平成 16 年度 公立はこだて未来大学卒業論文

コンピュータはイメージできるか

寺田　健

情報アーキテクチャ学科　　1201024

指導教員　　迎山和司

提出日 2005 年 1 月 31 日

Can you imagine something, Mr.computer?

by

Takeru TERADA

BA Thesis at Future University-Hakodate, 2005

Advisor:　Kazushi MUKAIYAMA

Department of Media Architecture
Future University - Hakodate

January 2005

Abstract– This research aims to find a way to make video from words. I paid attention to the function of mental images that we humans can imagine.

We humans can imagine a scene when we are reading novels. For example, we can imagine the smell for "offensive smell", we can imagine the image for "a touching scene". Now we call the visual image , that we imagine, "mental visual image". "Mental visual image" has been studied in the area like cognitive psychology. The studies for mental visual image progressed by technologies like fMRT (functional magnetic resonance imaging), PET (positron emission tomography). We can see the activity of nerves in brain using these technologies.

I paid attention to mental visual image and set a goal to make computers do the same way as we imagine from the words. In the model made in this research, words are given by Chinese characters (KANJI) drawn by user. And mental visual image is given by the video in Hard Disk. When user draw some characters on the screen, the computer searches for the video that has the meaning of the characters. For example, when you draw "牛" (means cow) and "見" (means see), computer searches for the video that cow is seeing and display it on the screen.

The problem of this model is that the relation between words and videos are given by me, the writer. But we learn the relation between words and mental visual image from our experiences and saved in the place, called long term memory (LTM). In 1986, Paivio considered the idea, called dual coding theory, the memory is coded in dual way, words and images. Thus, there is another problem on the model made in this research. The problem is that visual mental image is defined directly from videos in Hard Disk. When the object in the world is memorized, the visual information of the object is coded into features like colors and shapes.

The process that we imagine visual mental image is the process follows next. (1) understand the word, (2) imagine the features that means the word. Thus, when we imagine mental visual image, we need to do learn words and code the visual information we sense.

There are a problems above in the model made in this research, but it is meaningful to pay attention to the mental visual image that we imagine.

Keywords: Video composing, Mental image

概 要: 本研究の目的は、コンピュータを用いて言語から映像を作る方法についての提案を行うことである。その上で、筆者は人間の脳のイメージ機能に着目した。我々人間は小説などを読むとき、小説に書かれた文章があらわすシーンを想像しながら読むことができる。例えば「鼻を刺す臭い」と書かれていれば、我々は「鼻を刺す臭い」の嗅覚的なイメージを想像することができるし、「瞼に焼きつくような光景」と書かれていれば、「瞼に焼きつくような光景」の視覚的なイメージを想像することができる。このように、我々が想像するイメージの中でも視覚的なものを視覚イメージと呼ぶ。本研究において、筆者はこの視覚イメージに注目し、人間が言語入力によって視覚イメージを想像するのと同様の過程をコンピュータにさせることを目標とした。

視覚イメージについては主に認知心理学や知覚心理学と呼ばれるような分野で研究がなされている。近年の fMRT (functional magnetic resonance imaging) や PET (positron emission tomography) と呼ばれる脳の神経活動を画像化する技技術の発展により、イメージ研究についても新たな展望が開かれた。これらの技術を用いて、Kosslyn らは我々が視覚イメージを想像するときには、実際に実世界の対象を見ているときと同様の脳の部野が活動することを確認した。[1] この結果にともなって Kosslyn は、イメージは一種のワーキングメモリであるという考えを述べている。ワーキングメモリとは、感覚入力と長期記憶を結びつける役割を担うものである。ワーキングメモリがどのような機構をもつかといったモデルについては、Baddeley らによって提案がなされている。[9]

本研究において著者は、人間が与えられた言語から視覚イメージを想像する過程のモデルを作成した。言語は漢字を手書き入力によって与えるものとし、人間の視覚イメージに対応するものを、コンピュータのハードディスクに溜めこまれた映像とした。言語入力において手書き文字を用いた理由は、スクリーン上にそれぞれの文字の持つ概念の配置を行うためである。作成したモデルにおいては、いくつかの文字をスクリーン上に描くと、それらの文字の組み合わせにより、文字のもつ概念を含むような映像を検索し、出力する。最もシンプルな組み合わせは主語と述語の二文字である。例えば「牛」,「見」という二つの文字をスクリーン上に描くと、牛がこちらを見ているような映像を検索し、出力する。また複数の主語が描かれたあと、そのうちの一つの近くに対応する述語が描かれる、といった少し複雑な組み合わせでは、その述語が描かれた場所に一番近い主語と、その述語の組み合わせが「状況」として定義され、他の主語はその状況のもとで映像検索を行う。例えば「虎」,「馬」と描いた後で「虎」の近くに「歩」と描くと、虎が歩いているような映像と、それにともなって馬が逃げ出すような映像が出力される。

今回作成したモデルと人間がイメージを想像する過程を比べると、いくつかの相違点が存在する。大きな相違点として考えられるのは、人間は言語とイメージの関係を成長にともなう経験によって獲得していくのに対し、今回作成したモデルでは言語とイメージの関係を、著者が最初から与えてしまったことである。人間においては、言語とそれに対応するイメージは経験によって獲得され、長期記憶と呼ばれるものの中に蓄えられていく。もう一つの問題点は、イメージに対応するものとして、ハードディスクに溜め込まれた映像をそのまま用いたという点である。我々が実世界において見た対象をイメージ化するときには、対象の視覚的情報の符号化が行われる。(なお、自閉症と呼ばれる症状を持つ人間においては符号化が行われない。) 以上のように、言語入力から視覚イメージを想像する過程は、入力された言語を理解し、言語に対応する対象の、符号化された視覚的特長を想像するという過程であると考えられる。よって言語入力から視覚イメージを想像するためには、言語の習得と、対象の視覚的情報の符号化が必要になると考えられる。

今回作成したモデルと人間が言語入力から視覚イメージを想像する過程と照らし合わせると以上のような問題点が生じたが、言語から映像を作るという目的のもとで、人間が視覚イメージを想像するに注目したことは、映像構成の手法について考える上でも有意義であったと考えている。

キーワード: 映像論, 視覚イメージ

目 次

第1章　序論

本研究の目的は、言語から映像を構成する手法についての提案を行うことである。その上で、筆者は人間の脳のイメージ機能に着目した。

以前にも、言語に注目して映像を作ろうとする試みはあった。例えば、ソビエトのモンタージュ派と呼ばれる映画監督であるプドフキン、エイゼンシュタインらは、個々のショットは詩人が詩を作り出すための単語のようなものであると考えた。[15] 彼らは、ショットの中身や長さ、リズム等を、別のショットの中身や長さ、リズムと衝突させることでメッセージを生み出そうと考えた。しかし彼らは、言語の持つ文法構造のような特性を映像に結びつけようとしたのであり、我々が普段使用している言語そのものを映像と結びつけようと考えたわけではなかった。それに対し著者は、我々が使うような言語を直接映像と結び付けようと考えた。そこで、著者は人間の脳のイメージ機能に注目したのである。

イメージの本質に関しては現在でも研究が続けられ、明確なことは言えない。[1] しかし、我々が言語入力からイメージを想像することができるのは言うまでもない。例えば、我々人間は小説などを読むとき、小説に書かれた文章があらわすシーンを想像しながら読むことができる。

イメージはいずれも、感覚世界の情景と結びついている。例えば、文学作品等に見られる「瞼に焼きつくような光景」、「耳に残る響き」、「鼻を刺す臭い」といった表現はいずれも感覚世界にあるような情景と結びついている。「瞼に焼きつくような光景」に関しては、「瞼に焼きつくような光景」の視覚的な映像のようなもの（視覚イメージ）を想像することができるであろう。また Paivio は、我々が日常の場面を知覚し、記憶するときには、イメージ化され、言語化されるという二重符号化説を唱えた。[12] 以上のように、言語とイメージは深い結びつきを持っている。

本研究では、言語から映像を構成するという目的のもとで、イメージの中でも視覚的なもの、つまり視覚イメージに注目し、人間が言語入力から視覚イメージを想像するのと同様の過程をコンピュータにさせることを目標とする。また、人間が言語から視覚イメージを想像する過程のモデルの作成を通して、人間は何のためにイメージを想像するのか、イメージと言語の関係は何に依存するのかについても考察する。

