博士論文

目的や対象に応じたモデル化による画像認識手法に関する研究

指導教官　坂内 正夫　教授

東京大学大学院　工学系研究科　電子工学専攻

27124　孟 洋

1996年12月20日　提出

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 研究の背景と目的

近年、コンピュータや通信衛星、光ファイバの普及に伴う放送型情報のディジタル化により、テレビ放送の多チャンネル化に代表されるように、極めて大規模な映像を中心としたマルチメディア情報の流れが出現しつつある。また、ビデオカメラやビデオデッキなどの映像情報の入力・蓄積装置の普及、ハードディスクや光ディスク、CD-ROMなどの情報蓄積媒体の低価格化、大容量化などに伴い、利用者自身による映像情報の獲得や、出版など多様なメディアによるマルチメディア情報の配布が可能となり、放送など様々なメディアをとおして映像を中心とした多くの情報を利用者が獲得できる環境が整ってきている。しかし、このようなマルチメディア環境の出現の一方で、その情報量の膨大さや情報の広汎な分布などのために、利用者がその情報の海を有効に利用することが困難な状況にもなりつつある。このため、情報空間と利用者の間に介在し、利用者が必要とする情報の獲得を可能とする情報処理技術、特に映像情報の高度な処理を可能とする画像認識技術の研究、開発が重要になってきている。

画像認識技術とは、人間の視覚認識能力の工学的手段での実現を目指すもので、一般に、計算機による図面の認識や、画像・映像に示された物体やシーンの認識、画像・映像からの三次元構造の復元などを可能とする技術を意味する。通常、計算機による画像認識は、認識対象に関する知識に基づき、認識戦略に従って実行される。このため、認識対象に関する知識をどのように表現するか、認識をどのように行うかが重要な問題となる。これらは、認識の目的が何であるのか、認識の対象は何であるのか、どのような画像を入力とするのか、どのような情報を獲得できればよいのかなど、認識における様々な要件から決定されるものである。現在までに、多くの認識システムが考えられてきているが、その多くは、厳密な幾何学的形状をあらわすモデルを用いて、システムに組み込まれた認識戦略に従って認識を実行するものであった。これは、三次元構造の復元を中心に研究が進められているロボットビジョンをはじめ、はやくから画像認識技術が必要とされていた工業分野での応用において、物体の幾何学的形状が重要な役割を担っていたためである。

マルチメディア環境における映像情報処理との観点から画像認識技術について考えてみると、その大きな特徴の一つに扱う対象の世界が広範であるということがあげられる。従来の工業分野などにおける画像認識システムでは、ある特定の条件で撮影された画像から、

製品や部品などの特定物体の存在や位置、姿勢を認識することが目的とされ、対象の世界が強く限定されていた。このため、前述したように、多くの画像認識における研究、開発が、幾何学的形状による物体のモデル化や、画像からの幾何学的構造の抽出など、対象の幾何学的情報の取り扱い方を中心に行われていた。しかし、マルチメディア環境においては、その対象とする映像は、一般のテレビ映像にはじまり、WWW(World Wide Web) の利用に代表されるインターネット上を流れる映像、家庭で撮影されたビデオ映像など多彩なものが考えられ、目的においても、多くの映像情報からの必要な映像情報の選択、映像からの特定物体の抽出、映像情報の内容理解に伴う意味的情報の獲得など様々なものが考えられる。そして、目的に関して更に言及すれば、従来の画像認識技術が、危険な作業の自動化や、製品の組み立て、管理、検査の自動化など、作業の安全性、効率化に重点をおいたものであったのに対し、マルチメディア環境では、人間と同様の高度で柔軟な画像認識の実現、言い換えれば、利用者の様々な要求への対応や利用者の利便性の向上を可能とする情報の選択、獲得の実現に重点がおかれる。これは、従来の幾何学的情報の取り扱いを中心とした画像認識技術とは別に、マルチメディア環境をにらんだ映像情報処理という観点からは、様々な目的、様々な対象を扱える画像認識技術の研究、開発が望まれることを意味する。

現在、映像情報に対し様々な情報を付加する試みがなされており、放送メディアにおいては映像情報とともに文字情報を提供する文字放送などが行われている。また、放送番組のタイムスケジュールやニュースの内容、ドラマのシナリオなどの番組情報などは、一部インターネット上などにも流されており、自由に利用することが可能となっている。これらの情報は、利用者の利便性の向上をはかるために付加、あるいは、提供されているものであるが、しかし、これらの情報は情報提供者側の立場で付加されたもので、利用者の要求に細かく答えることができるものではない。利用者の要求に答えるためには、必要に応じて必要な情報を獲得できる技術、つまり、画像・映像情報から直接様々な情報を獲得できる技術が必要となる。マルチメディア環境における画像認識の目的は、利用者の要求そのものであるといっても過言ではない。このため、今後、どんなに利用価値の高い情報が映像とともに提供されるとしても、画像・映像情報から直接必要な情報の獲得を実現する画像認識技術への期待は高まる一方であるといえる。

以上、マルチメディア環境における画像認識技術の重要性を述べた。マルチメディア環

境においては、利用者の要求が多彩であるため、様々な目的や対象に対応できる画像認識技術が必要である。しかし、現在の画像認識技術の水準を考えると、人間同様の高度な画像認識能力を期待することはできない。このため、対象の世界を限定し、画像認識に必要な機能を特化、簡略化することが必要となり、汎用なモデル化手法や認識手法を検討するのとは別に、特定の目的、対象に応じた、つまり扱う情報に応じた認識手法の検討を行うことが重要となる。これは、多くの映像から必要な映像を選択するような場合には、映像のシーンを判別できる技術が必要であるし、映像の内容を理解するような場合には、自然画像の処理技術や動画像の処理技術が必要であるというようにである。

本研究では、利用者の要求に応じた高度な映像情報処理の実現との観点から、映像情報処理におけるいくつかの具体的な目的や対象を想定し、その状況における画像認識モデル、認識手法のあり方の検討を行い、それら状況に応じた画像認識の枠組を提案することを目的とする。

本論文では、はじめに、われわれの研究室で開発された様々な認識に利用可能な状態遷移モデルを取り上げる。状態遷移モデルとは、認識プロセスを「状態の遷移」という形でモデル化するもので、対象に依存しない枠組を提供する。このため、「状態」の定義の仕方により、図面の認識から画像の認識まで様々な対象、目的に利用することが可能である。続いて、状態遷移モデルを定義するルールの作成を容易にするルール作成支援システムについて取り上げる。このシステムでは、状態遷移ルールのうち、ボトムアッププロセスを定義するルールの作成を対象としており、グラフィカルユーザインターフェースを用いた人間と機械の対話をとおして、ルールの作成が可能である。第3章において、地図図面の認識を例に、状態遷移モデルによる認識の枠組とルール作成支援システムについて述べる。

次に、画像・映像のシーン/カットの認識を実現する、画像データベースを用いたシーン/カットの同定手法を取り上げる。一般に、画像のシーン/カットの認識は画像を構成する物体が理解できてはじめて可能となると考えられるが、物体の認識はシーン/カットが認識できていないと困難である。そこで、画像のシーン/カットを画像自体でモデル化し、画像のシーン/カットの認識を物体の認識なしに行う手法を提案する。この手法は、同様のシーン/カットの画像同士は類似している可能性が高い点に着目し、画像データベースと類似画像検索技術を用いることで、画像のシーン/カットの同定を試みるものである。映像への対応を含め、詳細は第4章において述べる。

続いて、認識対象の映る大きさに依存しない画像認識を試みるモデル化手法について述べる。一般の画像では、向きや大きさの違いなど、認識対象の映り方は様々である。現在までに、映る向きの違いへの対応については多くの研究がなされてきたが、映る大きさの違いについてはあまり考えられていなかった。そこで、映る大きさの違いへ柔軟に対応できるように、映る大きさによる特徴の違いを階層的に表現し、これらを統合的に利用できるモデル化手法の検討を行った。本研究では、このモデル化手法によるモデルを階層型距離モデルと呼ぶことにする。第5章において、人間の認識を例に取り上げ、階層型距離モデルの枠組について詳しく述べる。

最後に、提案したモデル化手法、及び、これらモデル化による画像認識の枠組の有効性を示し、今後のモデル化手法の一つの方向性について論じることで、本論文のまとめを行う。

1.2 本論文の構成

本論文の構成は、以下のとおりである。

本章である第1章では、本研究の背景と目的、及び、本論文の構成について述べる。

第2章では、画像認識とモデル化の技術について概説する。ここでは、画像認識におけるモデル化のあり方を、目的や対象に応じたモデル化という観点から整理し、本研究の立場、及び、位置づけを明確にする。

第3章では、図面や画像などの認識を可能とする多目的型モデルである状態遷移モデルを取り上げる。また、状態遷移ルールの作成を容易にするために構築した人間機械協調型のルール作成支援システムについて述べる。

第4章では、シーンレベルでの画像の認識を可能とする画像・映像のシーン/カットの同定手法について述べる。

第5章では、認識対象の映る大きさに依存しない画像認識を試みる階層型距離モデルについて述べる。

最後に、第6章で、本研究のまとめについて述べ、結論とする。

第 2 章

画像認識技術と対象のモデル化

2.1 概要

近年の情報処理機器の発達や情報通信基盤の整備などに裏付けられたマルチメディア環境の出現に伴い、画像・映像情報の高度な利用を可能とする技術への要求が高まってきた。このため、人間の優れた画像・映像情報の処理能力を、計算機で実現しようとする画像認識技術の研究、開発が盛んに行われている。

計算機に画像認識を行なわせるためには、対象とする世界に関する知識、認識対象に関する知識、認識戦略に関する知識など様々な知識が必要となる。これらは、どのような画像を対象とするのか、どのような情報を認識したいのかという画像認識の目的と密接に関連する。現在までに、様々な目的、様々な対象に対して、認識対象の表現方法、認識の実行方法などが検討されてきたが、その多くは、工業分野での利用を前提に物体の検出を目的としたもので、認識対象を幾何学的形状にて表現するものであった。このような幾何学的形状に重点をおく考え方は、実世界において人工物の多くが決まった形状を持つことから考えて理にかなっているといえるが、しかし一方で、自然画像などを扱う場合、形状では明確に定義できない認識対象も数多く、また、高度な画像認識という観点からは概念などの意味的な認識の実現が必要不可欠であり、様々な情報を扱える技術が必要とされている。これら様々な情報をどのように表現すれば良いのかという問題は、人工知能などの知識処理技術と深く関連し、幾何学的形状の表現に比べ、より多くの難しい側面を持っている。このような中、現在までに、人工物体の認識に機能情報の利用を試みた機能ベースの認識システム、自然風景などの認識に属性と相互関係情報の利用を試みたコンテキストベースの認識システムなどの検討が行なわれている。

本章では、画像認識の背景技術をふまえ、画像認識の形態を、知識の表現、特に、目的や対象に応じたモデル化という観点から整理し、マルチメディア環境をにらんだ画像認識技術に要求される要件を明らかにすることで、本研究の位置付けを明確にする。また、背景技術としては、画像検索において、検索が画像認識技術に基づいて行われることや、検索された画像は何らかの意味付けがなされたと捉えることができるなど、画像検索技術は画像認識技術と密接に関連しているといえることから、画像認識技術に加え、画像検索技術の現状についても取り上げる。

2.2 マルチメディア環境での画像認識技術

2.2.1 画像認識技術の現在

画像認識の研究は、現在までに工業分野をはじめとする様々な分野で行なわれ、多くのモデル化手法、様々な認識手法が提案されてきているが、古くは画像の特徴量による分類などを目的としたパターン認識の研究にはじまり、特に三次元構造の認識という点ではMITのRoberts(1965)が先駆けであるといわれている。

一般に、画像認識はボトムアップ解析とトップダウン解析の二つのプロセスにより実行される。ボトムアップ解析とは、一般的な光学的性質や幾何学的性質の知識をもとに、画像からシーンの再現を行なうものであり、トップダウン解析とは、認識対象のモデル、あるいは、認識対象に関する知識をもとに、モデルと画像との照合を行うことで対象の認識を行うものである。

トップダウン解析では、認識対象のモデル化、あるいは、認識対象に関する知識の表現が重要となり、工業製品のような固定形状の認識対象をあらわす幾何学的形状モデル[23]や、風景などの不定形状の認識対象を認識するためのコンテキストモデル[27]などが提案されている。なかでも幾何学的形状モデルにいたっては多くのものが検討され、三次元形状を直接モデル化する一般化円筒や、表面によりモデル化する拡張ガウス像、二次元の様々な様相によりモデル化するアスペクトグラフなどがある。最近では、意味的認識の実現との観点から、認識対象の機能性に着目した機能モデル[24][25][26]の考え方も提案されている。また、複雑な物体の構造や概念間の関係などの表現のため、多重解像度や上位-下位関係、全体-部品関係、類似-差異関係などの階層的な構造関係や、配置などの空間的関係の記述方法も重要な課題となっている。一方、ボトムアップ解析では、一般的な知識によるシーンの再現との観点から、物体の拘束条件による積木の世界の線画の解釈や、視点のズレを利用し三次元構造を復元するステレオビジョンなどの研究が古くから行なわれている。

また、最近、ロボットの視覚システムの実現を試みるロボットビジョンにおいては、画像の持つ三次元構造を復元して利用すべきであるというMarrのパラダイム[40]に対し、三次元構造の完全なる復元を目指すのではなく、目的に応じた情報のみを取り出し利用すればよいという目的指向型ビジョン[20][21]の考え方や、従来の単に与えられた画像を受動的に解析するパッシブビジョンに対し、カメラを積極的に制御し、解析に都合のよい画像

を取り込もうとするアクティブビジョンの考え方が示されており、画像認識研究の一つの流れをつくっている。

2.2.2 マルチメディア環境における画像認識技術への要求

「百聞は一見にしかず」という言葉から明らかなように、画像・映像情報は、人間に対し、瞬時に多くの情報を与えてくれる。このため、情報処理機器の発達に伴う、画像・映像情報を扱うためのコストの削減により、このような優れた情報伝達能力を持つ画像・映像情報を利用することへの要求が高まり、今では情報伝達の中心的役割を演ずるまでになった。

現在、多くの画像・映像情報が身のまわりに氾濫してきており、利用者が多くの情報を獲得できる環境が整いつつある。しかしその反面、画像・映像情報の氾濫は、情報の取得、管理を困難とし、せっかくの情報を利用者が上手に利用できないという問題も生じている。画像・映像情報は、今後もますます増加する一方であると考えられるため、より一層この傾向は強くなる。このため、利用者がこれら多くの情報から、必要とする情報を発見、獲得できる技術への期待が高まっている。画像認識技術による画像・映像情報空間へのアクセスの概念図を図 2.1に示す。

