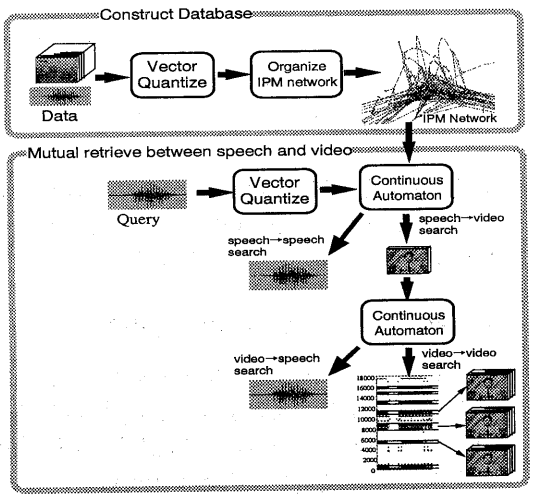

図 6: Concept of mutual spotting retrieval between speech and video

ある。Continuous Automaton はネットワーク内の最適 path を切れ目のない入力時系列中からスポッティング的に取り出す方式で、図 9 に DP における local path との対比を示す。図１０に、query sequence としての音声波形中で、「消費税」という部分が音声データベース中の「「消費税」に類似しての部分」が複数区間スポッティングされ、その時間区間に対応する動画（図１０では１フレーム部分のみ表示）が検索されている様子が示されている。

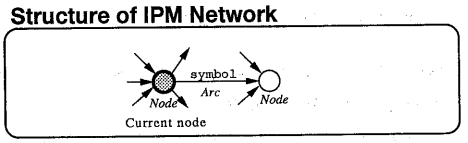

図 7: Basic structure of IPM network

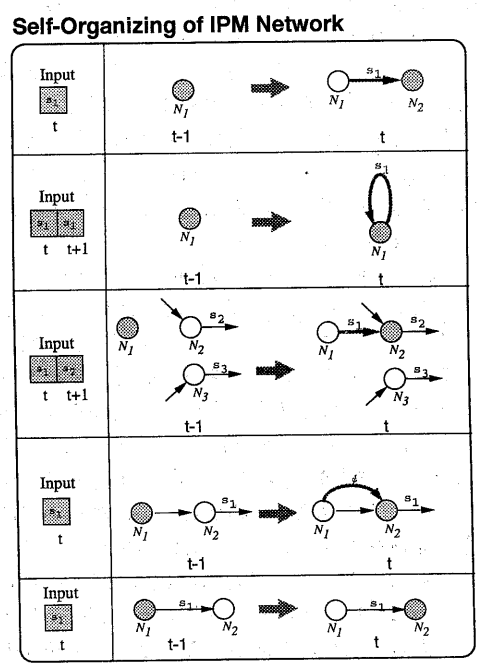

図 8: Rules of Self-Organization by IPM

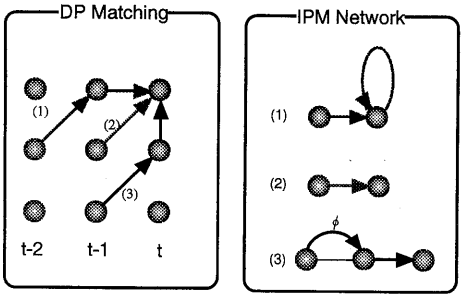

図 9: Comparision of local paths between DP and Continuous Automaton. Continuous Automaton is applied to an input time sequence for spotting retrieval in the IPM network database.

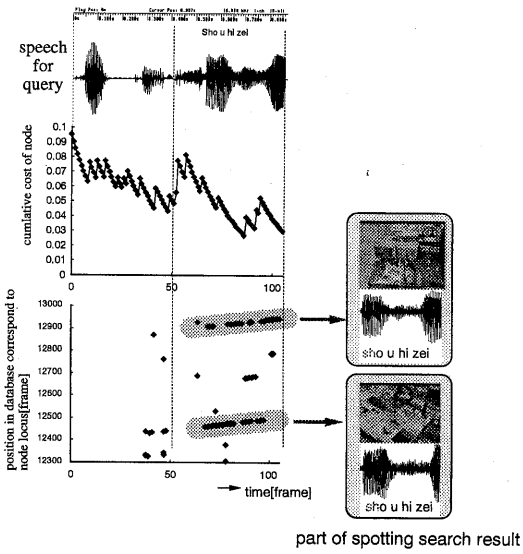

図 10: Video spotting search result from speech

6 テキスト - テキスト検索

記号化された情報であるテキストは、従来検索システムの典型的な対象になっている。その検索システムは、テキスト以外の画像や音声との相互検索システム構築を考えるとき、新たな評価をうける。その評価とは、テキストのデータベース組織化と検索方式が他のメディアとの相互検索に向いているかというものである。データベース組織化の従来手法として、各テキストについて出現単語を単語ベクトル中でその有無を表現し、このベクトル空間の固有空間化の後に距離計算や clustering 処理によってテキスト間の近さを表現するのが典型としてある。検索法としては、keyword 検索や全文検索などがある。