第2章　関連事項

2.1　イメージとは何か

一言でイメージと言っても、視覚的イメージ、聴覚的イメージ、嗅覚的イメージ等のさまざまな感覚のイメージが存在する。本研究では映像を構成するために視覚イメージのみに注目したが、この章においてはイメージ全般についての研究を紹介する。なお、本研究で扱う視覚イメージとは、現実世界における対象の、色、形体を含む映像であるとした。その理由として、近年の MRI や PET と呼ばれる脳神経活動を画像化する技術を利用した研究により、視覚イメージが映像のようなものであることが確認されたからである。

2.1.1 では、まずイメージ研究について関連の深い心的表象系と呼ばれる概念について紹介する。2.1.2 以降では、主に知覚心理学分野におけるイメージに関する研究を紹介する。

2.1.1　イメージと心的表象系

イメージは歴史的に、心的表象系と呼ばれる心の中の記号体系に含まれる記号であると考えられてきた。心的表象系に属する記号は、現実世界における対象の、知覚可能な特徴を表している。例えば「車」に関して知覚可能な特徴は、「車」の形、色、存在する場所等である。(図 2.1)

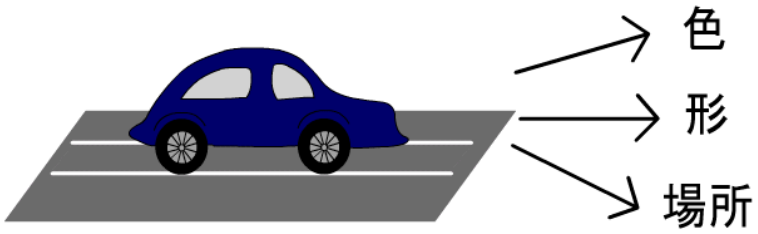

図 2.1:「車」の知覚可能な特徴

留意が必要なのは、イメージなどの心的表象系に属する記号は、現実世界における対象の特徴であり、対象そのものではないという点である。

2.1.2　知覚心理学分野におけるイメージの研究成果

近年、MRI(magnetic resonance imaging) や PET(positron emission tomography) 等の脳神経活動を画像化する技術の発達により、我々がイメージを想像するときには、実際に

感覚情報が入力されたときと多くの神経機構を共有していることが明らかにされてきた。[1]

視覚イメージに関しても、ある段階からは、あたかも実際の視覚情報が脳に入力されたときと同じように処理されていくことが明らかにされた。視覚情報の受容では、大きく二つの経路があると考えられている。(図 2.2)

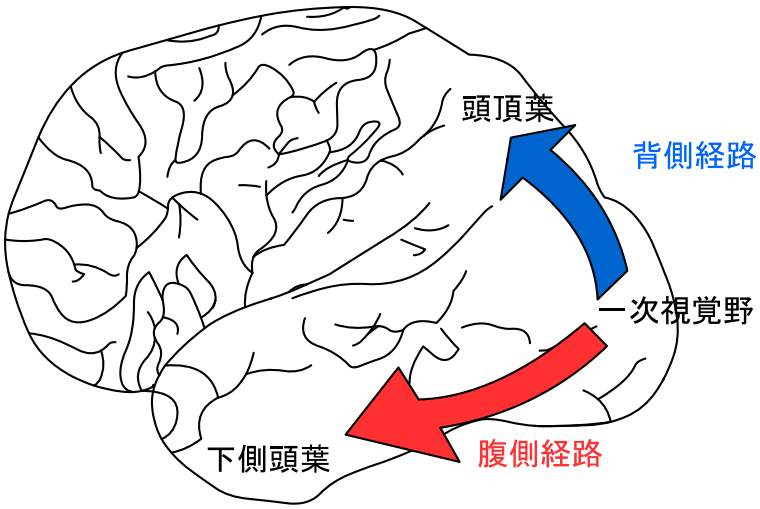

図 2.2: 視覚情報が一次視覚野に入力されたあとの情報の流れ

一つは一次視覚野から下側頭葉に到る腹側経路であり、対象物を同定するための情報を処理していると考えられている部分である。もう一つは、後頭葉の初期視覚野から頭頂葉に到る背側経路であり、位置や動きの情報を処理する部分であると考えられている。言い換えれば、腹側経路が"WHAT"に関する情報を処理するのに対し、背側経路は"WHERE", "HOW"といった情報を処理している。

Mellet らは聴覚によって与えられる指示に従った方向へと、頭の中で積み木を組み立てていく課題を用いて、対象物を同定するための背側経路が活動することを示した。[2] また、Kosslyn は PET を用いた実験により、人間が視覚イメージを想像するときには、脳の視覚情報の入り口である一次視覚野に活動が認められることを示した。[3] しかし、一次視覚野が視覚イメージの想像に際して活動するかどうかについては研究者によって異なった見解があり、決着がついていない。

以上のように、視覚イメージを想像する際に実際に視覚情報を処理する部野が働くという事実は、視覚イメージが絵や映像のようなものであることを表している。

Kosslyn は、視覚イメージは一時的に視覚情報を貯蔵するところであり、長期記憶からの情報を呼び出しながら現実の知覚との間を調整する(現実の知覚対象を予測する)機能も持っているのではないかと考えている。例えば、人間は 2 次元の網膜に投射された光学象から、3 次元構造をもつ実世界の様子を推測し、把握しなければならない。この作業は与えられた条件(網膜に投射された光学象)のみから一義的に説くことのできない、いわゆる不良設定問題である。この問題を効率よく解決するために、以前の経験に基づいた知識が何らかの役割を果たしているのではないかと考えられている。[1]

2.1.3　イメージの操作

この節では、脳の中に蓄えられたイメージを操作することによって現実世界の因果関係を予測するような行動、いわゆるメンタルオペレーション [1] について述べる。

本田らは脳内に形成されたイメージの操作を感覚運動制御に関連付けて考察するにあたり、そろばんの熟練者による暗算に注目した。[1] 感覚運動制御とは、我々がそろばんを、実際に手にとって指ではじくといった、現実世界における実際の運動のことを指す。

本田らは、そろばんの熟練者が暗算を行うときには、視覚運動制御と共通の神経機構がはたらくのではないかと考え、脳神経活動を画像化できる fMRI(functional magnetic resonance imaging) を用いて、そろばんの熟練者が暗算を行うときの脳神経活動を検討した。その結果、3 桁の見取り暗算を行う場合には、両側の背外側運動前野、後部頭頂皮質にほぼ対象性に強い信号増強が観察された。両側の背外側運動前野は、後部頭頂皮質と連動することにより、視覚運動制御に重要な役割を果たしていることが知られている。この結果は、現実世界の視覚運動制御に用いられていた運動前野の何らかの機能が訓練によって脳内世界のイメージ機能へと移行後、引き続き利用されていることを示唆しているものと考えられる。

そろばんの熟練者の例を含め、本田らはイメージ操作と運動制御の関係について、次のような仮説を立てた。進化的に初期の段階には、外界から入力された情報を一定のルールで処理し、外界の物体を操作する運動として発現するはたらきを担っていた神経機構が、進化に伴う脳システムの複雑化と記憶の発達によって、記憶から情報を引き出したり、逆に出力を記憶に書き込んだりすることが可能になった。その結果、外界との情報のやり取りを行わずに、脳の内部で一定のルールに基づいて情報を処理することが可能になったのがイメージの操作ではないかという仮説である。[1] つまり、人類の進化過程の中で、感覚運動制御からイメージの操作へと展開してきたプロセスが、個人の中で訓練により急速に誘導されたのがそろばんの熟練者の例だということである。

2.2　イメージと記憶の関係

2.1.2 項では、イメージは一時的に感覚情報を貯蔵するところであり、長期記憶から呼び出した情報を用いて、現実の知覚との間を調整する機能も持っているのではないか、という Kosslyn の考えを述べた。これは、イメージは一種のワーキングメモリであるという考え方である。ワーキングメモリとは、記憶システムの一種である。2.2 節では、イメージ・言語と記憶の関係について、主にワーキングメモリについての研究を紹介する。

2.2.1　ワーキングメモリ

ワーキングメモリとは、言語処理のような認知過程の中で、一時的に必要となるような記憶の働き、あるいは、これを実現している機構やシステムのことを言う。[4]

例えば小説を読むとき、我々は言語認知を必要とする。我々が言語を認知するとき、ほとんどの言語情報は時系列的に与えられる。例えば以下のような文章を読んでみよう。

「窓の外にはどんよりと曇った空が見える。冬の天気は毎日曇り空。布団の中

から出した顔を、刺すような寒さが襲う。机の上で目覚まし時計が鳴りだしたが、私は布団の中に顔を引っ込めた。」

以上のような文章を読んでいるとき、つまり、我々が時系列的に言語情報を受容しているとき、我々が文章の全体的な構造をつかむためには、現在読んでいる箇所以前に与えられた情報を覚えておく必要がある。この、以前与えられた情報を蓄えておくための機構がワーキングメモリである。