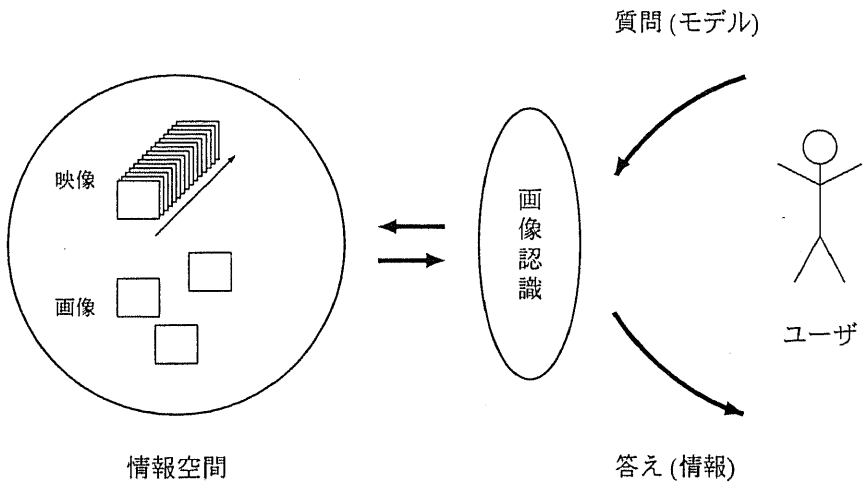


図 2.1: 画像・映像情報空間へのアクセス

マルチメディア環境における高度な画像・映像情報処理の実現との観点からは、(1) 利用者の多様な要求に対応できること、(2) 扱える対象世界が広範であることが重要なポイントとなる。これは、特定の映像だけでなく、テレビ映像などより一般的な映像を扱え、また、特定の物体を検出するだけでなく、映像内容の意味的な認識により、「野球の試合の結果は？」、「～のニュースを見せて」など、利用者の様々な要求に答えることができる技術が望まれていることを意味する。

このような認識を画像認識技術で実現するためには、最終的に概念レベルのシーンの記述を得ることができなければならない。しかし、シーンの記述を求めるためには、何らかの知識を用いて、概念と画像情報の対応づけを行なわなければならないが、現在の画像情報処理、知識情報処理のレベルでは、対象とする問題をできる限り特化、簡単化しても扱うことが難しい。これは、概念などの表現が困難であること、情報の種類が様々でこれら情報を統一形式で表現することが困難であること、対象や獲得すべき情報に応じて画像認識手法を選択する必要があることなどの理由による。

マルチメディア環境においては、人間主体の情報空間へのアクセス技術が望まれ、なかでも画像・映像情報を扱える画像認識技術が重要となる。特に、高度なアクセスを実現するためには、様々な情報を扱えることが必要であり、画像認識技術の現状から考えると、対象世界の限定と知識の活用の方法がポイントになるといえる。本研究の主題である、目的や対象に応じたモデル化の必要性は、ここから導かれる。

2.3 画像認識技術

2.3.1 画像認識の枠組

図 2.2に画像認識の枠組を示す。一般に、画像認識は、画像からの特徴抽出、記述作成、そして、認識対象をあらわすモデルとの照合という流れにそって行われる。はじめの画像からの特徴抽出では、画像の色や明るさが変化する部分であるエッジや、色や明るさが一様な領域などの特徴を、画像の局所的な情報を利用して抽出する。次の記述作成では、特徴抽出で求めた局所的な特徴を利用して、それらの関係から、線分や面、三次元構造などの対象のシーンの記述の作成を行う。最後の照合では、認識対象のモデルと求められたシーンの記述の照合を行うことで、認識目的にそった情報を獲得し、認識結果として出力する。

通常、このモデルとの照合は、画像レベル、二次元特徴レベル、三次元特徴レベルなどと様々なレベルで行なわれる。これら画像認識の各段階では、画像の撮影条件に関する知識にはじまり、物理的条件に関する知識、認識対象に関する知識など、様々な知識が利用される。

基本的な画像認識の枠組は以上のようなものであるが、一般に画像認識は大きく二つのプロセスにより実現される。一つはボトムアップ解析、もう一つはトップダウン解析である。

ボトムアップ解析

入力画像から出発し、画像からの特徴抽出、記述作成などというように、順次、情報のレベルを高度化し、最終的にシーンの記述を求めるデータ駆動の解析。

トップダウン解析

認識対象のモデルに基づいて、必要な特徴の抽出や、構造化などを行なうモデル駆動の解析。

原理的には、ボトムアップ解析のみでシーンの記述を求めることができ、画像の認識が可能であるはずであるが、(1) シーンには多くの認識対象が含まれ、モデルと画像特徴の照合が困難、(2) 画像に含まれる、意味のある特徴を完全に抽出することが困難などの問題があり、一般にボトムアップ解析だけでは十分な結果は得られない。このため、トップダウン解析が必要となる。トップダウン解析では、何を抽出すればよいのか、何を認識すればよいのかなどが明確なため、効率的で効果的な解析が実行できる。

2.3.2 画像認識における知識

人間が画像を見て、それが何であるか理解できるのは、画像から適切な特徴を抽出し、種々の知識を用いてその特徴を解釈することができるためである。このため、計算機による画像認識においても同様に、画像認識のための様々な知識が必要となる。一般に、画像認識で必要な知識は、物理的条件に関する知識、認識プロセスに関する知識、認識対象のモデルというように大きく三つに分けて考えることができる。物理的条件に関する知識とは、照明条件などの撮影環境における種々の物理的な条件に関する知識、認識プロセスに関する知識とは、画像からどのような特徴を抽出し、認識対象のモデルとどのように照合

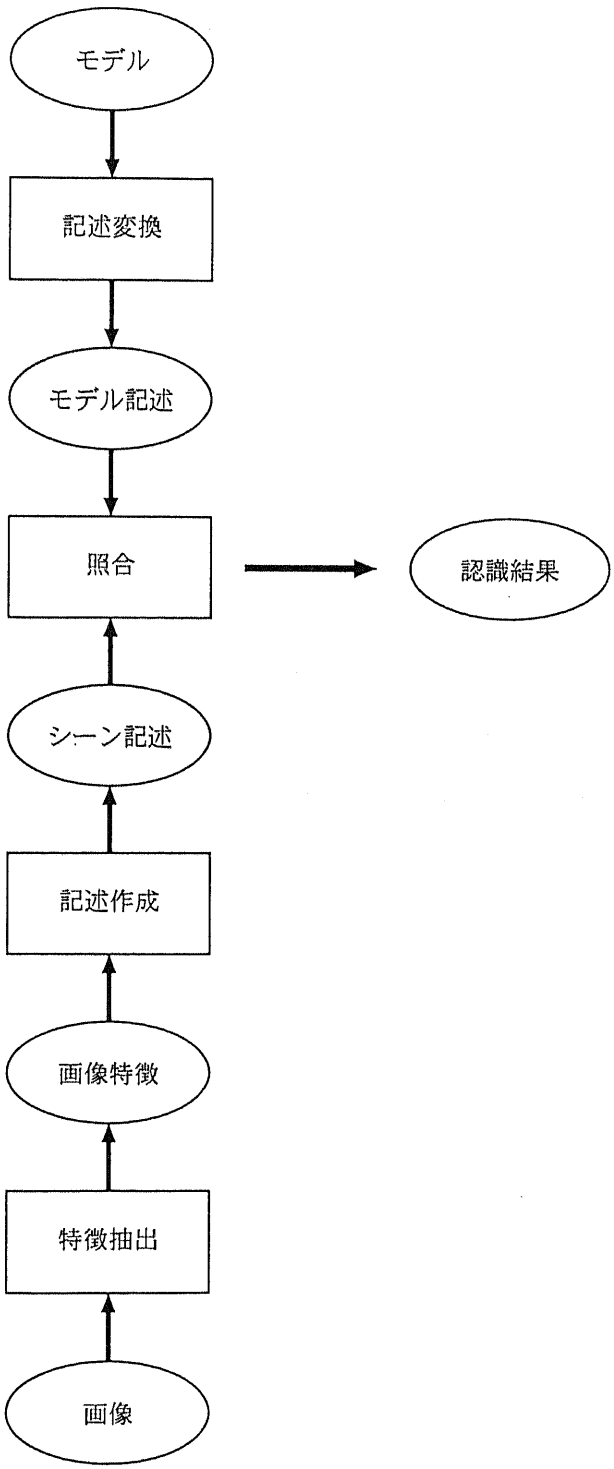


図 2.2: 画像認識の枠組

を行なうかなどの認識戦略に関する知識、認識対象のモデルとは、認識対象がどのようなものであるのか、どのような情報を獲得したいのかをあらわす知識のことである。次節では、知識の中でも対象に特に深く関わる、認識対象のモデルの表現の現状について取り上げる。

2.4 認識対象のモデルの表現

2.4.1 モデルの表現

初期の画像認識システムは、扱う画像を強く限定し、また、獲得すべき情報も限られたものとして開発されていた。このため、対象のモデルや認識戦略など認識に必要な知識はシステムに組み込まれた形で表現され、汎用性の乏しいものであった。このようなシステムは、特定の目的に対しては、効果的、効率的な認識を実現するが、対象の変更などには対応することが困難であった。

本節では、幅広い対象を扱うシステムの開発のためには重要となる、認識対象のモデルの表現について考える。認識対象のモデル化の方法は、現在までに様々な研究を通して多くのものが検討されてきた。そこで、ここでは、認識対象のモデルの表現について、表 2.1 に示すように、知識の記述手法、階層構造の表現手法、対象のモデル化手法と三つの観点から整理することにする。

表 2.1: 認識対象のモデルの表現

	例
知識の記述	フレーム、プロダクションルールなど
階層構造の表現	上位-下位関係、部分-全体関係、多重解像度関係など
対象のモデル化	幾何学的形状モデル、機能モデル、コンテキストモデルなど

2.4.2 知識の記述手法

知識の記述として特に重要なのは汎用性の問題である。ここでは、対象に関する知識を一様な形式で表現できる、プロダクションルールとフレームについて取り上げる。

プロダクションルール

プロダクションルールを用いたシステムはプロダクションシステムと呼ばれる。プロダクションシステムは、プロダクションルールにより記述される知識ベースを持ち、この知識と前向き推論を用いて問題解決をはかる推論システムである。プロダクションルールは、IF (条件部) THEN (実行部) の形式で記述されるもので、宣言的知識表現を試みるものである。認識システムでの利用においては、プロダクションルールにより、対象物に関する断片的な知識を表現し、対象物認識のための各種知識を記述する。プロダクションルールでは、知識が記号を用いて宣言的に表現できるため、知識の追加、修正が容易であるといえる。

プロダクションシステムを利用したものとしては、D.M.Mckeown による航空写真の画像の認識システム [28] があげられる。このシステムは、空港の画像を認識するためのもので、空港の画像に依存した部分をプロダクションルールにより記述でき、認識はプロダクションシステムにより実現されている。このシステムでは、ルールとして認識を実現する認識ルール、及び、検証を行なうための検証ルールを与えるようになっている。

フレーム

プロダクション規則は断片的な知識を表現するには都合が良いが、画像に含まれる対象物の認識を行なうような場合には、ある特定の対象物に関する知識は一箇所にまとめて記述できるほうが都合がよい。人工知能の分野では、ある概念に関するまとまった知識を記述する方法としてフレームがよく用いられる。フレームはいくつかのスロットから構成され、各スロットはそのフレームが表す概念のもつ属性や他の概念との関係を表すのに用いられる。このため、汎化-特化関係や全体-部分関係などの知識の階層的記述が可能である。

フレームを利用したものとしては、R.A.Brooks による画像認識システム ACRONYM[23] があげられる。このシステムは、空港の画像から飛行機を見つけ、機種の判定を行なうものである。対象のモデルはフレームにより記述され、フレームを用いることで、クラスと具体例を同時に表現するための上位-下位関係に基づいた階層関係をうまく表現している。

2.4.3 階層構造の表現手法

認識対象のモデル化を考える上では、いかにすれば認識対象をうまく表現できるかが重要な問題となる。ここでは、複雑な知識を表現するための手段である階層化について取り

上げる。知識を階層的に表現するといっても、何をどのように階層化するのかにより、様々な形態のものが考えられるが、大きくは、構造的知識の階層化と情報の記述レベルの階層化の二つに分けて考えることができる。これら階層関係を用いると、階層的関係を利用することによる認識の効率化や、階層的協調を利用することによる認識の高度化を実現できる。

構造的知識の階層化

構造的知識の階層関係としては、認識対象の上位–下位関係や全体–部分関係などがある。これは、例えば、家具と椅子や椅子と座面などの関係のことであり、図 2.3に示すような階層構造で表現される。

このような階層構造を利用したシステムとしては、R.A.Brooks による ACRONYM[23] がある。このシステムは、三次元モデルと二次元画像の照合を行う汎用的なシステムで、フレームのところでも述べたが、モデルをパラメータ化して階層的に記述することで、クラスと具体例という上位–下位関係をうまく表現している。

情報の記述レベルの階層化

情報の記述レベルの階層関係としては、認識対象の局所的特徴と部分的特徴の関係や多重解像度間の関係などがある。これは、例えば、エッジ、線、多角形や物体の輪郭、部品の形状などの関係のことであり、図 2.4に示すような階層構造で表現される。

このような階層構造を利用したシステムとしては、R.M.Bolle らによるシステム [30] がある。このシステムは、下位層の特徴から上位層の特徴を構成し、その際生じる仮説間の関係をネットワークで表すことにより、グローバルに一貫した認識を実現している。

2.4.4 対象のモデル化手法

現在までに様々な認識対象に対してモデル化が検討されてきた。ここでは、いくつかの目的や対象を取り上げ、代表的なモデル化手法について述べる。認識対象とモデル化の例を表 2.2に示す。

幾何学的形状のモデル化

幾何学的形状のモデル化とは、その名のとおり、個々の認識対象の形状をモデル化するもので、一般的な認識システムでよく用いられている。この方法は、認識対象そのものの

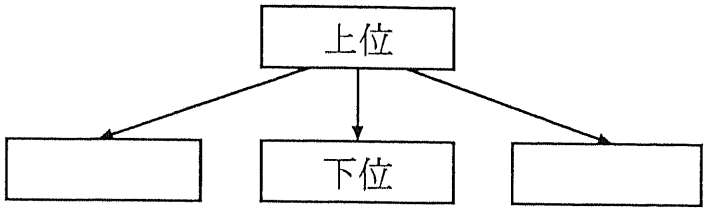


図 2.3: 構造的知識の階層化

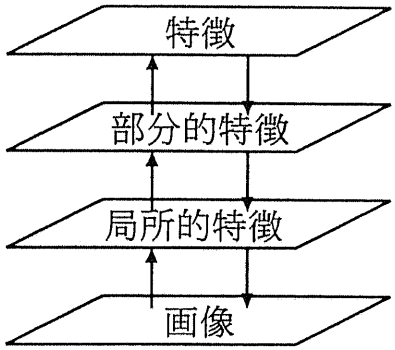


図 2.4: 情報の記述レベルの階層化

表 2.2: 対象のモデル化手法

対象	評価特徴	例
物体	幾何学的形状	テンプレート、アスペクトグラフ、一般化円筒など
物体のカテゴリー	機能性	機能モデルなど
自然風景	属性、相互関係	意味ネットワーク、コンテキストモデルなど
動画像	位置、動き	VSDL(Video Scene Description Language) など