さて、我々はテキストの組織化データベース作成のために、Galaxy Clustering[13] とよぶ 2 項関係データの非線形クラスタリング手法を用いている。これを使うとテキストの組織化データベースの作成と、画像とテキストの相互関係を埋め込むことに優れている。図 11 は、テキスト文を Cha -Sen という構文解析ソフト [14] で単語系列に分解し、その 2 項関係を新和度として、単語の空間配置を非線形クラスタリングを伴って行なわせたものである。ここで、出現単語数は約 12 万であり、これは 1995 年の毎日新聞の約 11 万 1 千記事から得られたものである。このとき、親和度の大きい単語は互に近くに配置される。また、一つの記事はこの空間で軌跡を描く。query の文または文章が入力されたとき、ChaSen 処理後に空間軌跡を描くが、その軌跡上および近傍軌跡上にある単語を含む記事群の中でその個数の最も多い記事が検索出力されることになる。

7 静止画像 - テキストの相互検索

query として与えられる静止画像でテキスト（文または文章）を検索することを考える [15]。いま用いるデータベースとして、平凡社の百科辞典を想定する。これには、項目が約 6 万（約 2 0 0 字／項目）あり、そのうち約 9 千項目に画像が含まれている。ある項目がテキストと画像の双方をもつとき、テキストと画像は関連していると想定することは自然である。従って、このような項目間において、

(1) テキスト間の距離が近ければ画像間の距離も近いことが望ましい

(2) 画像から特徴が取り出され、この特徴間の距離で画像間の本来の距離が定められる

という状況が生まれる。画像が空間に配置するとき、上記の (1) と (2) の条件が拮抗して全体として安定するところに配置されると自然であるとしよう。いま、百科辞典の各項目データを用いるとして、それらの項目で画像のあるなしにかかわら

図 11: Word space (over 120,000 words) after Galaxy Clustering

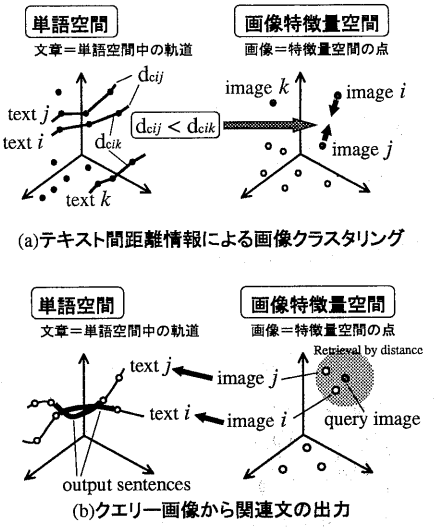

図 12: Text-Image organization and text retrieval by query image.

ずすべてのテキストから前述の Galaxy （単語）空間をつくるとしよう。平凡社の「マイペディア」という百科辞典を用いるとき、その出現単語数は約１１万となる。次に、画像の配置空間を上述の条件でもって作成する（図１２-(a)）としよう。未知の query 画像が与えられたとき、これを画像空間におき、近傍の画像を複数取り出す。この取り出された画像を含むテキストの trajectory を単語空間にもとめ、この trajectory が近接する部分（類似単語を共有する部分）を文として取り出し、これを出力とする (図１２-(b))。このようにして得られた文章の例を表１に示す [15]。

テキストを query 入力にして、その trajectory に近い単語空間内での trajectory をもつ百科辞典の項目のテキストで、かつ画像をもつものがあれば、この画像を出力とすることができる。

また、図１３に単語空間の Galaxy Clustering の演算による単語の分布の収束の様子を示し、図１４に画像空間における分布の収束の様子を示す。

ここで行なわれるような画像とテキストの相互検索がより広い対象について良好に働くためには、静止画とテキストの組からなるデータを大量に用いることが必要になってくる。

8 音声による静止画検索など

音声から静止画へのチャネルはテキストを媒介することでつけられる。３章に述べたように、音声によるテキストの検索はなされる。さらにテキストから静止画の検索は６章の考え方で行なう。このように２つのメディア間での相互検索には、３つめのメディアを介して行なうことで実現されるものもある。別の例としては、静止画によるテキストの検索を行ない、検索されたテキストの合成音声を新たな query として、音声から動画を検索することで、動画を最終的に検索出力とすることも考えられる。このように、音声と静止画と動画とテキストの相互を自由に行き来することの実現が最終的な相互検索システムである。