ワーキングメモリの機能がどのようにして実現されているのかについては、まだ明確ではないが、複数のタイプがあるということが一致した見解として認められるようになってきている。なかでも広く認められているワーキングメモリは、視覚的情報を保持する視覚的ワーキングメモリと、言語・音韻的情報を保持する言語的ワーキングメモリである。[4] ワーキングメモリの形成については、長期記憶から意図的に想起された情報がアクティブに保存されるとワーキングメモリになるという考えがある。[8] 長期記憶については次節で紹介する。

2.2.2 長期記憶

長期記憶とは長期間保持される記憶で、意味記憶、手続記憶、エピソード記憶の３つに分類することができる。意味記憶は言葉の意味や概念の記憶であり、「鯨は哺乳類である」といった一般的な知識としての記憶を指す。手続記憶は言葉を解さず、体で覚えているような記憶であり、たとえば「車の運転の仕方」といった手続きに関する記憶である。エピソード記憶は、特定の時間的・空間的文脈の中に位置付けることの出来る出来事、言い換えれば個人的体験の記憶である。

2.2.3 ワーキングメモリのモデル

ここでは、Baddeley のワーキングメモリのモデルについて述べる。Baddeley はワーキングメモリについて、次のようなモデルを提案している。（図 2.3）[9]

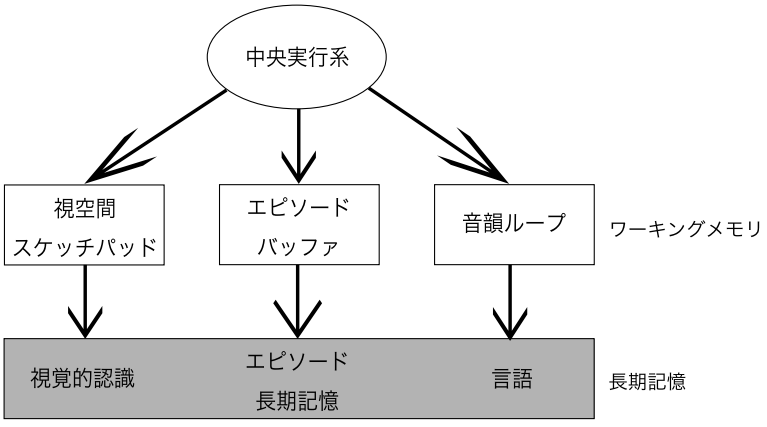

図 2.3: Baddeley(2000) のワーキングメモリ・モデル

Baddeley はワーキングメモリを一つの中央実行系と三つのサブシステムに分けて考えている。三つのサブシステムはそれぞれ、視空間的スケッチパッドとエピソード・バッファ、音韻ループと呼ばれる。中央実行系は、長期記憶の表象を活性化する役割を持つ。視空間的スケッチパッドは空間の理解に関わるサブシステム、音韻ループは言語の理解に関わるサブシステムとして想定されている。視空間的スケッチパッドは、言ってみれば心の中のキャンバスであり、今回の研究で問題となる視覚的イメージは、視空間的スケッチパッドに描かれるものであると考えられる。

しかし視空間的スケッチパッドや音韻ループは、それぞれ異なった表象を持つ。よって、これらの情報を統合し、意味を形成させるための機構が必要である。そこで複数の情報源から得られる、異なる様相で保持されている表象を統合する機構として導入されたのがエピソード・バッファである。エピソード・バッファとは、視覚的スケッチパッドや音韻ループ、長期記憶からの情報など、複数の情報源からの表象を保持することができるという特徴を持つ。

しかし、Baddeley のモデルにも問題は残されている。たとえば、エピソード・バッファと視空間的スケッチパッド、音韻ループの境界はどの程度明確なのか、また、エピソード・バッファとエピソード長期記憶は完全に分離しているのかなどである。[4]

2.3 イメージと言語の関係

2.1 節, 2.2 節を踏まえた上で、本研究において問題となる、言語とイメージの関係についての研究例を紹介する。

2.3.1 Pavio の二重符号化説

我々が日常の場面が知覚され、記憶されるときには、心的表象系に記号化され、言語化されるという Paivio の二重符号化説がある。[12] この説によると、具体的な言葉はイメージとともに言語的にも記憶されるが、抽象的な言葉は言語的にのみ記憶される。

近年、この説を裏付けるかのように、Mazoyer らは、我々が言語から視覚イメージを想像するときには脳のなかの言語情報を扱う部野と視覚的情報を扱う部野が活動することを実験によって確かめた。[14] 言語情報を扱う部野が活動する際に、視覚的情報を扱う部野が活動するということは、言語と視覚イメージが深い関連をもっていることを示している。

2.3.2 言語からイメージを想像する過程でのワーキングメモリの役割

我々が文章を読むとき、状況モデル（situation model）を構成することが知られている。状況モデルとは、読者の推測や知識によって予測された状況の表象である。例えば、「どんよりと曇った空」について思い浮かべる状況は人によって異なる。

Friedman と Miyake は、状況モデルが多次元的特長を備え、いくつかの異なった性質の情報を表現していることに注目した。[10] 例えば、どこの部屋に誰がいるといったといった空間的状況や、誰々がその部屋を出た（結果、その部屋には誰もいなくなった）といった因果的状況である。Friedman と Miyake は、実験によって状況モデルの構築がワーキン

グメモリの働きに大きく依存していることを確認した。特に、空間的情報には空間的ワーキングメモリ（Baddeley のモデルでいう視空間スケッチパッド）が、因果的情報には言語的ワーキングメモリ（Baddeley のモデルでいう音韻ループ）が大きく関与していることがわかった。

2.4 言語の習得

言語の習得に関しては、今回の研究の中では問題としていなかった。なぜならば、今回目標とした部分は、既に言語とイメージの関係が構築されたことを前提として、我々が言語入力から視覚イメージを想像する部分をモデルにしようと考えたからである。しかし言語と視覚イメージの関係は、人間は経験によって構築していく。この点については 4 章の考察においてさらに深く述べるが、作成したシステムの考察にあたり、言語の習得について考慮する必要が生じた。よってこの 2.3 節において関連事項として、人間の言語習得について主に認知心理学分野における研究を紹介する。

言語の習得について説明するにあたり、2.3.1 でまず言語に対して、いくつかの見方を紹介する。2.3.2 では言語習得に必要となる要素、2.3.3 では人間の言語習得過程について説明する。2.3.4 では、コンピュータに言語獲得させるための研究例について紹介する。

2.4.1 言語に対する見方

言語に対する視点としては主に、記号論的視点、精神分析的視点、言語使用論的視点の 3 つが存在する。[11]

記号論的視点とは、言語をひとつの記号システムとしてとらえる見方である。ある記号は、他の記号との関係の中で意味を持つ。例えば漢字の「馬」は、漢字という体系の中である意味を担っている。

精神分析的視点とは、言語の意味を、話す主体の何らかの心的過程、意図によって規定しようという見方である。これは、例えば、我々が「食べ物ある？」と聞くことには、「食べ物をくれ」という意図が含まれていることがあるように、言語は話者の状況によって意味が変わるからである。

言語使用論的視点とは、言語を行為としてとらえる見方である。言語がコミュニケーションの場において、どのように使用されているかをみることによって、言語をとらえる。

2.4.2 言語獲得の要因

言語の獲得には、遺伝（生物学的な賦与）と環境（我々が経験する世界）が大きな役割を果たしている。 遺伝と環境と、どちらが重要かにより理論は異なってくる。

Chomsky は、生成文法という立場から、子供が言語を学習する方法は、生得的であると述べている。[4] 生成文法とは、人間はあらゆる言語の基礎となる普遍文法を持って生まれてくるとし、子供たちは発達によって、それぞれの文化に個別な文法を得て、定常状態に落ち着くという考えである。Chomsky は人間が複雑な文法を獲得するには、外部からの手本の入力だけでは十分でないと主張している。記号論的視点から見た言語は、この

ような Chomsky の説にもとづいて獲得されると考えられる。

他方、Piaget や Bruner は子供の言語獲得には、経験が大きな役割を果たしていると述べている。Bruner は言語習得に関して、社会的な相互作用、つまり外部とのコミュニケーションが重要であると考えている。[4] 精神分析視点や言語使用論的視点から見た言語の獲得はこのように、経験を重視する立場から説明される。