形状をモデル化するため、同一カテゴリーに属する似たようなものを認識する場合でも、対象ごとにモデルを設定しなければならない。

幾何学的形状をモデル化する方法は、数多く考えられてきた。これらは大きく、見かけの形状で表現するものと、三次元座標系での形状で表現するものに分けることができる。

- 見かけの形状での表現

  見かけの形状での表現手法としては、面関係グラフ法、アスペクトグラフ法、Winged-edge 法、拡張ガウス像法などがある。面関係グラフ法は、三次元物体での見える面とそれらの関係を、面の属性と面をノードとするグラフにより記述するものである。アスペクトグラフ法は、三次元物体をトポロジー的に異なる形状の集合で記述するもので、トポロジー的に異なる形状は、その変化の関係をあらわすグラフにより関係づけられる。

- 三次元の座標系での形状での表現

  三次元の座標系での形状での表現手法としては、一般化円筒法、CSG(Constructive Solid Geometry) 法、超二次曲面法、Oct-tree 法などがある。一般化円筒法は、三次元物体を一本の軸とその軸のまわりにおける物体の太さで記述するもので、空間中における軸の変化と軸のまわりにおける物体の太さの変化を表す二つのパラメータで記述する。CSG 法は、複雑な三次元物体を単純な物体の組合せで記述するもので、組合せは物体の和や差といった集合演算で記述する。

幾何学的形状モデルを用いた認識システムには、R.A.Brooks による ACRONYM システム [23] などがある。詳しくは次節で紹介する。

機能のモデル化

幾何学的形状のモデル化では、同一カテゴリーに含まれる物体を認識する場合でも、物体ごとにモデルを用意する必要がある。このため、一般的な物体の認識を行うことは困難であった。機能のモデル化とは、このような問題を解決するために、同一カテゴリーに含まれる物体の共通の機能をモデル化することで、物体をカテゴリーという枠で統一的に扱おうとするものである。

物体が持つ様々な機能的特性を表現するために必要な知識としては、次のものが考えられている [24]。

- 物体の機能

  物体の機能は、その物体を構成する部品の小機能が集積されることにより、実現されていると考えられる。このため、物体の機能を小機能へ分割することで、機能の階層的なモデル化が可能となる。一般に、小機能を部分機能、部分機能を提供する部品を機能要素という。

- 機能環境

  物体がそれ自身の機能を果たすためには、その機能が要求する周辺環境が整っている必要がある。このような周辺環境のことを、機能環境という。

- 使用者と作用対象

  物体の機能を考える場合、機能を中心として使用者と作用対象が存在する。使用者と物体は、使用部位、使用形態により連結され、また、作用対象と物体は、作用部位, 作用形態により連結される。このため、物体の形状、位置などの特徴が制約されることになる。ここで、使用部位とは使用者と物体との介在部分、使用形態とはその介在部分の部分機能、また、作用部位とは作用対象と物体との介在部分、作用形態とはその介在部分の部分機能のことである。

- 機能の相補性

  一般に、シーンは複数の物体により構成される。このため、物体同士の相性や組み合わせによる機能的特性の変化について考慮する必要がある。このような機能の相互関連のことを、機能の相補性という。

機能モデルを用いた認識システムには、L.Stark らによる GRUFF2 システム [25] などがある。詳しくは次節で紹介する。

コンテキストのモデル化

自然の風景にあらわれるような木や空といったものは、形状や機能では定義することが困難である。そのため、属性やシーン全体のコンテキストにより、認識対象をモデル化す

る方法が考えられている。コンテキストを用いた認識では、対象に関する知識だけでなく、周囲の幅広い知識が必要となる。このため、認識を行なうためには、多くのコンテキストが必要であり、その応用においては、認識時の探索空間の増大化や知識の妥当性の検証などへの十分な配慮が必要となる。

コンテキストは、脈絡、つまり、対象自身の属性や周囲との相互関係を記述するもので、プロダクションルールのような形式で記述される。コンテキストで表現できる情報としては、次のようなものが考えられる。

- 対象自身を表す情報

  これは、対象の候補を取り出すためのもので、対象の形状や色などの属性情報のことである。

- 対象と周囲の関係を表す情報

  これは、対象の候補から可能性の高いものを取り出すためのもので、周囲と対象の位置関係などの関係情報のことである。

コンテキストモデルを用いた認識システムには、T.M.Strat らによる Condor システム [27] などがある。詳しくは次節で紹介する。

動画像構造のモデル化

動画像構造のモデル化とは、画像のどの位置にどのようなオブジェクトが映っているのか、また、そのオブジェクトはどのような動きをしているのかなどといった、オブジェクト個々の情報やレイアウトの情報をモデル化しようとするものである。このようなモデル化は、画像中の個々のオブジェクトの認識が難しいことから、色や形状情報に注目して、複数の領域の位置関係や動きにより、記述される。

現在、記述要素としては、位置、形、大きさ、色、位置の変化、形の変化、大きさの変化、オブジェクトの数、オブジェクトの分布、数の変化、分布の変化などが考えられている [60]。

動画像構造モデルを用いた認識システムには、Y.Gong らによる映像シーン記述言語 VSDL を用いたシステム [59] などがある。

2.5 認識システムの実例

2.5.1 幾何学的形状モデルを用いた認識システム

ここでは、R.A.Brooksにより開発されたACRONYMシステム[23]を取り上げる。

このシステムでは、幾何学的形状を用いて、画像中から物体を認識することができる。認識対象のモデルは、三次元の幾何学的モデルで表現され、三次元の形状は、スピン、断面、スウィーピングルールにより定義される一般化円筒を用いて階層的に記述される。ここで、スピンは軸を、スィーピングルールは断面の形状変化を示すものである。各モデルは、全体-部分関係を表す物体グラフと、上位-下位関係を表す関係グラフより記述される。また、各部のパラメータは、各関係グラフにおいて、制約の伝搬が行なわれることにより、決定される。

このシステムの認識プロセスを次に示す。

1. 物体グラフと関係グラフから、見え方の予測を行ない、予測グラフを生成する。
2. 画像のエッジ情報から、一般化円筒の二次元への投影であるリボン記述を生成し、記述グラフを生成する。
3. 記述グラフに予測グラフを満たすものがあるかどうか調べる。

このシステムは、空港を上空から撮影した画像からの航空機の認識に適用された。このときの、一般化円筒により記述された航空機のモデルを図2.5に、認識結果を図2.6に示す。

2.5.2 機能モデルを用いた認識システム

ここでは、L.Starkらにより開発されたGRUFF2システム[25]を取り上げる。

このシステムでは、機能的特徴を用いて、物体の属するカテゴリーを決定することができる。カテゴリーの機能的特徴は、階層的なグラフでモデル化され、機能自体は機能プランにより定義される。機能プランは、物体の形状を質的に評価する知識要素の組み合わせにより記述され、機能は、三次元情報を元に評価される。

このシステムでは、形状を質的に評価する知識要素として、相対的配置、寸法、近傍、クリアランス、安定性の五つが考えられている。

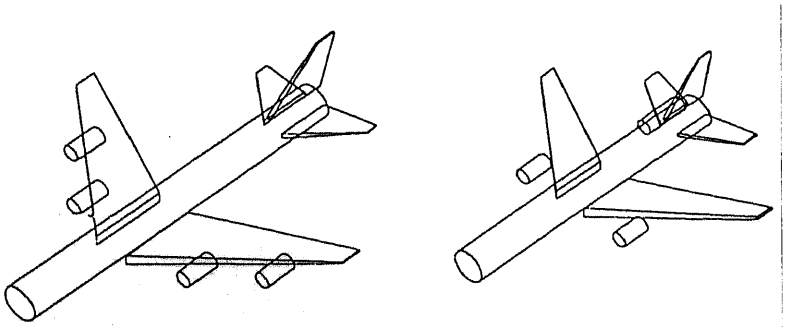


図 2.5: 一般化円筒による航空機のモデル

- 形状を質的に評価する知識要素

  1. 相対的配置 (relative orientation)

     面の法線ベクトルを用いて、二つの面の配置が条件を満足するかどうかを調べる。

  2. 寸法 (dimensions)

     可能性のある機能要素の位置、大きさが条件を満足するかどうかを調べる。

  3. 近傍 (proximity)

     二つの面の距離が条件を満足するかどうかを調べる。

  4. クリアランス (clearance)

     機能を生かすために必要な空間が存在するかどうかを調べる。

  5. 安定性 (stability)

     それがとるべき姿勢において、安定した支えとなりえるかどうかを調べる。

カテゴリー「ベンチ」のモデルを図 2.7に示す。モデルは、機能的記述で定義され、特定の幾何学的、構造的記述は必要としない。グラフの各ノードは、名前、タイプ、実現、機能プランという四つのフィールドを持つフレームで表され、実現フィールドは、機能を実現するために満足されるべき知識要素のリスト、機能プランフィールドは、そのノードに定義されたサブカテゴリーへのアークを持つ。

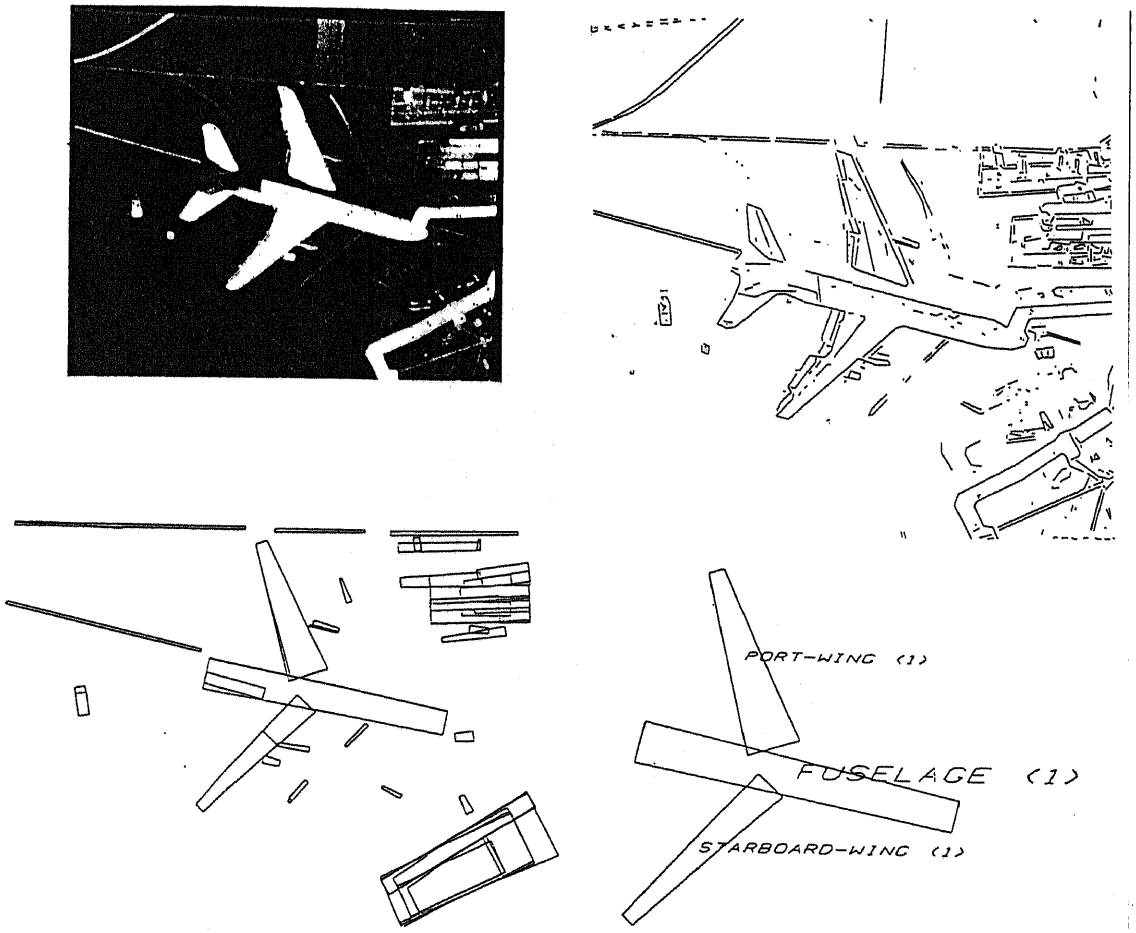


図 2.6: 認識結果

このシステムは、以下の手順で認識を実行する。

1. 機能要素と考えられる部分を抽出する。例えば、物体の単一面や構造の三次元モジュールなどの部分である。

2. 物体の形状を質的に評価する知識要素を用いて，各部の機能を判定し、属するカテゴリーの決定を行う。

認識結果として、カテゴリー「ベンチ」と認識されたものの一部を図 2.8に示す。

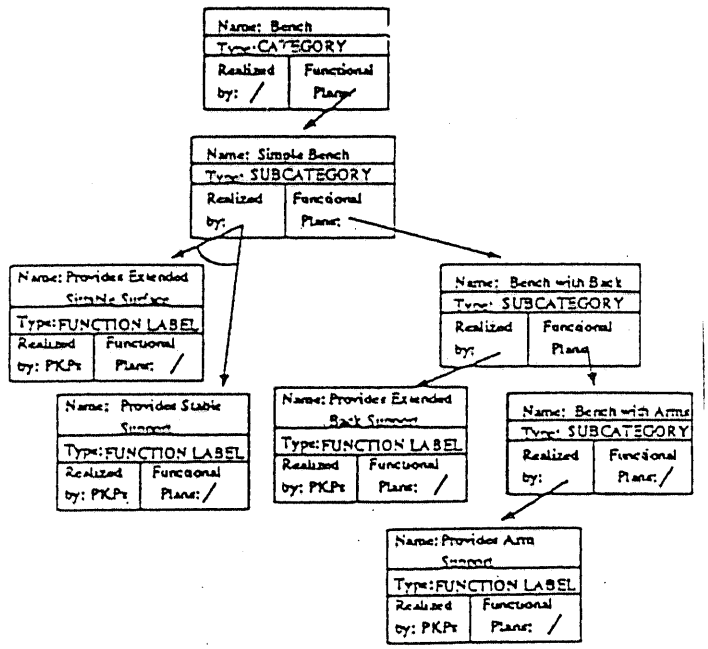


図 2.7: 「ベンチ」のモデル

2.5.3 コンテキストモデルを用いた認識システム

ここでは、T.M.Strat らにより開発された Condor システム [27] を取り上げる。

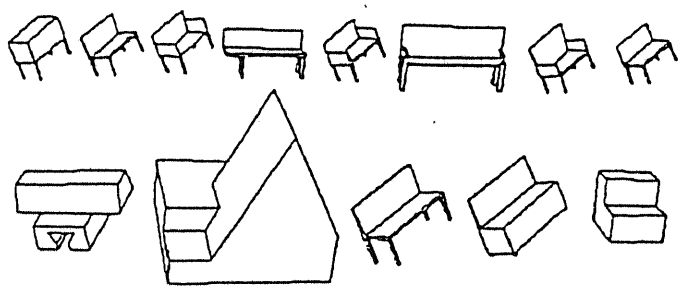


図 2.8: 認識結果

このシステムは、コンテキスト情報を用いて、自然の風景を認識することを目的に開発された。このシステムでは、コンテキストを次のように記述する。

$L : \{CE_1, CE_2, \cdots, CE_n\} \Rightarrow A$

ここで、L は対象とするオブジェクトのラベル、CE_iはコンテキストの概念を定義するコンテキスト要素、A は行動を表す。各コンテキストは、すべてのコンテキスト要素の条件が満足されたときに、行動を起動する。