9 おわりに

ここでは、音声、静止画、動画、テキストという４つのメディアの任意の２つの間での相互検索システムがどのように実現できるかの方式を示した。この検索システムの実現には、多くの未解

入力画像	出力文
	おしべは6本，黄色の葯がある．東北地方の山地に多く，晩秋にブナなどの広葉樹に束生する．
	1912年に東京の白木屋が客寄せに考案した少女歌劇団が最初．〈ベルリンのバッハ〉，〈ハンブルクのバッハ〉と呼ばれる．

表 1: 未知画像とそれから検索される文章

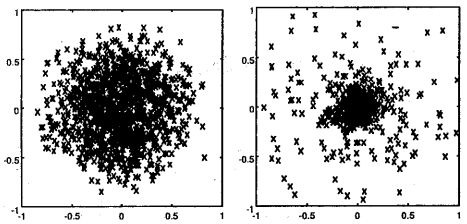

図 13: Clustering of word space : initial (left), final (right)

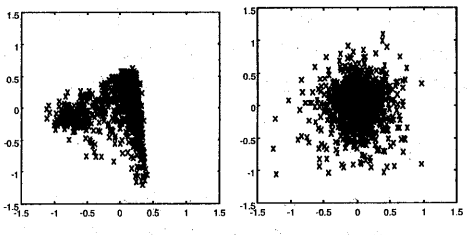

図 14: Clustering of image space : initial (left), final (right)

決の課題があるが、この技術開発は次世代の情報処理システムとして極めて有用であるといえよう。

謝辞

本稿の作成に協力戴いた、RWC の情報ベース機能つくば研究室／マルチモーダル機能つくば研究室の皆さんに深謝します。

参考文献

[1] 長屋茂喜、古川清、岡 隆一,"マルチモーダルデータベース検索に基づく音声とジェスチャのモーダル変換", 第3回知能情報メディアシンポジウム論文集, pp.173-179 (1997).

[2] 栗田 多喜夫、加藤俊一、福田 郁美、板倉 あゆみ," 印象語による絵画データベースの検索", 情処論、Vol.33, No.11, pp.1373-1383 (1992).

[3] 小野 敦史、天野 督士、斗谷 充宏、佐藤隆、坂内 正夫," 状態遷移モデルとシーン記述言語による自動キーワード付与機能をもつ画像データベースとその評価", 信学論、**J79-D-II**, No.4, pp.476-483 (1996).

[4] 長坂 晃朗、宮武 孝文、上田 博唯," カットの時系列コーディングに基づく映像シーンの実時間識別法", 信学論、**J79-D-II**, No.4, pp.531-537 (1996).

[5] Mark T.Maybury, Editor, "Intelligent Multimedia Information Retrieval", The MIT Press, 1997.

[6] 木山 次郎、伊藤慶明、岡 隆一," Incremental Reference Interval-free 連続 DP による任意話題音声の要約と話題境界検出", 信学論、**J79-D-II**, No.9, pp.1464-1473 (1996).

[7] 岡 隆一," 連続 DP を用いた部分整合法によるフレーム特徴の音素片認識", 信学論 (D)、**J70-D**, No.5, p.917 (1987-05).

[8] 中沢正幸, 古川清, 遠藤隆, 豊浦潤, 岡隆一, "音声波形からの音素片記号系列を用いた音声要約と話題要約の検討", 信学技報, **SP97-15-21**, pp.119-124 (1997).

[9] 遠藤 隆、中沢 正幸、長屋茂喜、高橋 裕信、岡 隆一,"音声と動画像の自己組織化ネットワークによるデータ表現とスポッティング相互検索", 人工知能学会、合同研究会 "AI シンポジウム '97, **SIG**-J-9702-3, pp.15-20 (1997-12).

[10] 豊浦 潤、岡 隆一,"テキストの知識ベース化のための自己組織化ネットワークの提案", 信学技報, **NLC**-96-59, pp.23-30 (1997).

[11] Ryuichi Oka : "Spotting Method Approach Towards Information Integration", Proceedings of 1997 Real World Computing Symposium, pp.175-182, 1997.

[12] Ryuichi Oka , Yoshiaki Itoh, Jiro Kiyama, Jian Xin Zhang : "Concept Spotting bye Image Automaton", Proc. Spring Meet. of Acoust., Soc. Japan, **3-4-12**, 1995 (in Japanese).

[13] 高橋 裕信、新田 義貴、岡 隆一,"非線形クラスタリングによるパターンの分類", 信学技報、**PRMU**-98-13 (1998).

[14] ChaSen, "http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html."

[15] 森 靖英、高橋 裕信、新田 義貴、岡 隆一,"画像とテキストの自己組織化データに基づく画像理解方式の提案", 信学技報、**PRMU**-98-74 (1998-09).