Karmiloff と Karmiloff-Smith は、遺伝と環境の二分は有用ではなく、両者のインタラクションに焦点を当てるべきだと述べている。[5] 彼らは、進化が果たした役割に注目し、次の二点が重要であると述べている。まず、進化は人間で生後の脳の発達期間を非常に長くし、環境からの入力が、発達する脳の構造を形成できるようにした。第二に進化は、たくさんの学習メカニズムを与えてきた。この学習メカニズムは、さまざまな環境入力とのインタラクションによって、生かされる。言語は生得的な能力ではなく、認知や社会性なおどのほかの領域相互に作用しあい、獲得されるのである。

2.4.3 人間の言語獲得過程

Tomasello は、語彙獲得の認知的基盤として、(1) 他者が何について話しているのか、その指示対象を認知しカテゴリーをつくることを可能とする子供の能力、(2) 他者が言語のさまざまな部分を使用している際に、その他者の意図が何かを理解できる子供の能力を挙げている。[6] これらの能力を、われわれは乳児期から幼児期にかけて獲得する。生後 8-12 か月で、乳児は関心を他者に向ける。この時期の乳児には、渡す・見せる（他者の注意をひきつける行為）、追随凝視（他者が他のものを見ているときに他者の凝視を追う）、社会的参照（新奇な事物への大人の情動反応をモニターする）といった行為が見られる。[5] 初期の語彙発達段階（生後 15 か月）においては、乳児は場面にあった音声を一貫した形式で、システマティックに繰り返し発する。これは、一語発話と呼ばれる。語彙が 100 語程度に達すると、乳児は二語発話をするようになる。例えば「ママ 来る」といったようなものである。

2.4.4 コンピュータに言語を獲得させるためのモデル「Rhea」

錦見らはコンピュータの言語獲得についてのモデル、Rhea(neaR Human languagE Acquisition) を提案した。[7] Rhea はことばの意味の獲得を重視したものであり、そのための道具として文法の獲得を同時に行う。ここで言うことばの意味とは、環境を認識するための注意のプロセスという一定の手続きを指す。

Rhea は言語的入力と非言語的入力を受け取り、それぞれを統語構造（ことばとことばの関係）とフィルタ（そのことばに対応付けられている意味）という内部的な記述に変換し、それらにもとづいた入力のクラスわけを行う。この過程はことばの構造と意味、構造を記するための統語の規則、意味を文節するための概念、そして意味と構造にもとづいたことばのカテゴリーの獲得とみなすことができる。

第3章　作成したモデル

本研究において目標とするのは、言語入力から、視覚イメージをコンピュータに想像させることである。これは、我々人間においては、例えば文章を読んだときに、それぞれの単語（馬、草原、走る）の組み合わせから、実際に馬が草原を走っているような視覚イメージを想像することに対応する。本研究では、言語入力から視覚イメージを想像する過程をモデル化した。モデル化とはある現象についての過程や推測の具体的な形を表現するための手段の一つである。[7] 3 章では、作成したモデルについて紹介する。

3.1　作成したモデルの概要

言語を与えて、与えられた言語のもつ概念を含む映像を出力するプログラムを作成した。言語は漢字を手書き入力によって与えるものとし、人間の視覚イメージに対応するものを、コンピュータのハードディスクに溜めこまれた映像とした。言語入力において手書き文字を用いた理由は、スクリーン上にそれぞれの漢字の持つ概念の配置を行うためである。作成したモデルにおいては、いくつかの漢字をスクリーン上に描くと、それらの漢字の組み合わせにより、文字のもつ概念を含むような映像を検索し、出力する。

提案するモデルにおいては、主語を表す漢字と、述語を表す漢字の組み合わせが、映像の言語表現の基本単位となる。この基本単位は、言語の文法表現で用いられる記号（S：主語、V：述語）を用いると、「S・V」という、構造によって表わされる。例えば、主語に「牛」、述語に「見」を定義したときには、牛がこちらを見ている映像が呼び出される。(図 3.1）

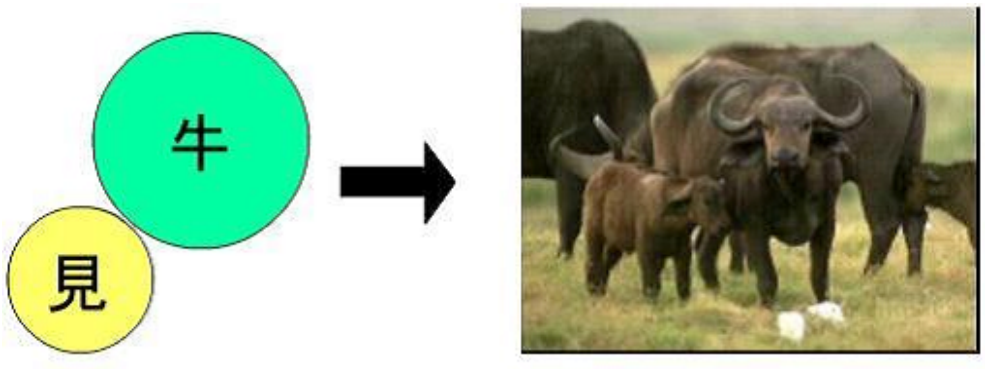

図 3.1: 「牛」と「見」という漢字を与えて映像を呼び出す。

作成したプログラムで上記のような「S・V」構造が定義されたときの動作を紹介する。

図 3.2 が起動画面である。図 3.2 では左側に配置されている白いキャンバスに「馬」・「走」という漢字を入力した。（図 3.3） その結果、馬が走っている映像が出力された。（図 3.4） 映像が出力される位置は、主語が描かれた位置であり、その映像の大きさは描かれた主語の大きさと同値である。