このシステムでは、コンテキストとして、以下の三つの type のものを設定している。システムは、これらのコンテキストにより、候補の生成、評価、そして、一貫性の評価という形で認識を実行する。

1. Type 1 : 対象の候補の生成を行うコンテキスト

   例 $SKY : \{CLIQUE - IS - EMPTY\} \Rightarrow$

   $BRIGHT - REGIONS$

2. Type 2 : 対象の候補の評価を行うコンテキスト

   例 $SKY : \{ALWAYS\} \Rightarrow$

   $ABOVE - HORIZON$

3. Type 3 : シーン全体での一貫性の評価を行うコンテキスト

例 $SKY : \{GEOMETRIC - HORIZON - KNOWN\} \Rightarrow$

$PARTIALLY - BELLOW - GEOMETRIC - HORIZON$

自然の風景の画像の例を図 2.9に、自然の風景の認識結果を図 2.10に示す。

図 2.9: 自然の風景の画像

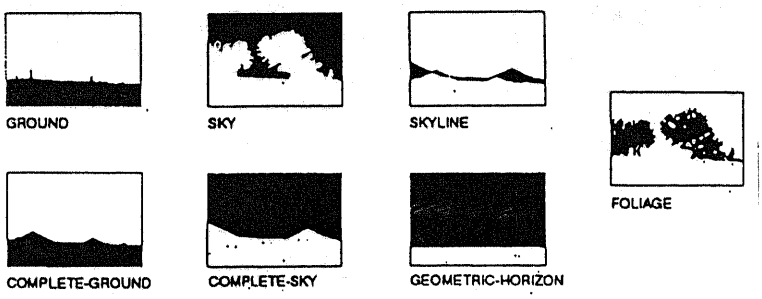


図 2.10: 認識結果

2.5.4 動画像構造モデルを用いた認識システム

ここでは、Y.Gong らにより提案された映像シーン記述言語 VSDL を用いたシステム [59] を取り上げる。

VSDL は、ユーザ定義により、所望のシーンにアクセスするための言語として提案されたものである。VSDL では、領域の色とその動きの情報を記述することで、多くの映像のなかから、ユーザが必要としているシーンを検索できる。

VSDL は以下に示すように階層的に記述を行なう。

1. 色定義階層: Color を用い、色を定義する。
2. 領域定義階層: Segment を用い、色領域を定義する。

   定義の中では、面積、位置、動きを指定できる。
3. シーンクラスの定義: Class を用い、シーンクラスを定義する。

   定義の中で、色領域の出現、色分布の変化、カメラワークを記述できる。

相撲シーンの検出例を図 2.11に示す。

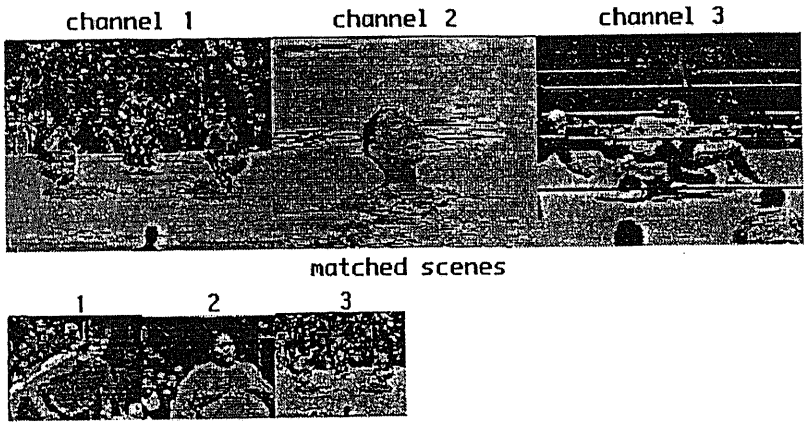


図 2.11: 相撲シーンの検出結果

2.6 画像検索技術

2.6.1 画像認識としての画像検索技術

画像認識技術と深く関連する技術に画像検索技術がある。画像検索のうち、内容検索といわれるものは、利用者の要求など検索意図に対し、画像を何らかの形で評価することで実現され、この評価は、基本的に何らかの画像認識による情報抽出により行なわれる。このため、画像検索技術は画像認識技術の上に成り立っているものといえる。一方、画像検索を画像認識の枠組の中で考えてみると、検索の実行により、検索画像から検索意図に基づく情報が抽出、認識されたと捉えることができる。これは、キーワードと画像、画像と画像など、検索意図と画像を結びつけ、画像の意味づけを行なっていることにほかならない。ここでは、検索意図と画像の結びつけという観点から画像検索技術について述べる。

2.6.2 画像検索における検索要求と索引情報

画像検索技術では、検索要求の与え方と索引情報の選択が重要である。検索要求とは、どのような観点で検索を行ないたいのかなど検索意図をあらわすもので、一般に、画像の意味をあらわすキーワード、部分パターンをあらわす画像特徴、全体パターンをあらわす画像そのものがよく利用される。画像内容に従った検索は内容検索といわれ、特に、キーワードを利用するものはキーワード検索、画像を利用するものは例示画像検索などといわれる。一方、索引情報とは、画像検索を行なう際に評価に用いる情報のことであり、情報の抽象度で考えると、画像特徴レベル、画像内容レベルに分けることができる。表2.3に代表的な研究例を示す。

表 2.3: 画像検索技術の分類

検索要求	索引情報	
	画像特徴レベル	画像内容レベル
キーワード	感性キーワード [47]	キーワード自動獲得 [42]
画像特徴	QBIC[58]	
画像	TRADEMARK [49]	連想検索 [54]

2.6.3 検索要求からみた画像検索

検索要求をあらわすものとしては、キーワード、画像特徴、画像など利用される。ここでは、それらの具体的システムの紹介を行なう。

キーワードによる検索

キーワードによる検索とは、与えられたキーワードと、データベース内の画像の何らかの情報を結びつけることで行なわれる検索のことである。

これには、画像に対しキーワードを予め付与しておくものや、画像から意味レベルのキーワードを自動獲得するもの、キーワードと画像の特徴量を関係づけておくものなどが考えられている。

キーワードと特徴量を関連づけを利用したものとしては、栗田らによる印象語による検索を行なう絵画データベース [47] がある。これは、印象語と画像から求められる特徴量を正準相関分析モデルを用いて関連づけを行なうものである。特徴量としては、カラー画像の RGB 値に対し、第 0 次と第 1 次の局所自己相関特徴を用いている。また、印象語としては、30 個の形容詞を用いている。

画像から意味レベルのキーワードを自動獲得するものとしては、山根らによるスポーツ画像を対象とした検索システム [42] がある。これは、画像認識技術の現状を踏まえ、完全な認識ができなくても、認識ができたところまでの認識結果を有効に活用しようとするものである。これは、本論文第 3 章で述べる状態遷移モデルの枠組を利用して実現されている。

画像特徴による検索

画像特徴による検索とは、画像特徴の範囲などを指定し、その条件を満足する画像を、データベースから検索するもののことである。

これには、画像特徴の範囲を指定するもの、画像のレイアウトを指定するものなどが考えられている。

画像特徴の範囲やレイアウトなどを指定による検索を行なうものとしては、Flickner らによる QBIC システム [58] がある。このシステムでは、グラフィカルユーザインターフェースを用いて、物体の色や形状、配置、そして、輪郭を描くことなどにより、検索要求を与えることができる。このシステムの特徴は、機械が得意とする特徴量の計算などは自動化

し、それ以上の内容に関する部分は人間の判断にまかせる点にある。

画像による検索

画像による検索とは、画像がもつ何らかの情報を用いて、データベース内の画像を検索するもののことである。

これには、画像そのものから自動的に抽出される特徴量を用いるもの、画像に予めキーワードを付与しておくものなどが考えられている。

画像そのものから自動的に抽出される特徴量を用いるものとしては、加藤らによる商標・意匠データベース TRADEMARK[49] がある。このシステムでは、検索要求は概略画像を用い指定する。このシステムでは、画像を階層的にメッシュに分割し、それら各メッシュを基本に、濃淡や概略形状、周波数特徴などをあらわす特徴を計算し利用している。

画像に予めキーワードを付与しておくものとしては、柴田らによる連想検索のシステム[54] がある。このシステムでは、画像に対し、予めキーワードを付与し、このキーワードから画像間の距離を算出することで、連想検索を行なっている。キーワードを用いた距離の計算は、キーワードをベクトルで表現することで実現している。

2.7 本研究の位置付け

前節までに、認識対象のモデルの表現に焦点をあて、画像認識技術の現状を述べた。マルチメディア環境における画像認識技術という観点からは、利用者の多様な要求に対応できること、扱える対象世界が広範であることが重要となり、テレビ映像などへ柔軟な対応が望まれる。しかし、従来までの画像認識技術は、産業応用を中心として研究、開発され、テレビ映像などのより一般的な画像の認識を意識したものは少なかった。一般的な映像を扱う場合には、その映像ゆえの新たに解決しなければならない問題や課題が生じる。本研究は、このような問題に対する一アプローチを検討するものである。本研究では、以下に示す三つの観点による課題を取り上げた。

- 一般的な映像では、対象の特性が限定できないため、多目的に利用可能な認識モデルが望まれている。

- 一般的な映像では、オブジェクトの認識とは別にシーンレベルでの認識の実現が望まれている。
- 一般的な映像では、撮像距離などにより対象の映る大きさが異なるため、それらの違いを統一的に扱えるモデル表現が望まれている。

2.8 まとめ

本章では、画像認識の背景技術について、知識の表現、特に、目的や対象に応じたモデル化という観点から整理し、いくつかの特徴的なモデル化手法と画像認識手法について述べ、最後に、本研究の位置付けを明確にした。

画像認識を行なうためには、対象とする画像の特徴、獲得すべき情報の種類などにより、様々な知識が必要となる。ここでは、認識対象のモデル化を中心に、これらの知識の表現手法について概説した。ここで述べたモデル化手法は、それぞれ大きな特徴を持っており、シーンから物体の認識、動画像にいたるまで、対応が可能である。しかし、一般的な映像を対象とする場合には、同じ対象であっても、対象の映り方が様々に変化したり、また、画像全体の意味を知りたい場合も少なくないなど、従来の方法では対応しがたい課題も存在する。

これに対し、本研究では、いくつかの具体的課題を取り上げ、それら目的や対象に応じたモデル化手法、認識手法の検討を行なった。以降、各章では、先に述べた具体的な課題に対してのモデル化手法、画像認識手法の提案を行い、検討を行なう。

第 3 章

多目的な認識を可能とする状態遷移モデルとルール作成支援システム

3.1 概要

従来の認識システムは、認識対象に依存した形で構築される場合が多く、新たな認識対象を処理するためには、システムを構築し直す必要があった。このため、認識対象の変更に柔軟に対応することが困難であった。また、マルチメディア環境をにらんだ画像認識の実現という観点からは、より多目的に利用できる認識システムの形態が望まれる。これに対し、推論エンジンの汎用化をめざしたプロダクションシステムや、知識の統一的表現を可能とするフレームを用いたシステムなど [28][29] が試みられているが、対象に依存した部分の記述がそれほど容易ではなかったり、認識戦略を表現することが難しいなどの問題があった。

これまでに、これらの問題を解決すべく、多目的型認識システムの実現を目指して、われわれの研究室では、状態遷移モデルを提案し、それを用いた認識システム [14]-[16] を開発してきた。この状態遷移モデルを用いた認識システムでは、認識基本要素 (図面では線分、画像ではセグメントなど) を認識のステップを表す「状態」という形で管理し、それら認識基本要素の状態を次々に遷移させることで認識を実現する。状態の遷移は、状態遷移ルールという形で記述でき、高度な認識を実現する上で重要なボトムアップ解析とトップダウン解析の実行が可能となっている。また、状態遷移ルールは作成が容易にできるように考慮されており、認識対象に依存した部分の状態遷移ルールを作成、変更することで、新たな認識対象へも柔軟に対応できるようになっている。

ところで、開発された状態遷移モデルを用いた認識システムでは、状態遷移のためのルールが必要である。この状態遷移のためのルールの作成は、人手により行う必要があり、枠組については容易に作成できるが、条件判断部のしきい値など微調整が必要な部分は試行錯誤で調整を行う必要があり、手間のかかる困難な作業となる場合もあった。このため、この問題を解決するため、グラフィカルユーザインターフェース (GUI) を用いた、ユーザとの対話によりルール作成を可能とするルール作成支援システム [1][3][13] を構築し、システムの拡張を行った。このルール作成支援システムでは、効率よくルールを作成するため、帰納推論による学習アルゴリズムを用いている。

本章では、はじめに、状態遷移モデルを用いた認識システムについて述べ、その後、状態遷移ルール作成のために開発したルール作成支援システムについて説明する。

3.2 状態遷移モデルを用いた認識システム

3.2.1 状態遷移モデル

状態遷移モデルとは、多目的な認識システムの開発を目指して提案された認識モデルのことである。この状態遷移モデルでは、認識ステップを「状態」、認識プロセスを「状態遷移」という形でモデル化する。状態遷移モデルによる画像認識の枠組を図3.1に示す。この枠組では、状態と状態遷移の形態を限定していない。このため、状態と状態遷移の方法の定義の仕方により、様々な目的、様々な対象に対して利用することが可能である。後で述べる状態遷移を用いた認識システムでは、これら状態と状態遷移の方法の定義を、ルールという形で記述することができるようになっている。このため、ルールの記述を変更することで、様々な対象を扱うことが可能である。