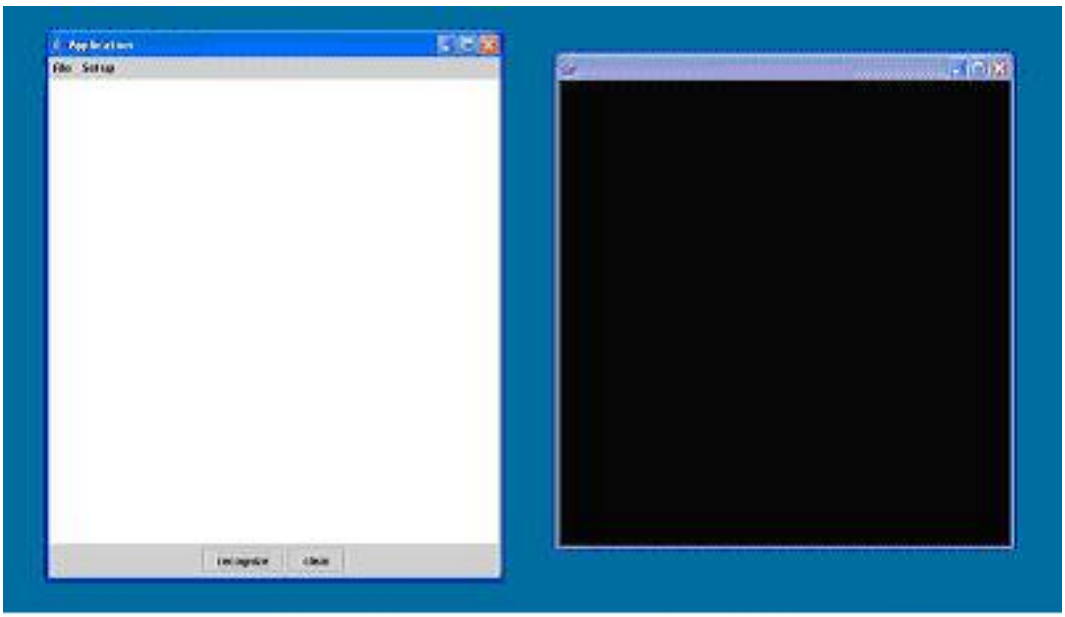

図 3.2: 起動画面

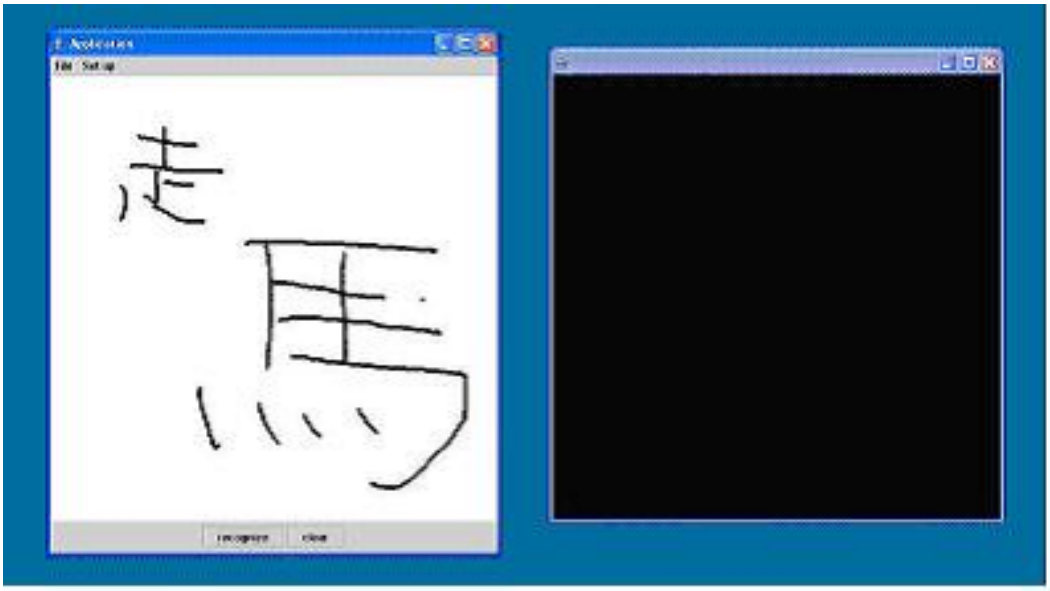

図 3.3: 文字を描いたところ

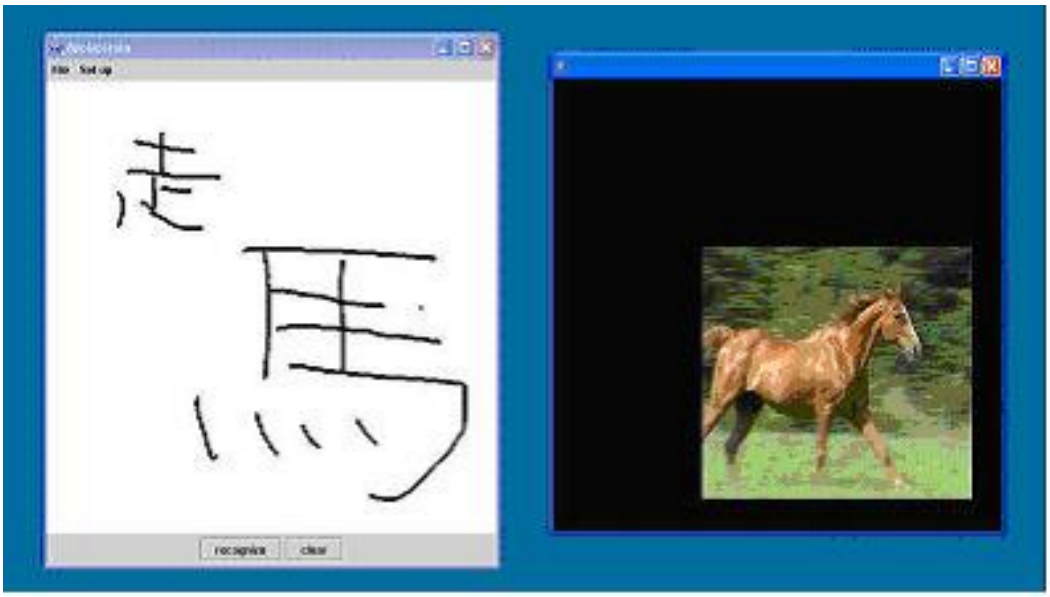

図 3.4: 描いた文字から視覚イメージを想起したところ

基本単位では、「$S \cdot V$」という単純な構造しか表せない。そこで拡張として、「状況 X では $S \cdot V$」という構造を導入した。状況 X は、基本単位「S・V」によって定義される。複数の主語が描かれたあと、ある述語が描かれると、その述語が描かれた場所に一番近い主語と、その述語の組み合わせが「状況」として定義され、他の主語はその状況のもとで映像検索を行う。例えば、起動画面において、「虎」と「馬」という漢字が与えられたとする。次に、「虎」の近くに「歩」という漢字が描かれ、「虎が歩く」という「$S \cdot V$」構造が定義されたとする。(図 3.5) これは、「虎が歩く」という状況が定義されたことと同値である。

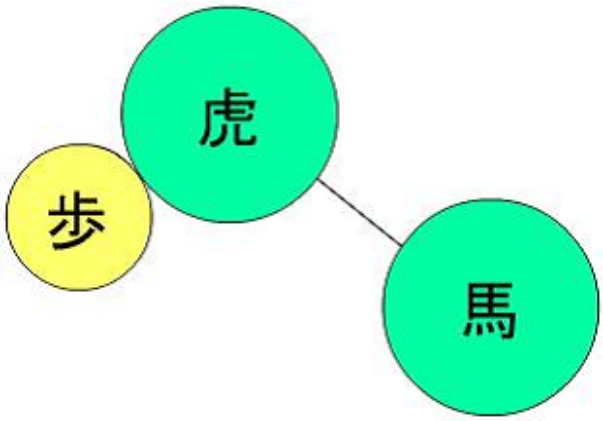

図 3.5: 「虎」と「馬」を描いた後、 「虎」に述語「歩」を定義

「虎が歩く」という状況が定義されると、「馬」には「逃」という述語が自動的に定義

される。(図 3.6)

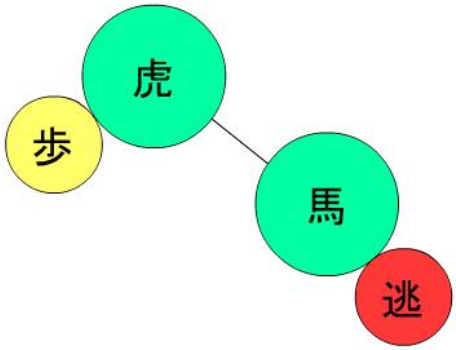

図 3.6: 状況：「馬」に「逃」という述語が定義される

すると、「虎が歩く」映像と「馬が逃げる」映像が呼び出される。(図 3.7)

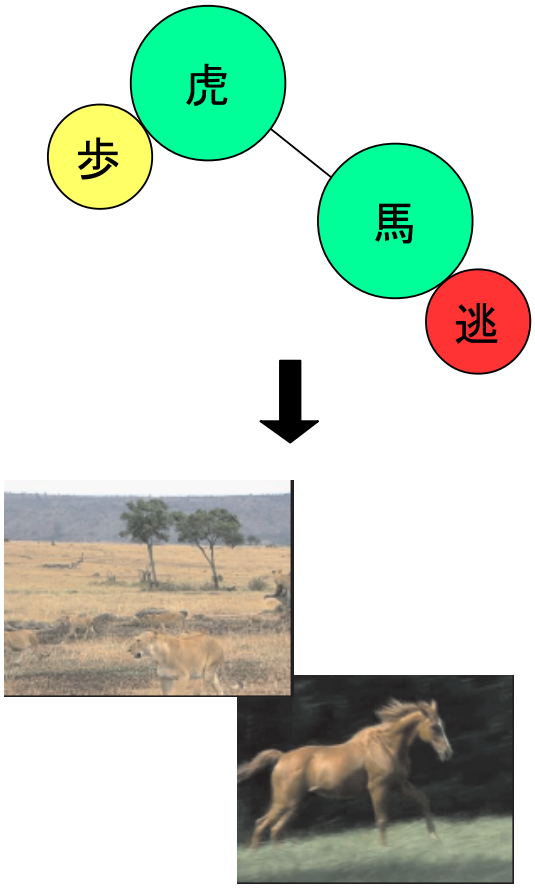

図 3.7: 「虎が歩く」映像と「馬が逃げる」映像

今回作成したモデルにおいて「虎・馬・歩」と順に描いたあとの、実際の動作の様子を図 3.8 と図 3.9 に表す。

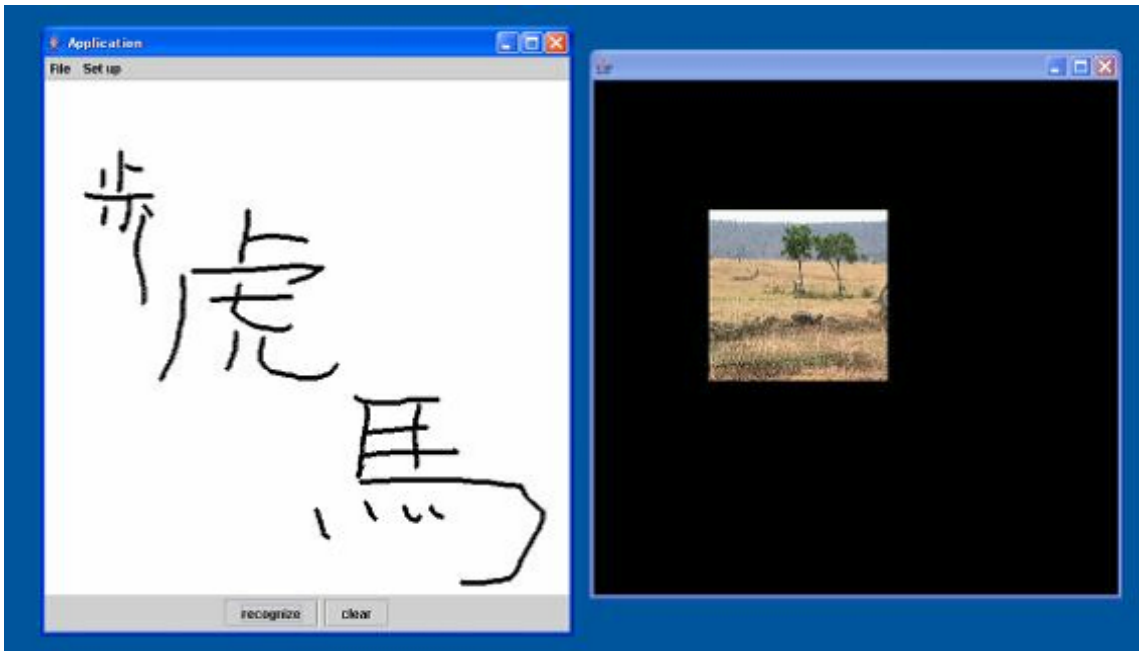

図 3.8: 虎が歩き出す

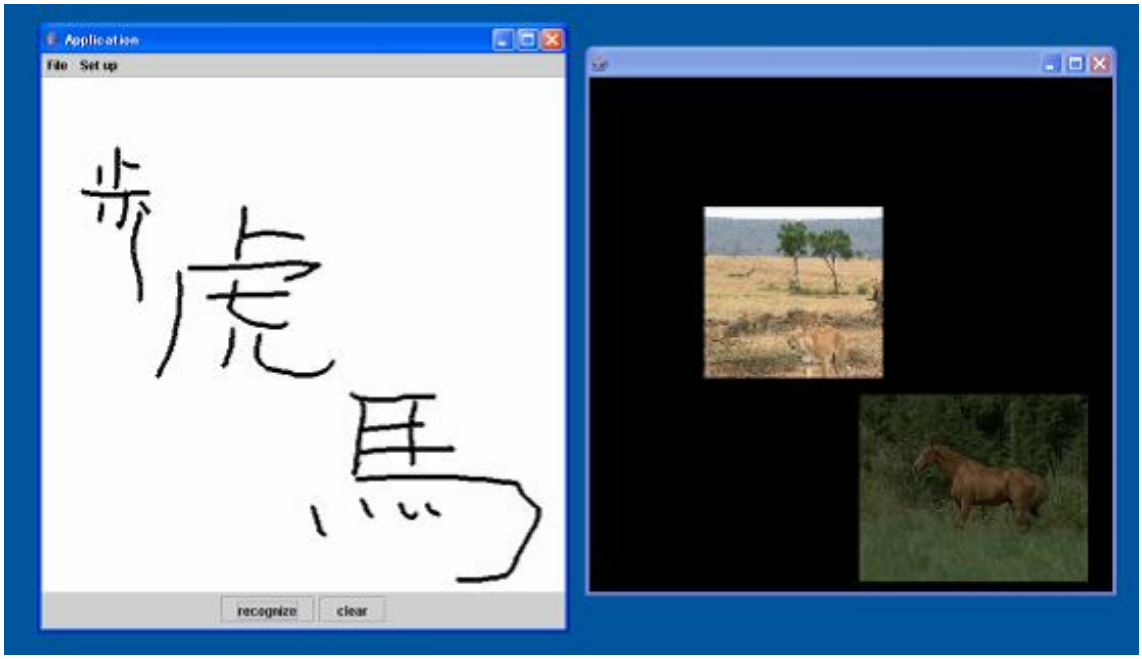

図 3.9: 馬が逃げ出す

3.2 モデルの実装方法

今回作成したモデルは Java 言語によって実装された。モデルの実装において必要となる技術は、文字の認識と映像の検索方法である。なぜ文字の認識が必要になるかというと、映像をどこに、どのような大きさで配置するか決定するために、利用者が文字を描いた位置とその大きさの情報を利用したためである。

3.2.1 文字の認識

文字の認識には、ストローク方向列の DP マッチングを用いた。ストローク方向とは、書き始めの点から見た、書き終わりの点の方向を表す。今回はストローク方向を 8 方向に定めた。（図 3.10）

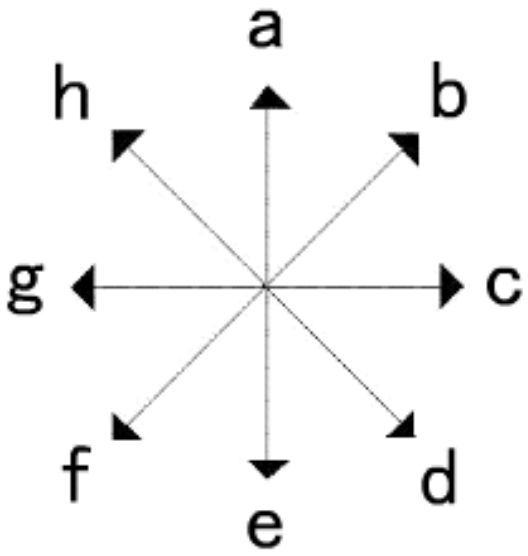

図 3.10: ストローク方向

例えば「犬」という文字はその書き順から、「一、／、＼、丶」と時系列的に分解できる。これを図 3.8 のストローク方向であらわすと、[c, f, d, d] となる。こういった時系列ストローク方向パターンを認識のためのデータとし、前段階としてそれぞれの文字についてストローク方向パターンを定義しておく。文字が描かれたら、そのストローク方向パターンから最も類似したパターンを探し出すことによって、文字の認識を行っている。ストローク方向を認識のためのデータとした理由は、今回のモデルに用いた文字は漢字であり、漢字学習者においては、ある程度書き順が一定して得られると考えられたからである。

DP マッチングとは、2 つのパターン間の距離を求めるアルゴリズムである。距離とは、2 つのパターンの違いを表す概念で、スカラー値で表される。例えばパターン Q と S が一次元符号列

$$Q = q_0, q_1, q_2, q_3$$

$$S = s_0, s_1, s_2$$

で表されるとしよう。

Q と S の間の距離 $D(Q,S)$ に関しては、以下の 2 つの条件を満たす必要がある。

$$D(Q,S) \geq 0$$

$$D(Q,S) \text{ の値が小さい} \leftrightarrow Q \text{ と } S \text{ は類似している}$$

Q と S の対応 $w(i)$ が与えられて、図 3.11 の Q と S' のように対応付けらるとき、

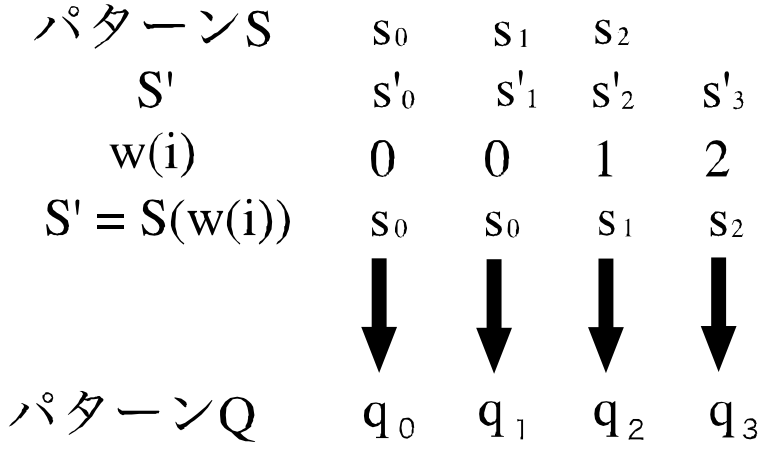

図 3.11: Q と S の関係を表すグラフ

Q と S の間の距離 ($D(Q,S)$ は以下のようになる。

$$D(Q,S) = \sum_{i=0}^{3} d(q(i), s'(i)) = \sum_{j=0}^{3} d(q(i), s(w(j))) \tag{3.1}$$