状態遷移モデルの例として、図面から三角形を認識する場合の状態遷移ダイアグラムを図3.2に示す。

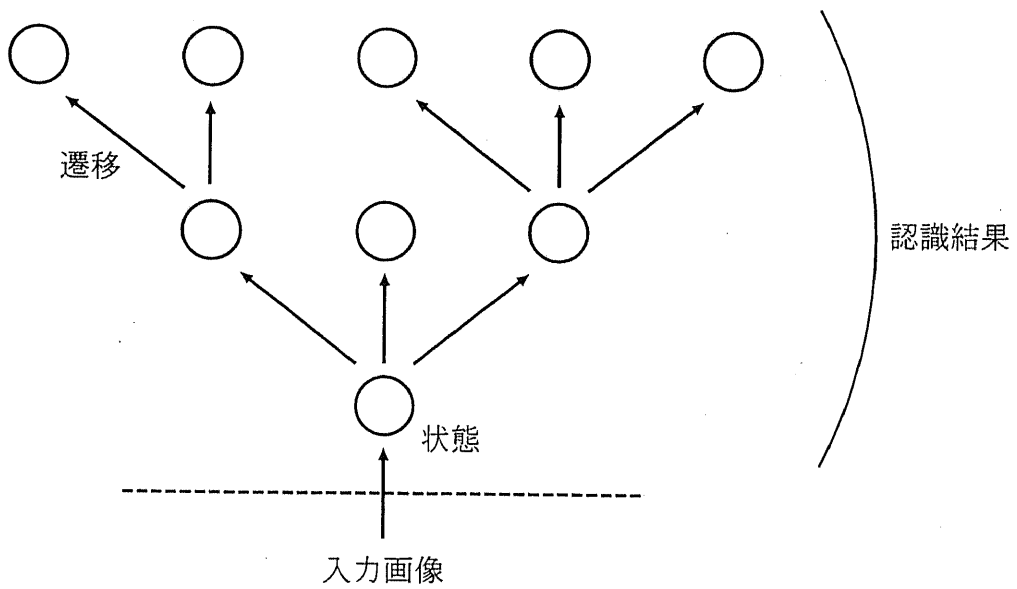


図 3.1: 状態遷移モデルによる画像認識の枠組

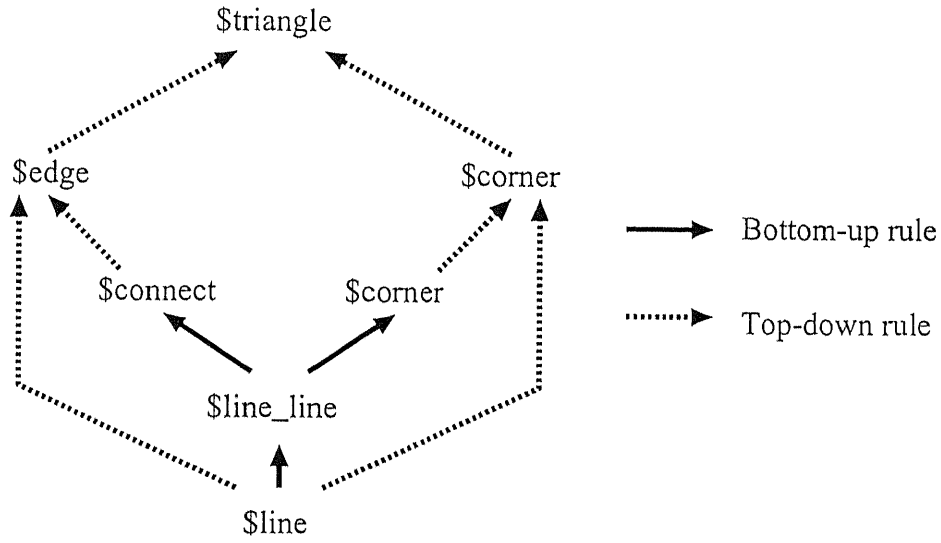


図 3.2: 三角形抽出のための状態遷移ダイアグラム

3.2.2 認識システム

状態遷移モデルを用いた認識システムでは、認識基本要素 (画像を構成する基本的な要素) を「トークン」と呼び、画像認識をこれらのトークンへの適切なラベルづけで実現する。このラベルは各トークンの持つ「内部状態」として表現され、画像認識の過程はトークンの内部状態の状態遷移によって実現される。すなわち、トークンは、画像を構成する基本要素そのものを実体として持ち、認識プロセスにおいて、その内部状態を自律的に状態遷移させることができる。このため、画像認識の最終結果はトークンの最終状態としてあらわされることになる。ここで、トークンの状態遷移は、画像認識の基本的戦略である、データ駆動のボトムアップ解析とモデル駆動のトップダウン解析を組み合わせて実行できるようになっている。

状態遷移モデルを用いた認識システムでは、システムを認識対象に依存する部分と依存しない部分に分けて構成することで、新たな認識対象を処理するために構築し直す部分を最小限に押えられるようになっている。対象に依存する部分は、状態遷移ルールという抽象度の高い形で記述できるようになっており、認識対象の変更に柔軟に対応することが可能である。状態遷移ルールは、基本的なボトムアップ解析を行なうボトムアップルールと、

対象のモデルをあらわすトップダウンルールの二種類のルールから構成される。対象に依存しない部分は、画像を構成するトークンの状態遷移を実現する部分であり、基本的には認識対象によらず、様々な対象に利用可能である。本システムでは、トークンの状態遷移を実現する部分を認識カーネルと呼ぶ。

状態遷移モデルを用いた認識システムの枠組を図 3.3に示す。

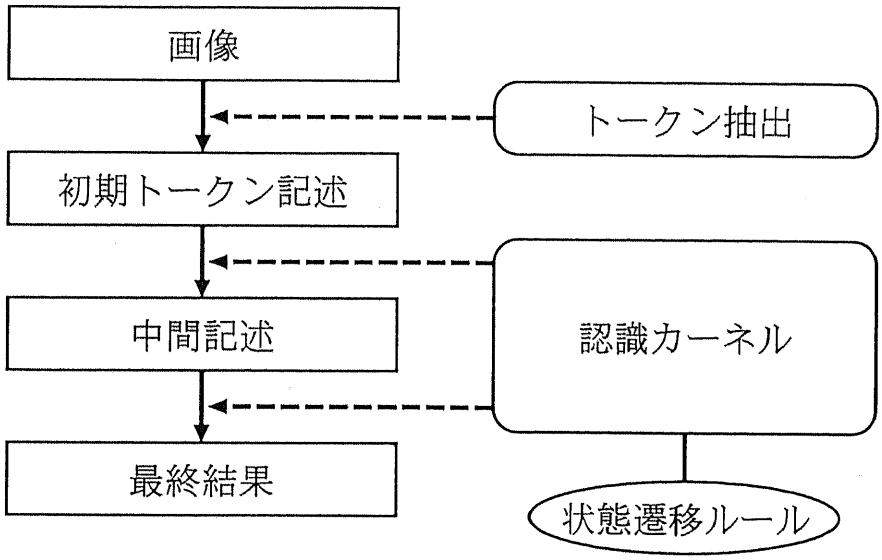


図 3.3: 状態遷移モデルを用いた認識システムの枠組

3.2.3 認識プロセス

はじめに、画像データを「トークンエクストラクタ」により処理し、画像を構成する基本的な要素からなる初期トークン記述に変換する。ついで、各トークンは「認識カーネル」の働きにより、状態遷移ルールに従って、他のトークンとは独立に周囲の状況を調べながら状態遷移を行なう。

認識カーネルの動作モードには、ボトムアッププロセスモードとトップダウンプロセスモードの二つがあり、それぞれが状態遷移を実現する。ボトムアッププロセスモードでは、各トークンの状態遷移を確定的に実行し、トップダウンプロセスモードでは、ボトムアッププロセスによる結果を用いて、非確定的な認識を実行する。これらは、意図する記述レ

ベルが得られるか、または、認識プロセスが進まなくなるまで繰り返される。

認識実行時、認識カーネルは、はじめにボトムアッププロセスのモードとなり、ボトムアップルールに基づいて処理を実行する。次に、各トークンの状態遷移がボトムアップ的に処理できなくなった時点で、トップダウンのモードとなり、トップダウンルールに基づいた処理を行なうよう動作する。通常、ボトムアッププロセスでは、トークンの面積などの特徴量を用いた検査、トークン間の近接性の検査など局所的な検査を行い、トップダウンプロセスでは、モデルとの大局的な照合を行なう。

- ボトムアッププロセス

  ボトムアッププロセスでは、トークンの自律動作をシミュレートして状態遷移を行わせ、トークンの局所的な周囲状況の調査のみでその意味的解釈が確定的に決定できる場合の意味づけを行う。

  ボトムアップルールは式 (3.1) のように記述される。トークンの現在の状態をヘッド部分に持ち、満たすべき制約条件、遷移すべき次の状態をボディ部に持つ。

$$cur\_state \longrightarrow cond_1, cond_2, \ldots, cond_n, next\_state\_pred. \tag{3.1}$$

  ボトムアッププロセスは以下の手順で実行される。

  1. すべてのトークンから任意のトークンを選択し、そのトークンの状態に対応するボトムアップルールを検索する。
  2. 検索されたルールの条件が満足されるかどうかテストし、満足される場合には状態遷移を実行し、満足されない場合には別のルールを検索する。
  3. すべてのルールで条件が満足されない場合には、他のトークンを選択し、ルール選択、条件テストを繰り返す。
  4. すべてのトークンで状態遷移が行えなくなるまで、以上の手順を繰り返す。ただし、一度停止したトークンも再度試みる。

通常、条件を満足するルールが複数存在する場合、効率の悪いバックトラックを行わせる必要があるが、ここでは、確定的に行える状態遷移のみを行なわせることで、この問題を回避している。

- トップダウンプロセス

  ボトムアッププロセスが終了すると、次にトップダウンプロセスを実行する。

  トップダウンプロセスでは、ボトムアッププロセスで生成された状態を用いて、与えられた認識対象のモデルとの大局的な照合を行う。ここでは、ボトムアッププロセスと異なり、バックトラックを行なわせることで、非確定的な認識を実現する。

  トップダウンルールは式 (3.2) のように記述される。ゴールモデルをヘッド部分に持ち、満たすべき制約条件、簡易化した下位モデルをボディ部に持つ。

$$goal \longleftarrow$$
$$cond_1, cond_2, \ldots, cond_n, subgoal_1, subgoal_2, \ldots, subgoal_n. \qquad (3.2)$$

  トップダウンプロセスは以下の手順で実行される。

  1. 認識すべき対象とマッチするヘッド部を持つルールを検索する。
  2. 検索されたルールの条件が満足されるかテストし、満足される場合にはそのモデルを構成するいくつかの下位モデルを再帰的に呼び出す。条件が満足されない場合には別の解釈の可能性が残っている部分へバックトラックし、照合を再開する。

3.2.4 地図図面の認識

ここでは、状態遷移モデルによる図面の認識について説明する。ここでの対象は図 3.4 に示すような地図図面とする。図面は、スキャナによりビットマップイメージデータとして、システムへ入力される。そして、トークンエクストラクタにより、シンボルレベルの記述である初期トークン記述へ変換する。ここで、トークンエクストラクタとして、当研

究室で開発されたソフトウェアベースの高速図面イメージプロセッサ AI-Mudams[38] を利用した。

ビットマップイメージデータは、AI-Mudams により、図形の輪郭線を折れ線近似した線分ベクトル群へと変換される。そして、それらを孤立した図形を表すベクトル群と、比較的長い線的な図形を表すベクトル群に分け、初期トークン記述とする。この初期トークン記述から認識を行うための状態遷移ルールの例を図 3.5に示す。これは、植生界$\$lub$ の認識のためのルールの例である。植生界とは、「畑」や「田」といった植生の境界をあらわす閉領域で、地図の上では小さな点の列、道路などにより構成されるものである。

この状態遷移ルールでは次のような認識を実行する。ボトムアッププロセスでは、各トークンの面積を計算し、面積がある値の範囲にあるトークンを$\$dot$ トークンとして識別する。そして、$\$dot$ トークンからある距離以内にある$\$dot$ トークンを近接しているとみなし、近接している$\$dot$ トークンがただ二つであるものを点列を構成するトークンとして抽出する。ボトムアッププロセスで判定する植生界の構成の様子を図 3.6に示す。トップダウンプロセスでは、ボトムアッププロセスで抽出された$\$dot$ トークンをもとに、その連結性の評価から、$\$lub$ になるものを探索する。

この状態遷移ルールによる認識結果を図 3.7に示す。

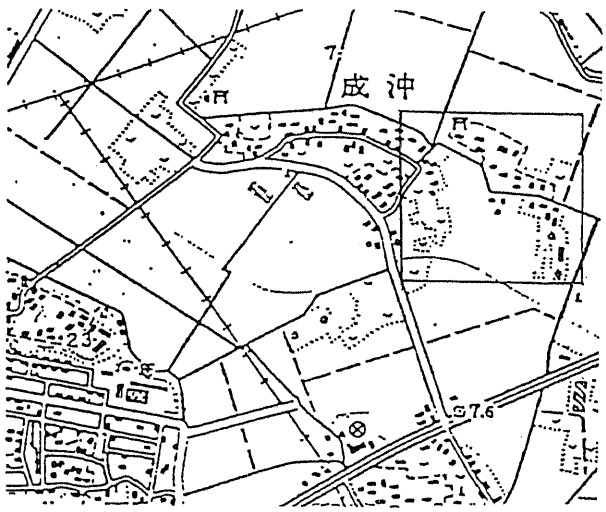


図 3.4: 地図図面の例 (破線内は図 3.7 に対応)

```
$loop ⟶
    $area(A),
    $transform($loop(A)).
$loop(A) ⟶
    $((A ≥ 0, A < 4000)),
    $transform($dot).
$dot ⟶
    $get_near_loop(120, A),
    $select(A, $dot, B),
    $transform($dot(B)).
$dot([A, B]) ⟶
    $get_self(S).
    $transform($dot(S, A, B)).
```

(a) ボトムアップルール

```
$lub ⟵
    $token($dot(B, A, C)),
    $lub($dot(B, A, C), B, C, [B], B, C).
$lub(_, L, _, _, _, L).
$lub(T, C, P, M, B, L) ⟵
    $linked(T, C, P, N, NT),
    $(notmember(N, M)),
    $lub(NT, N, C, [N|M], B, L).
$linked($dot(C, N, P), C, P, N, NT) ⟵
    $not_eq(N, P),
    $token_at(N, NT),
    NT = $dot(_, _, _).
```