上式において、$d(q_i, s_i)$ はパターン Q と S の i 番目の要素 q_i と s_i の間の距離を表す。パターン Q と S の対応 $w(i)$ がわからないとき、Q と S の間の距離は、全ての対応 w の中で最も距離が小さくなるときの値をパターン Q と S の間の距離と定義する。

$$D(Q,S) = \min D(Q,S;w) \tag{3.2}$$

DP マッチングは動的計画法と呼ばれる、グラフにおける最適経路を求める方法に基づいている。上式を説明するために、図 3.12 のようなグラフを考えよう。

ノード同士の間にある数字はそのノード間の距離を表す。例えば、 q_0 と q_1 間の距離は 1 である。このとき、例えば開始状態 (q_0, s_0) から (q_1, s_1) までの距離 $|(q0, s0) \leftarrow (q_1, s_1)|$ を $D(q_1, s_1)$ と表す。3.2 式で表される距離 $D(Q,S)$ は $D(q_3, s_2)$ と同値である。

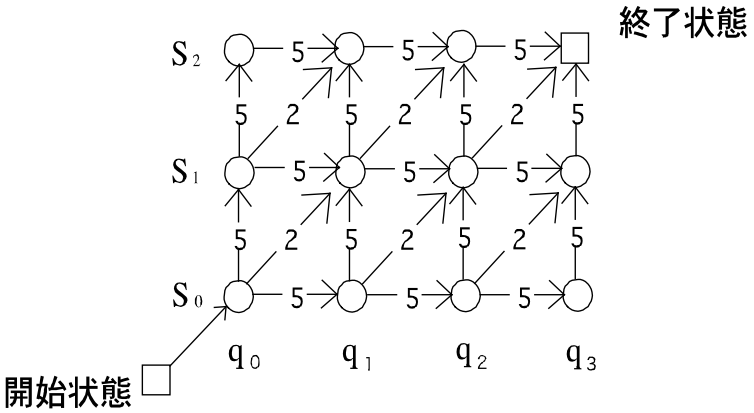

図 3.12: Q と S の関係を表すグラフ

開始状態 (q_0, s_0) から (q_1, s_1) までの経路は以下の 3 通りある。

$$(q_0, s_0) \rightarrow (q_1, s_0) \rightarrow (q_1, s_1)$$

$$(q_0, s_0) \rightarrow (q_1, s_1)$$

$$(q_0, s_0) \rightarrow (q_0, s_1) \rightarrow (q_1, s_1)$$

(q_0, s_0) から (q_1, s_0), (q_1, s_1), (q_0, s_1) までの距離は、それぞれ $D(q_1, s_0)$, $D(q_1, s_1)$, $D(q_0, s_1)$ で表されるため、上の 3 つの経路について、開始状態から (q_1, s_1) までの最短距離は

$$D(q_1, s_0) + |(q_1, s_0) \rightarrow (q_1, s_1)| = 1 + 1 = 2$$

$$D(q_1, s_1) = 0$$

$$D(q_0, s_1) + |(q_0, s_1) \rightarrow (q_1, s_1)| = 1 + 1 = 2$$

の最小値である。よって、(q_0, s_0) から (q_1, s_1) までの最短経路は $(q_0, s_0) \rightarrow (q_1, s_1)$ となる。こうした計算を開始状態（q_0, s_0）から終了状態 (q_3, s_2) まで芋づる式にしていって、$D(q_3, s_2)$ に最短距離を与える経路が最適経路である。$D(q_3, s_2)$ の最短距離が Q と S の間の距離、最適経路は Q と S の対応付けを表す。

図 3.9 のグラフにおいて、最適経路（最短経路）は例えば次のようになるだろう。

開始状態（ $\rightarrow$) $(q_0, s_0) \rightarrow (q_1, s_1) \rightarrow (q_2, s_2) \rightarrow (q_3, s_2)$ ($\rightarrow$) 終了状態

これは、q_0 と s_0 、q_1 と s_1、q_2 と s_2、q_3 と s_3 が対応することを意味している。このとき、開始状態から終了状態までの最短距離 $D(q_3, s_2)$ は 1 である。

以上のように、Q と S においては、符号 $q_0, q_1, q_2, q_3, s_0, s_1, s_2$ の間の距離が数値化できれば、DP マッチングによって Q と S の間の距離を求めることができる。

表 3.1: 各ストローク方向間の距離

	a	b	c	d	e	f	g	h
a	0							
b	2	0						
c	5	2	0					
d	5	5	2	0				
e	5	5	5	2	0			
f	5	5	5	5	2	0		
g	5	5	5	5	5	2	0	
h	2	5	5	5	5	5	2	0

今回の研究では、文字の認識のため、図 3.2 で示した、各ストローク方向の類似度を表 3.1 のように定義した。

表 3.1 においては、数値が大きいほど、その符号同士が類似していることを表す。例えば、a と b の距離は 2 であり、a と c の距離は 5 である。これは、b は c よりも a に類似していることを示している。

今回文字を認識するために用いたアルゴリズムでは、さまざまな漢字について時系列ストローク方向パターンを定義しておき、文字が描かれたら、最も距離が小さいパターンを探し出すことによって文字を認識している。

3.2.2 映像の検索

映像の検索のために、ビデオのメタデータを記述する方法として Mpeg7 を用いた。Mpeg7 とは、Mpeg（Moving picture experts group）により開発された ISO/IEC 国際規格のひとつで、マルチメディア・コンテンツの検索を容易にするために、コンテンツのメタデータを記述するための形式である。メタデータとは、そのマルチメディア・コンテンツの特徴を記述したものである。

Mpeg7 では、メタデータは自己拡張可能なマークアップ言語である XML(eXtensible Markup Language) を用いて表現される。例えば馬が走っている映像の Mpeg7 ファイルを付録 1 に添付する。付録 1 の Mpeg7 ファイルにおいて重要なのは以下の部分である。

$\langle Who \rangle$

$\quad \langle Namexml : lang = "en" \rangle$ horse $\langle /Name \rangle$

$\langle /Who \rangle$

$\langle WhatAction \rangle$

$\quad \langle Namexml : lang = "en" \rangle$ run $\langle /Name \rangle$

$\langle /WhatAction \rangle$

上の例は「馬が走る」映像のメタデータである。上の例のように、「何が」, 「何をしている」ビデオかについてのキーワードをメタデータと記述しておくことで、これらのキーワードをもとに映像検索を行っている。

第4章　考察

入力される言語から映像を構成する手法についての提案することを目的として、人間が視覚イメージを想像する過程に注目し、同様の過程を実現するためにモデルを作成した。4 章では、作成したモデルと人間が視覚イメージを想像する過程とを照らし合わせ、モデルの問題点等について考察する。

2 章で、人間のイメージと言語は、長期記憶と深く結びついていることを述べた。人間においては、イメージと言語の関係は長期記憶の発達に伴って形成される。しかし今回作成したモデルでは、筆者が最初から映像と対応する言語の関係を与えている。よって、作成したモデルは、人間が視覚イメージを想像する過程と同様の処理を行っているとは言えない。コンピュータが視覚イメージを想像するためには、長期記憶の形成を行わなくてはならないと考えられる。そこで、コンピュータに長期記憶を作らせ、言語入力から視覚イメージの想像に結びつけるための方法について今後の展開として考えてみる。

言語入力から視覚イメージの想像に関する課題として、大きくは以下の 3 点が考えられる。

(1). 視覚的情報の符号化と記憶
(2). 言語の習得
(3). 言語入力から視覚イメージを想像する

以上の (1), (2), (3) は、まず言語を学習して、次に長期記憶を形成して、次にイメージを想像してというように順番を持っているわけではない。言語の習得と長期記憶の形成は同時処理されるし、ある言語に対する長期記憶（意味記憶）がなくても人間は、その言語について類推することで、視覚イメージを形成することができる。よって、これら (1), (2), (3) は、同過程において処理されるべきである。