(b) トップダウンルール

図 3.5: 状態遷移ルールの例

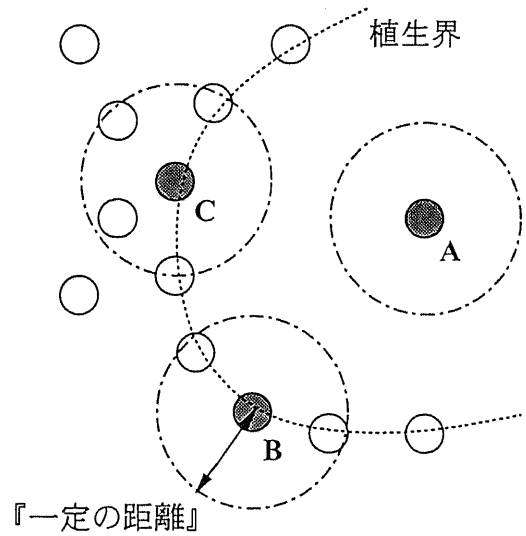


図 3.6: 植生界の構成

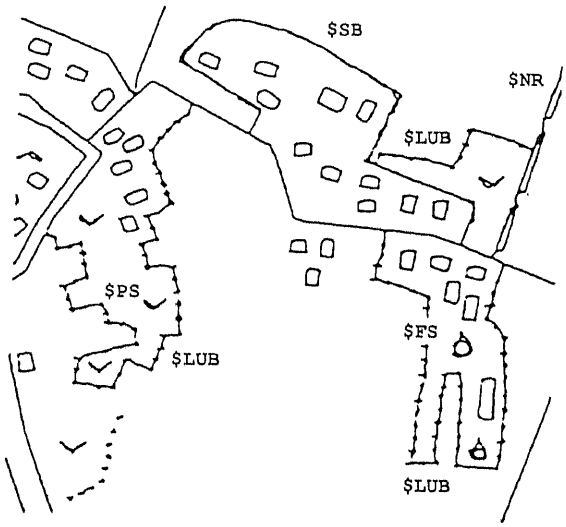


図 3.7: 認識結果

3.3 人間機械協調型ルール作成支援システム

3.3.1 システムの概略

前節までに述べたように、状態遷移モデルを用いた認識システムでは、認識知識を状態遷移ルールにより記述する。しかし、このルール作成において、制約条件における閾値など微調整を必要とする部分の作成が困難であるとの問題があった。そこで、このような問題を緩和するため、人間機械協調型のルール作成支援システムの作成を行なった。このシステムの枠組を図 3.8に示す。このシステムは、状態遷移モデルを用いた認識システム部とルール学習システム部から構成される。

このルール作成支援システムでは、GUI を用いたユーザとの平易な対話からルールを生成することができる。状態遷移ルールとしては、ボトムアップルールとトップダウンルールの二つのルールが必要だが、このシステムではボトムアップルールのみを扱うようになっており、閾値などの微調整が必要な部分を含むルールの自動獲得を行なうものである。

このシステムは、帰納学習により、ユーザからの一連の指示から一つのルールを作成す

るように動作する。このルールは、既存のルールによりある状態に至っているトークンに対し、所定の条件判定を行ない、ある状態に遷移し得るかを判定するものである。条件判定を行なうために必要なパラメータの測定のためのルールは、ユーザが予め作成しておく必要がある。

このシステムのルール学習の大まかな手順は次のとおりである。はじめに、システムは、認識システムの認識結果をユーザに提示し、その妥当性をうかがう。もし、その認識結果の中に誤った認識結果が含まれていれば、ユーザはシステムに対して誤っている箇所を指摘する。これに対し、システムはルールに変更を加えて、認識システムの認識結果がユーザの指摘に応えるものになるようにする。その際、ユーザとのセッションは事例データベースに蓄えておき、同じ問い合わせを繰り返すことを回避する。このようにして、システムはユーザとのセッションを繰り返し、ルールの修正、作成を行って、妥当な認識結果を得るようにする。

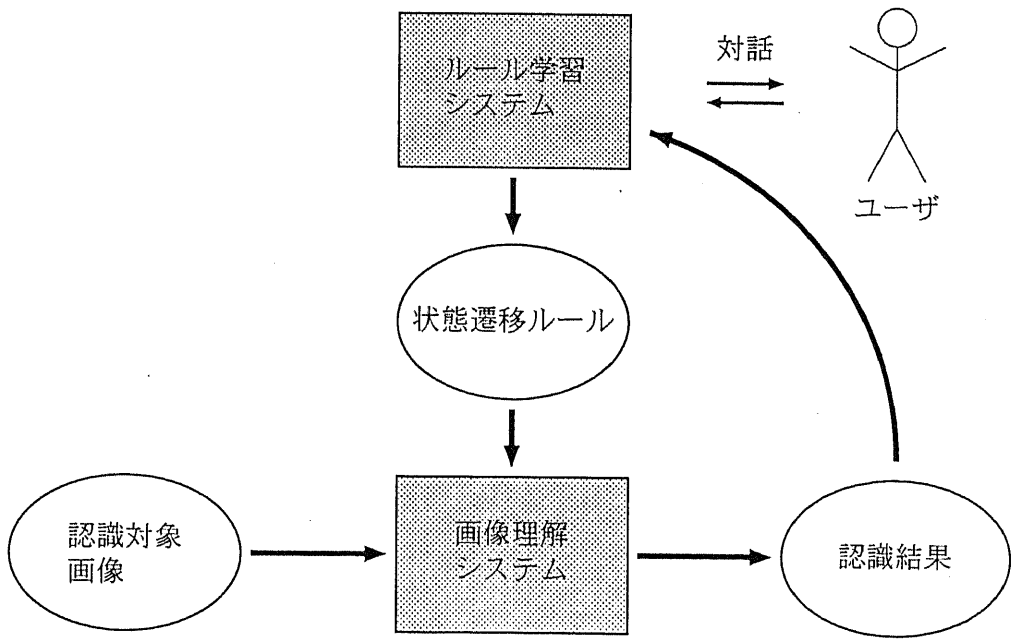


図 3.8: ルール作成支援システムの枠組

3.3.2 学習対象とするボトムアップルール

地図図面の認識のところで示したボトムアップルールからも明らかなように、状態遷移モデルにおけるボトムアッププロセスは、図3.9に示すように、パラメータ計測→状況判断→状態遷移の繰り返しで実現されている。これは、以下のように記述できる。

$$
\begin{aligned}
&\mathcal{S}_1(\pi_{in_1},\ldots,\pi_{in_n}) \longrightarrow \\
&\quad \mathcal{P}(\pi_{in_1},\ldots,\pi_{in_n},\pi_{out_1},\ldots,\pi_{out_m}), \\
&\quad \$\mathtt{transform}(\mathcal{S}_{11}(\pi_{in_1},\ldots,\pi_{in_n},\pi_{out_1},\ldots,\pi_{out_m})). \qquad (3.3) \\
&\mathcal{S}_{11}(\pi_1,\ldots,\pi_{n'}) \longrightarrow \\
&\quad \mathcal{C}_{111}(\pi_1,\ldots,\pi_{n'}), \\
&\quad \$\mathtt{transform}(\mathcal{S}_{111}(\pi_1,\ldots,\pi_{n'})). \qquad (3.4)
\end{aligned}
$$

ここで、$\mathcal{P}(\pi_{in_1},\ldots,\pi_{in_n},\pi_{out_1},\ldots,\pi_{out_m})$ は、パラメータの計測を行なう部分であり、$\mathcal{C}(\pi_1,\ldots,\pi_n)$ は、パラメータ$\pi_1,\ldots,\pi_n$が条件を満たしているかどうかを判断する部分である。また、$\mathtt{transform}(\mathcal{S})$ は、状態 $\mathcal{S}$に遷移させる部分である。したがって、式 (3.3) は、パラメータ計測のためのルール、式 (3.4) は、計測したパラメータによる状況判断を行なわせるルールとなる。

このようなボトムアップルールを作成する場合、状態遷移の仕方であるルールの骨組みは簡単に決定できるが、そのルールの調整に手間取ることが多い。ここで、骨組みとはその繰り返しのうち状況判断を除いた部分、調整とは状況判断の部分の調整のことである。そこで、今回のシステムは、この骨組みはユーザが与えるものとして、状況判断の部分 $\mathcal{C}(\pi_1,\ldots,\pi_n)$ の調整をユーザとの対話により実現するものとした。

3.3.3 ルールの学習

ルールの学習は以下の手順で行われる。ユーザからシステムへの指示は、＜S,V＞という対により行う。S はあるトークンの現状態を表し、V はそのトークンが想定した状態になるかならないかにより、true か false の値をとる。ここで、true を正の例、false を負の例と呼ぶ。対話は、＜S,V＞という形式の指示で行なわれるが、本システムでは GUI を

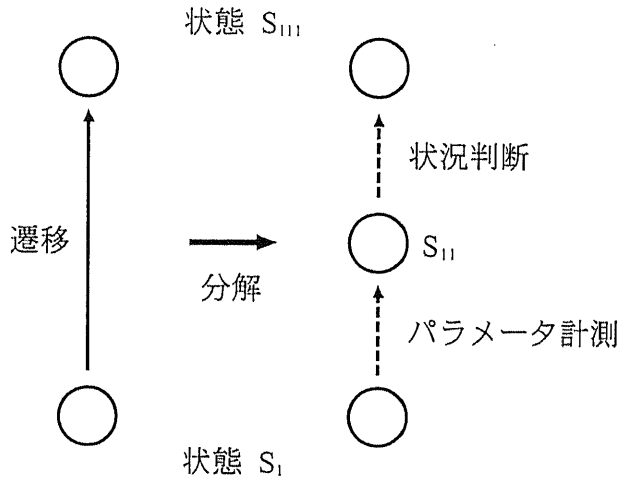


図 3.9: ボトムアッププロセスの分解

用いて、学習によるトークンの識別結果を確認しながら、マウスにより、指示するためのトークンを選ぶことができる。

1. 認識システムにより、与えられた画像のトークンを、学習に必要なパラメータを含む状態に遷移させる。

2. ユーザは学習に必要なトークンを指摘する。

3. 現在までの指示と矛盾なく、今の指示を実現するルールが存在するなら、そのルールを生成する。ルールが生成できない場合には、状況判断のために必要な新たなパラメータをユーザに求める。

4. 生成したルールによりトークンを分類した結果をユーザに提示する。

5. 結果がユーザの想定とあっている場合、生成したルールを確定して終了する。

6. 結果が想定と異なっている場合、生成したルールを無効にし、2 より繰り返す。

ルールの学習アルゴリズムとしては、数値とリスト構造を扱うことができる帰納的推論アルゴリズムを用いた。これは、Shapiro のモデル推論システム [37] を基ににし、条件と

しての数値の大小比較を扱えるように拡張したものである。基本的な動作は、与えられた状態のパラメータの型に適合する、一連の生成可能な条件つきルールを生成し、与えられた例をこのルールに適用することによって、すべての例に矛盾しないルールを求めるというものである。ルールの生成の例を図3.10に示す。

ルールの学習の流れは以下のとおりである。ここで、システムに与えられた正の例を$T_1, \cdots, T_n$、負の例を$F_1, \cdots, F_m$とする。

1. 任意の正の例T_iを選び、着目している正の例Tとする。
2. Tを満たし、すべてのF_jを満たさないルールのうち、もっとも一般的なものを生成し、Rとする。
3. RですべてのT_iが満たされるか調べ、満足されれば求めるルールとして出力する。
4. 満足されない場合、Rを無効にし、これまでに着目している正の例になっていないT_iをTとして2から繰り返す。
5. すべてのT_iを着目している正の例として試行した結果、満足されるルールが生成できなかった場合には、ルールの学習の失敗をユーザに告げて終了する。

基となるモデル推論システムでは、ルールの選言を扱うことが可能であるが、本システムの学習アルゴリズムでは扱えない。このため、ある状態のトークンからあらゆる状況を認識するためにシステムが生成できるルールは、その状態の各パラメータから生成される条件の連言によって表される単一のルールということになる。これは、複雑な条件が必要となるような局面は数少ないと考え、また、人間との対話を行うという観点から、収束が早い単純なアルゴリズムを用いることに重点をおいたためである。

3.3.4 ルールの生成アルゴリズム

ルールの生成アルゴリズムは、一つの正の例と複数の負の例から、正の例を満足し、すべての負の例を満足しないルールを、いくつかの制約条件の連言として生成するものである。

ルールの生成は次のように行なわれる。はじめに、もっとも一般的なルールとして恒真をあらわす[](空集合)を生成する。これに対して、精密化オペレータを適用し、ルールを

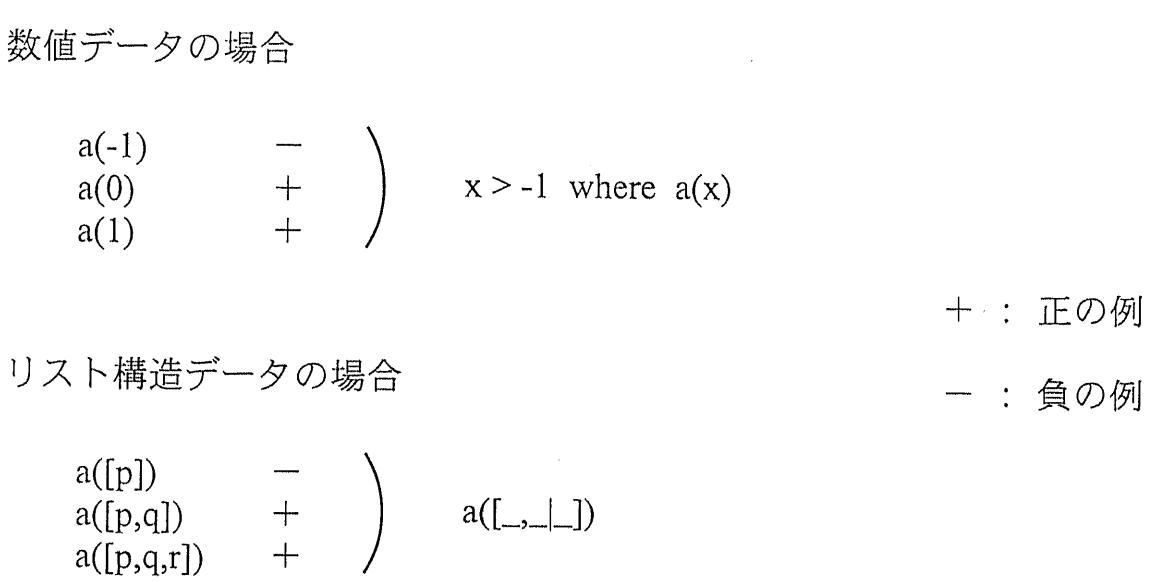


図 3.10: ルール生成の例

特殊化する。得られたルールに対して、更に精密化オペレータを適用することで、ルールは次々に生成されていく。ルールは正の例を満足するように生成されるため、生成されたルールの検定は負の例を適用することにより行なわれる。

本システムでは、モデル推論システムと異なり、パラメータの型を扱えるよう拡張されている。現在、扱えるパラメータの型は、リスト構造と数値データであり、パラメータの型は対話によりユーザから与えられるものとしている。リスト構造では、リストを構成する各要素の内容は参照しないが、任意の構造のリストを識別できるルールの生成が可能である。数値データについては、大小比較ができる一般の実数と、角度などのようなサイクリックになっているものを扱える。システムは必要になった時点である状態について各パラメータの型をユーザに問い合わせてくる。それに対しユーザは、リスト構造、一般の実数、サイクリックな数およびその値がとり得る上限と下限などといった情報を与える。

帰納的推論アルゴリズムでは、生成した条件を順次特殊化していく部分でパラメータのための条件を一つずつ付け加えていくことになるが、リスト構造に対する精密化は、リスト構造の複雑さが順次増していくよう行う。また、数値データに対する精密化は、ルール

におけるパラメータをXとすると、$L \leq X, X \leq H, L \leq X \leq H$のうちのいずれかのルールを生成させることにより行う。

通常の実数の場合のルールの生成を考えてみる。与えられた正の例に出てくる数値データの上限をT_{max}、下限をT_{min} とし、T_{max}以下にならない最小の負の例があればこれをF_{max}、T_{min}以上にならない最大の負の例があればこれをF_{min}とする。このとき、実数Xに対する条件は、次のようになる。

$$X_{max} = \begin{cases} \frac{T_{max}+Fmax}{2} & if F_{max} exists \\ T_{max} \end{cases} \tag{3.5}$$

$$X_{min} = \begin{cases} \frac{T_{min}+Fmin}{2} & if F_{min} exists \\ T_{min} \end{cases} \tag{3.6}$$

$$X \leq X_{max} \tag{3.7}$$

$$X_{min} \leq X \tag{3.8}$$

$$X_{min} \leq X \leq X_{max} \tag{3.9}$$

ルール生成アルゴリズムでは、単にこれらのルールのうちのどれを結果のルールに付加するか決定する。

3.3.5 地図図面におけるルールの学習

ルール作成支援システムを用いて、地図図面におけるルール生成の実験を行った。図3.7に示す地図図面を用いて、植生界を構成する点列を抽出するルールの生成を考える。実験により得ようとしているルールは、図3.5相当のものである。

1. 点列の構成要素となり得る小さい点を抽出するルールの生成

   はじめに、次のルールによりトークンの状態を遷移させておく。

$$\$loop \longrightarrow$$

$$\$area(A),$$

$$\$(A \geq 0),$$
$$\$\mathtt{transform}(\$\mathtt{loop}(A)). \tag{3.10}$$

このルールにより、面積値をパラメータに持つ点の候補のトークンが生成される。ここで得たいのは、面積の閾値である。システムは、ユーザから2つの正の例と2つの負の例の指示を受けて次のルールを生成した。これにより、小さい点を抽出するための面積の閾値を求めることができたといえる。

$$\$\mathtt{loop}(A) \longrightarrow$$
$$\$(A \leq 3200),$$
$$\$\mathtt{transform}(\$\mathtt{dot}). \tag{3.11}$$

2. 点トークンから点列を構成するものを識別するルールの生成

システムは、パラメータを含む状態を正、及び、負の事例として与えられて、これを識別するためのルールを作成するように働くが、点列を構成するものを識別するルールは、パラメータ計測と状況判断を別のルールで行う必要があるため、直接パラメータ計測のルールを生成することができない。そこで、次に示すようなルールを使い、ルール生成を行わせる。

$$\$\mathtt{dot} \longrightarrow$$
$$\$\mathtt{get\_near\_loop}(100, P_1), \$\mathtt{select}(P_1, \$\mathtt{dot}, Q_1),$$
$$\$\mathtt{get\_near\_loop}(105, P_2), \$\mathtt{select}(P_2, \$\mathtt{dot}, Q_2),$$
$$\$\mathtt{get\_near\_loop}(110, P_3), \$\mathtt{select}(P_3, \$\mathtt{dot}, Q_3),$$
$$\$\mathtt{get\_near\_loop}(115, P_4), \$\mathtt{select}(P_4, \$\mathtt{dot}, Q_4),$$
$$\$\mathtt{get\_near\_loop}(120, P_5), \$\mathtt{select}(P_5, \$\mathtt{dot}, Q_5),$$
$$Q = [Q_1, Q_2, Q_3, Q_4, Q_5],$$
$$\$\mathtt{transform}(\$\mathtt{dot}(Q)). \tag{3.12}$$

このルールにより、求めたい「近接を表す距離」となり得るいくつかの候補を用いて、それぞれの距離以内に存在する点を検索し、これらをリスト構造で表現し、周囲状況を表すパラメータ Q としている。このルールを適用したあと、ルールの生成を行わせる。1 つの正の例と 4 つの負の例から次のルールを生成した。

$$\$\mathtt{dot}([\_,\_,\_,[A,B]|\_]) \longrightarrow \$\mathtt{transform}(\$\mathtt{dot}(A,B)). \tag{3.13}$$

このルールにより、パラメータ Q の第 4 の要素が長さ 2 のリストであることが示される。これは、近接をあらわす距離が 115 であり、自分からその距離の範囲内に二つだけの点が存在するものということをあらわしている。生成されたルールを適用したところ、認識結果は満足のいくものとなった。

3.3.6 ルールの学習の評価

学習途中のルールによるエラーを定義し、本システムの評価を行った。

いま、ある状態 A になるべきトークンを分類するルールの学習過程を考える。候補トークン全体の集合を U、ユーザの想定する A という状態になるべきトークンの集合を X_A とする。一方、学習途中のルールで分類されるトークン集合を $\tilde{X}_A$ とする。ルール学習により、これらの値を近づけていく必要がある。ここで、学習中のルールによるエラー ϵ を定義する。

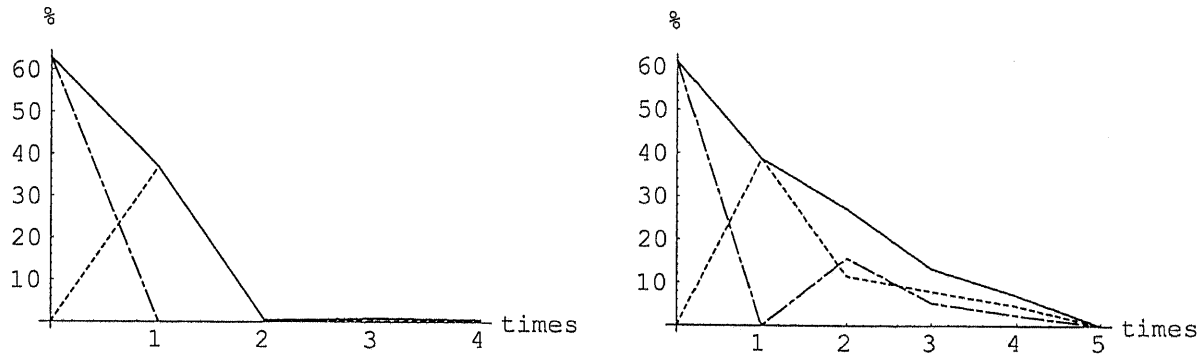
$$\epsilon_1 = \frac{\left|X_A - \tilde{X}_A\right|}{|U|} \tag{3.14}$$

$$\epsilon_2 = \frac{\left|\tilde{X}_A - X_A\right|}{|U|} \tag{3.15}$$

$$\epsilon = \epsilon_1 + \epsilon_2 \tag{3.16}$$

ここで $|\cdot|$ は集合の濃度を表し、$-$は差を表す。前者はルールにより識別されるトークン集合が小さいことを表し、後者はルールにより生成されるトークン集合が大きいことを表す。先に示した点の抽出と点列の抽出でルール生成を行った場合の結果を図 3.11に示す。

(a) の結果からは、2 回の指示でほぼ妥当なルールの生成できていることがわかる。エラーが 0 にならないのは、面積だけでは分類できない点が存在しているためである。また、(b) の結果からは 5 回の指示でほぼ妥当なルールの生成ができていることがわかる。これにより、本ルール作成支援システムの効果を確認することができた。



(a) ドットの分類　　(b) ドット列抽出

図 3.11: 実験結果 (鎖線はϵ_1、点線はϵ_2を表す)

3.4 まとめ

本章では、はじめに、多目的型認識システムの実現を目指して提案された状態遷移モデルを用いた認識システムについて概説し、状態遷移ルール作成のために開発したルール作成支援システムについて述べた。

状態遷移モデルは、状態と状態遷移の方法の定義の仕方により、様々な対象の認識に利用できる。ここでは、地図図面の例を示したが、セグメントなどを基本認識要素として利用すれば、後の章でふれるように、画像への適用も容易である。また、第 2 章で少し紹介したが、画像のシーンをあらわす枠組としても利用可能である。このように、状態遷移モデルは多くの可能性を持っているが、モデルをルールという形で記述できなければならないという問題も抱えている。これは、ルールで記述するものすべてに共通の問題であるが、認識対象によってはルール化することが困難なものも多い。また、現在は、ボトムアップ

プロセスが終了したあとトップダウンプロセスが起動するようになっているが、これらのプロセスを繰り返し組み合わせて実行できるようにすると、より柔軟で高度な認識が可能となると考えられる。ただし、この場合、現在、ボトムアップにおける状態の遷移が一意に行なわれるのに対し、一つの可能性として状態の遷移が行なわれなければならならず、組み合わせ問題が生じる可能性がある。

続いて、ボトムアップルールの作成を容易にするルール作成支援システムについて述べた。このシステムでは、GUI を用いて、平易な対話によりルールの生成が可能である。ルールの学習アルゴリズムとしては、帰納的推論アルゴリズムを用いた。ここで用いた帰納推論アルゴリズムは、Shapiro のモデル推論システムをもとに、数値とリスト構造が扱えるように拡張したものである。実験では地図図面の認識のためのルールの生成実験を行ない、十分実用的なルールの生成が可能であることを確認できた。このシステムでは、現在、ボトムアップルールのみを対象としており、画像認識で重要となるグルーピングや対象のモデルをあらわすトップダウンルールを扱うことができない。複雑な構造を持つ対象を扱う場合には、これらへの対応が強く要求される。このため、今後の課題の一つとして、より一般化したルール作成支援についての検討があげられる。

第 4 章

画像データベースを用いたシーン/カットの同定手法

4.1 概要

現在、多くのテレビ番組が日夜をとわず提供されている。これは、映像情報のデジタル化や、通信衛星や光ファイバの普及に代表される通信手段の充実により、テレビ放送の多チャンネル化が進むなど、ますます増加する傾向にある。このような中、利用者が必要なテレビ番組を自動的に発見、選択できたり、ダイジェスト版を自動的につくるなどの映像処理技術への要求が高まってきている。そこで、本章では、映像フィルタリングや画像のトップダウン解析を可能とする画像のシーン/カットの認識技術について取り上げる。

現在までに、シーンの同定や映像フィルタリングを行う研究としては、様々な形態のものが行われてきた。映像シーケンスを特徴量により記号列に変換し、その記号列の照合を行うことでシーンの同定を実現するもの [45] や、画面内のオブジェクトの動きや配置などをルールという形で記述し、利用者の要求にそった柔軟な映像フィルタリングを実現するもの [59] などである。しかし、これらは、コマーシャルのようにいつでも映像シーケンスが全く同じものである必要があったり、人間が画面構成などのシーン/カットの特徴を細かくルールで記述しなければならないなど、一般の多くの映像、例えば、テレビ番組などに適用するには難しい問題を抱えている。

このため、このような問題を考慮し、多くのシーン/カットに対応できるように、シーン/カットを同定する知識として実例である画像そのものを利用できる画像データベースを用いたシーン/カットの同定手法 [12]-[6] の枠組を提案する。この手法は、代表的なシーン/ カットをあらわす画像をデータベース化し、画像データベース内の画像と対象画像を比較、対応づけを行なうことで、対象画像のシーン/カットの同定を実現しようとするものである。画像の対応づけには、画像の類似性に基づいた対応づけを可能とする類似画像検索技術を用いる。画像の類似性を評価するための特徴としては、映っているオブジェクトの認識には立ち入らないとの立場から、画像の全体的特徴をあらわす統計的特徴量を利用する。

本章では、はじめに、提案する画像データベースを用いたシーン/カットの同定手法について述べ、映像フィルタリングや画像認識への応用について説明する。

4.2 画像データベースを用いたシーン/カットの同定

4.2.1 シーン/カットの同定の枠組

典型的なシーン/カットの画像の場合、画像の全体的特徴からシーン/カットの内容が理解できることが少なくない。(シーンとは内容の意味的なまとまりを、カットとは画面構成の同一性をあらわすものとする。) また、テレビ画像などの作為的に撮影された画像では、カメラ位置の制約や演出意図などの観点から、画面内でのオブジェクトの映る大きさや位置、そして、背景など画面構成がほぼ定まっている場合が多いと考えられる。このため、図 4.1に示すように、画面構成など画像の全体的特徴からシーン/カットを同定できる可能性があるといえる。ここで提案する手法はこのような考えに基づくもので、予め代表的なシーン/カットの画像をデータベース化し、その画像データベースから対象画像との類似画像を検索することで、対象画像のシーン/カットの同定を行なおうとするものである。画像データベースを用いたシーン/カットの同定の枠組を図 4.2に示す。ここで、代表的なシーン/カットの画像から構成する画像データベースのことを、シーンデータベースと呼ぶことにする。