まず (1), (2) について同過程において処理することを前提として解決方法を考えてみる。2 章で Tomacello は語彙獲得の認知的基盤として、他者が話している対象を認知しカテゴリーをつくることを可能とする能力と、他者が言語を使用している際に、その意図が何かを理解できる能力の獲得を挙げていることを述べた。よって、コンピュータが言語と対応する視覚的特長を学習するためには、視覚的情報と言語的情報のペアを取得し、以前与えられた視覚的情報と言語的情報との相違点を見つけ、カテゴリー分けすることが必要である考えられる。ここで視覚的情報をカテゴリー分けするために、視覚的情報の符号化が必要となる。

人間においては、視覚的情報は、形態・色情報と空間・運動情報に符号化されると考えられている。[13]　コンピュータも、取得された視覚的情報列を形体・色情報と空間・運動情報に符号化した情報をもとにカテゴリー分けを行い、外部と内部からのフィードバックを通じて、言語を獲得していくことができるのではないかと考えている。外部的なフィー

ドバックとは、外部の人間が「それは違う」あるいは「それは正しい」と教えてやることである。内部的なフィードバックとは、実世界のある時間・状態において、次の時間・状態において得る言語的情報と視覚的情報を予測し、その後実際に得られた言語的情報と視覚的情報との比較を行うことである。

(3) が今回の研究で目標とした部分であったのだが、今回実装したモデルの問題点は、以上の (1),(2) で述べたような点を考慮していなかったことにある。そのため、今回実装したモデルでは、言語と視覚イメージを直接結びつけてしまった。ここでは、(1), (2) を考慮した上で、2.3.4 節で述べた Baddeley のワーキングモデルをもとに、言語入力から視覚イメージを想像する方法について考えてみる。図 4.1 は、2.3.4 節で述べた Baddeley のワーキングモデルである。言語入力から視覚イメージを形成する過程として、(a) 言語が

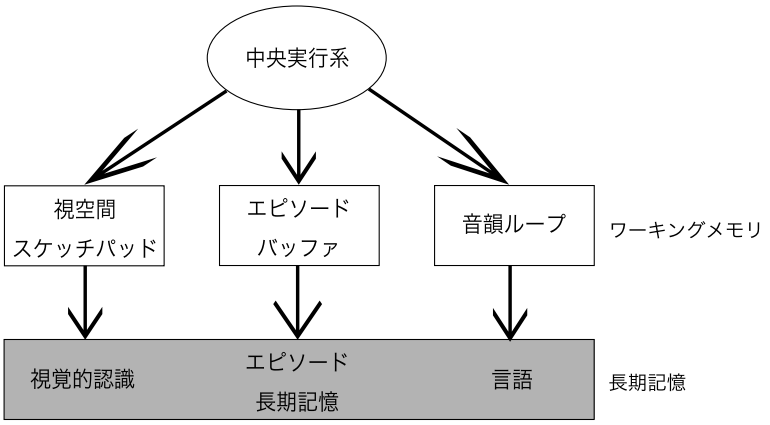

図 4.1: Baddeley(2000) のワーキングメモリ・モデル

入力されたとき、長期記憶（意味記憶）をもとに、言語が認識される。(b) 言語が認識されると同時に長期記憶が活性化する。(c) 長期記憶から視空間的スケッチパッドと呼ばれる視覚的ワーキングメモリに、視覚イメージが描かれるの 3 段階が考えられる。(3) を実現するためには、上に述べた (a), (b), (c) を実行する必要がある。

また、言語が理解される、文章が理解されるということは、このようにして長期記憶から呼び出された情報が統合されることであると考えられる。

第5章　結言

本研究の目的は入力される言語から映像を構成する手法について提案することであった。その上で著者は人間の脳のイメージ機能に注目した。今回の研究においては、映像は、イメージのなかの視覚的なもの、視覚イメージに該当する。コンピュータに、人間と同様、視覚イメージを想像させることを目標とし、そのためのモデルを作成した。

今回作成したモデルと人間が視覚イメージを想像する過程を照らし合わせると、いくつかの問題点があった。大きな問題は、今回作成したモデルにおいては筆者が言語と映像の関係づけを行っているのに対し、人間は経験によって言語とイメージの関係を構築していくという点である。視覚的情報の記憶と言語情報の記憶の関連など、厳密に人間が視覚イメージを想像する過程をモデルにするには、明らかにしなくてはならない点が多くある。しかし、言語入力から映像を構成する手法について提案を行うために、人間の脳のイメージ機能に関して注目したことは、映像構成の手法の一つとしても有意義であったと考えている。

謝辞

本研究を行うにあたってご指導賜りました、迎山和司先生（公立はこだて未来大学）に感謝の意を表します。また、迎山研究室のメンバーにはいろいろとお世話になりました。本研究の初期に有意義な助言をいただいた、柳英克先生（公立はこだて未来大学）と美馬義亮先生（公立はこだて未来大学）に感謝します。

参考文献

[1] 乾敏郎・安西祐一郎編: 認知科学の新展開４　イメージと認知, 岩波書店, (Aug. 2001)

[2] Mellet, E., Tzourio, N., Crivello, F., Joliot, M., Denis, M., and Mazoyer, B.: Functional anatomy of spatial mental imagery generated from verbal instructions. Journal of Neuro-science, 16, 6504-6512 (1996)

[3] Kosslyn S.M., Thompson, W. L., Kim, I. J., and Alpert, N. MM.: Topographical representations of mental images in primary visual cortex. Nature, 378, 494-498 (1995)

[4] 乾敏郎・安西祐一郎: 認知科学の新展開３　運動と言語, 岩波書店, (Sept. 2001)

[5] Karmiloff, K. and Karmiloff-Smith, A.: Pathways to language: From fetus to adolescent. Harvard University Press, (2001)

[6] Tomasello, M.: The pragmatics of word learning. 認知科学, 4(1), 59-74, (1997)

[7] 錦見美貴子: 認知科学モノグラフ　　言語を獲得するコンピュータ, 共立出版, (Sept, 1998)

[8] 芋阪直行編著: 認知科学の探求　意識の認知科学－心の心の神経基盤, 共立出版, (Aug, 2000)

[9] Baddeley, A.D.: Working memory and distributed vocabulary learning. Applied Psycholinguistics, 19, 537-552, (2000)

[10] Friedman, N. P. and Miyake, A: Differential roles for visuospatial and verbal working memoryu for situation model construction. Journal of Experimental Psychology, General, 129, 61-83

[11] 立川健二・山田広昭: 現代言語論　ソシュール　フロイト　ウィトゲンシュタイン, 新曜社, (Jun. 1990)

[12] Paivio, A.: Mental Repre sentations, A dual coding approach. Oxford Press, (1986)

[13] 太田信夫, 多鹿秀継編著: 記憶研究の最前線, 北大路書房, (Feb, 2000)

[14] Mazoyer, B., Tzurio-Mazoyer, N., Mazard, A., Denis, M., Mellet, E.: Neural bases of image and language interactions, International Journey of Psychology, 37(4), 204-208 (2002)

[15] 岡田晋著: 映像学・序説, 九州大学出版会, (1996)

付録その1

$\langle ?xmlversion = "1.0" encoding = "iso-8859-1" ?\rangle$
$\langle Mpeg7 xmlns = "urn:mpeg:mpeg7:schema:2001" xmlns:xsi = "http://www.w3.org/2001/XMLSchema-instance" xmlns:mpeg7 = "urn:mpeg:mpeg7:schema:2001" xsi:schemaLocation = "urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd" \rangle$
$\langle Description xsi:type = "ContentEntityType" \rangle$

$\langle MultimediaContent xsi:type = "VideoType" \rangle$

$\langle Video \rangle$
$\langle MediaLocator \rangle$
$\langle MediaUri \rangle$
$C:/ProgramFiles/eclipse/workspace/C2V/moviesamples/horse_run/$
$\langle /MediaUri \rangle$
$\langle /MediaLocator \rangle$
$\langle TextAnnotation \rangle$
$\langle FreeTextAnnotation xml:lang = "en" \rangle$
if lion walk
$\langle /FreeTextAnnotation \rangle$
$\langle StructuredAnnotation \rangle$
$\langle Who \rangle$
$\langle Name xml:lang = "en" \rangle$ horse $\langle /Name \rangle$
$\langle /Who \rangle$
$\langle WhatAction \rangle$
$\langle Name xml:lang = "en" \rangle$ run $\langle /Name \rangle$
$\langle /WhatAction \rangle$
$\langle /StructuredAnnotation \rangle$
$\langle /TextAnnotation \rangle$
$\langle /Video \rangle$

$\langle /MultimediaContent \rangle$

$\langle /Description \rangle$
$\langle /Mpeg7 \rangle$

図 目 次

表 目 次