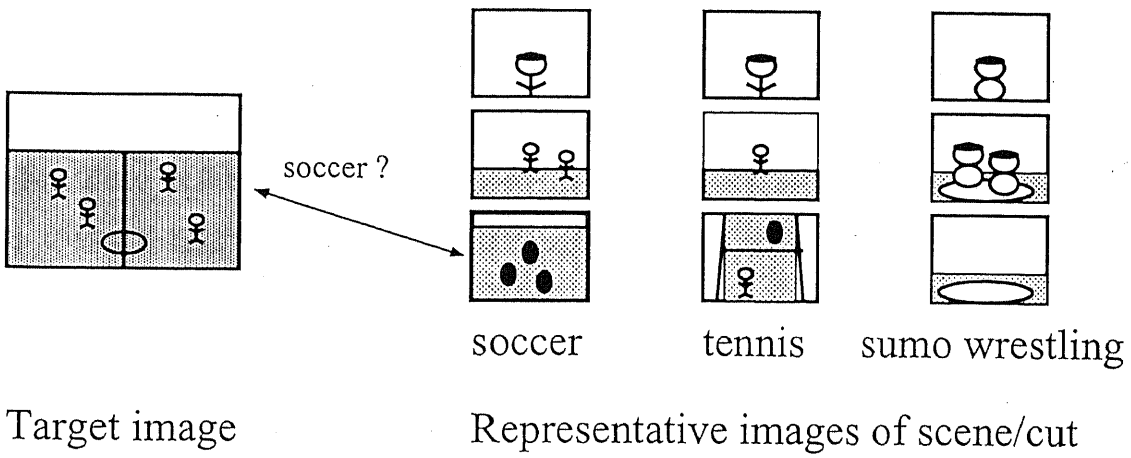


図 4.1: 画像の類似性によるシーン/カットの同定

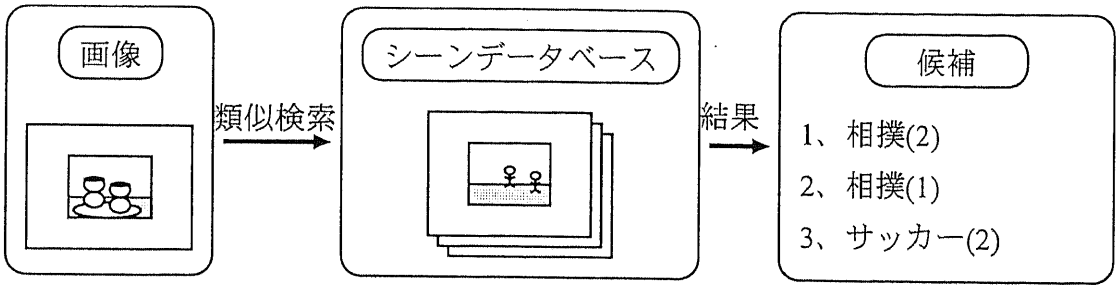


図 4.2: 画像データベースを用いたシーン/カットの同定の枠組

4.2.2 本手法の特徴

一般に、画像のシーン/カットは、画像の意味的構成から考えて、画像に映っているオブジェクトや、オブジェクト同士の関係などから定義され、それらの情報が認識できてはじめてシーン/カットの認識が可能となる。しかし、画像認識では、シーン/カットとオブジェクト、オブジェクトとセグメントなど階層構造のなかでの相互依存性があり、通常、セグメンテーションの実行にはどのようなオブジェクトの認識を行ないたいのか、オブジェクトの認識にはその画像のシーン/カットが何であるのかなど、上位の概念が理解できてはじめて可能となるという側面がある。画像認識におけるシーン/カットの認識とオブジェクトの認識の相互依存性と本手法の位置付けの関係を図 4.3に示す。

本手法では、対象画像をシーンデータベース内の画像へ直接対応づけることで、シーン/カットの同定を行なう。このため、オブジェクトの詳細な認識は基本的に必要とせず、以下のような応用が期待できる。

- シーン/カット情報による画像のトップダウン解析

  同定により得られたシーン/カットの情報により、シーン/カットの情報を用いたトップダウン解析、あるいは、シーン/カットに応じた画像認識手法の選択、利用が可能となる。これにより、対象画像を幅広く設定することも可能で、また、直接オブジェクトの認識を行なうものに比べ、より柔軟な認識を実現できる。

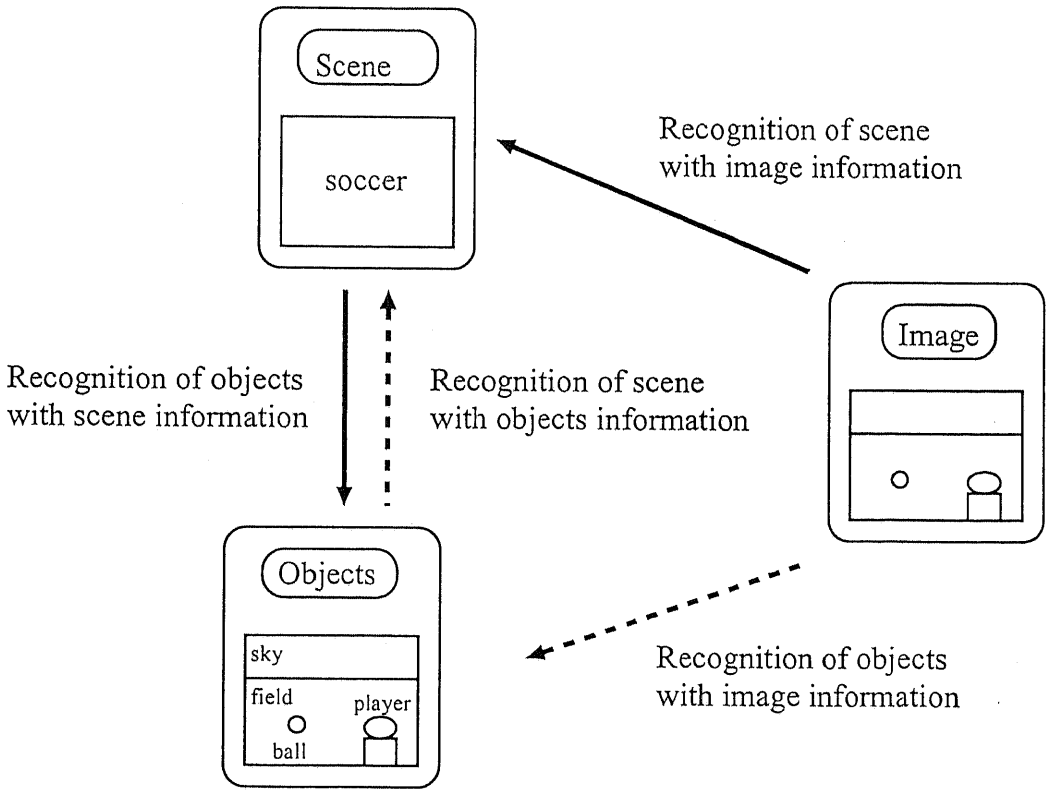


図 4.3: 画像認識における相互依存性

- 映像フィルタリング

  予め必要とする映像に対応するシーンデータベース内の画像を指定し、与えられる映像とその画像との類似性を評価することで、必要な映像を選択、抽出する映像フィルタリングが実現できる。

また、本手法では、シーン/カットの同定のための知識として、体系化されたものでなく、実例である画像そのものを利用できるため、拡張が容易に行なえる点も大きな特徴の一つである。これは、シーンデータベース内の画像そのものが、シーン/カットをあらわすモデルとなるため、対象画像の変更や、多様な画像への対応など、シーンデータベース内の画像を追加、変更することで簡単に実現できるということである。

4.3 シーンデータベースの構成と類似画像検索手法

4.3.1 シーンデータベースの構成

シーンデータベースは、画像・映像のシーン/カットをあらわす複数の代表的な画像から構成する。各画像に対しては、検索実行時の時間的負荷を軽減するため、また、柔軟な画像認識を実現するため、検索に用いる特徴量や、その画像に対する認識モデルなどを予め付与しておく。詳しくは後述する。

シーンデータベース内の各画像は、類似画像検索により検索された際、対象画像のシーン/カットの意味づけを行なえなければならない。このため、図4.4に示すように、各画像は、シーン、カット、画像という階層的な構成により管理し、その画像が示すシーンやカットを明確にしておく。また、階層的に管理することで、特徴量のバラツキへの対応や、後述するように、映像のシーン/カットの同定への対応も可能となる。

4.3.2 シーン/カットにおける画像の類似性

画像のシーン/カットの同定を実現する上では、同一シーン/カットにおける画像の類似性について検討しておかなければならない。類似性を検討するにあたり、以下の点について注意する必要がある。

- オブジェクト特徴の多様性の影響

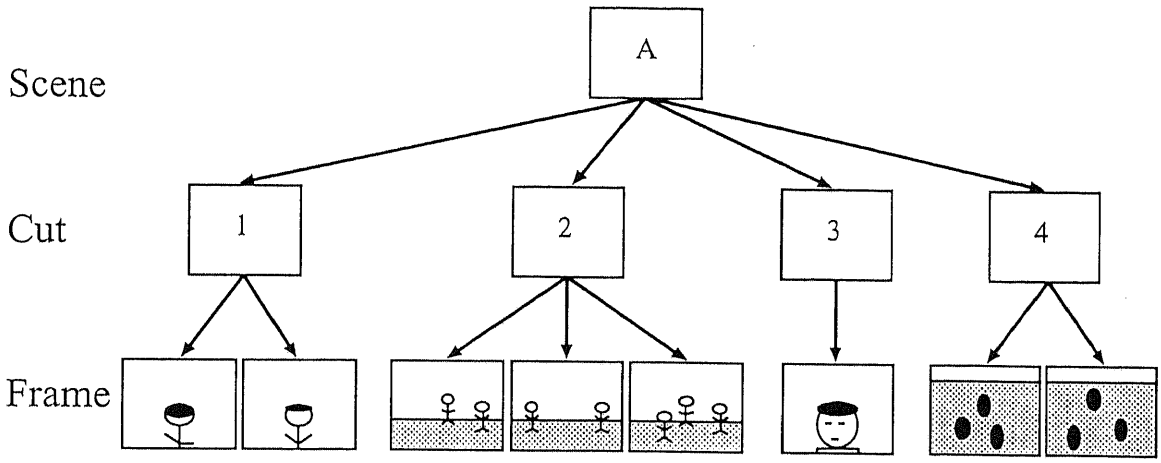


図 4.4: シーンデータベースの階層的構成

同一のシーン/カットと思われる画像でも、映っているオブジェクトの色や形、向きなどに違いがあり、画像の特徴に違いが生じる。これは、例えば、同じ人間であっても服装により色情報などが異なることを意味する。

- オブジェクトの動きの影響

  同一のシーン/カットと思われる画像でも、映っているオブジェクトの位置や姿勢は様々であり、画像の特徴に違いが生じる。これは、例えば、同じシーン/カットであっても人間の映っている位置などに違いがあることを意味する。

- 撮影条件やカメラパラメータの違いの影響

  同一のシーン/カットと思われる画像でも、光の加減やカメラワークなどにより、画像特徴に違いが生じる。これは、例えば、同一の人間であっても人間の映る色や大きさなどに違いが生じることを意味する。

本手法では、これらの影響に対し、適切なシーン/カットの定義、及び、各シーン/カットにおける有効特徴量の選択により対応する。

4.3.3 類似画像検索のための特徴量

シーンデータベースを用いたシーン/カットの同定は、対象画像に対して類似画像検索を行なうことで実現される。類似画像検索は、一般に何らかの画像から直接抽出できる情報、多くは特徴量の類似性を評価することで行なわれる。このため、類似画像検索にどのような特徴量を利用するのか、また、類似性をどのように評価するのかが大きな課題となる。類似画像検索に用いる特徴量としては、現在までの様々な画像検索システムの開発 [43]-[56] に伴い、色情報や形状情報、テクスチャ情報に基づいたものなど多くのものが検討されてきているが、その選択はシステムの開発者にまかされ、どの特徴量がどのような目的に適しているかなど、選択の基準はあまり明確ではない。そこで、ここでは、以下に示すようないくつかの画像から直接抽出できる特徴量を採用し、それら特徴量を複合的に利用することを検討する。ここで用いる特徴量としては、信号レベルのもの、つまり、画像から直接求めることができるものを利用することにした。これは、セグメンテーションなどによる構造的特徴の抽出は重要であるが、その不正確さが問題になる場合も多いためである。

- 色情報

  - 色相ヒストグラム (図 4.5参照)

    HSV(色相、彩度、明度) データにおける色相を 18 分割、及び、無彩色である白、灰色、黒を合わせた、計 21 次元。

$$\{H_1(1), H_1(2), \cdots, H_1(21)\} \tag{4.1}$$

  - 隣接色相ヒストグラム (図 4.5参照)

    HSVデータにおける色相を 6 分割、対象性を考慮し、計 21 次元。

$$\{H_2(1,1), H_2(1,2), \cdots, H_2(6,6)\} \tag{4.2}$$

  - RGBデータの平均、分散

画像を $f(x,y) = (r(x,y), g(x,y), b(x,y))^T | x \in X, y \in Y$ とする。

$$E = \frac{1}{XY} \sum_{x \in X, y \in Y} f(x,y) \tag{4.3}$$

$$\sigma^2 = \frac{1}{XY} \sum_{x \in X, y \in Y} (f(x,y) - E)^2 \tag{4.4}$$

RGBに対し、各 1 次元、計 6 次元。

- テクスチャ情報

画像の濃度 i の点から一定の変位$\delta = (r, \theta)$ だけ離れた点の濃度が jである確率 $P_\delta(i,j), (i,j = 0,1,\cdots,n-1)$ とする。

- コントラスト

$$F_1 = \sum_{k=0}^{n-1} k^2 \{ \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P_\delta(i,j) \}_{k=|i-j|} \tag{4.5}$$

- 角 2 次モーメント

$$F_2 = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \{P_\delta(i,j)\}^2 \tag{4.6}$$

- エントロピー

$$F_3 = -\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \{P_\delta(i,j) log P_\delta(i,j)\} \tag{4.7}$$

- 相関

$$F_4 = \frac{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} ij P_\delta(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{4.8}$$

$$\mu_x = \sum_{i=0}^{n-1} i \sum_{j=0}^{n-1} P_\delta(i,j), \qquad \mu_y = \sum_{j=0}^{n-1} j \sum_{i=0}^{n-1} P_\delta(i,j)$$
$$\sigma_x^2 = \sum_{i=0}^{n-1} (i - \mu_x) \sum_{j=0}^{n-1} P_\delta(i,j), \qquad \sigma_y^2 = \sum_{j=0}^{n-1} (j - \mu_y) \sum_{i=0}^{n-1} P_\delta(i,j)$$

$\delta = (1, 0°)$、$\delta = (1, 90°)$ に対し、各 1 次元、計 8 次元。

– 局所自己相関

画像の濃度 $f(x,y)$ における 1 次 (3 × 3 領域) の局所自己相関。

$$F_5 = \frac{1}{XY} \sum_{x \in X, y \in Y} f(x,y) f(x+a_x, y+a_y) \tag{4.9}$$

変移方向 $(a_x, a_y) = (0,0), (-1,-1), (0,-1), (1,-1), (1,0)$ に対して、各 1 次元、計 5 次元。

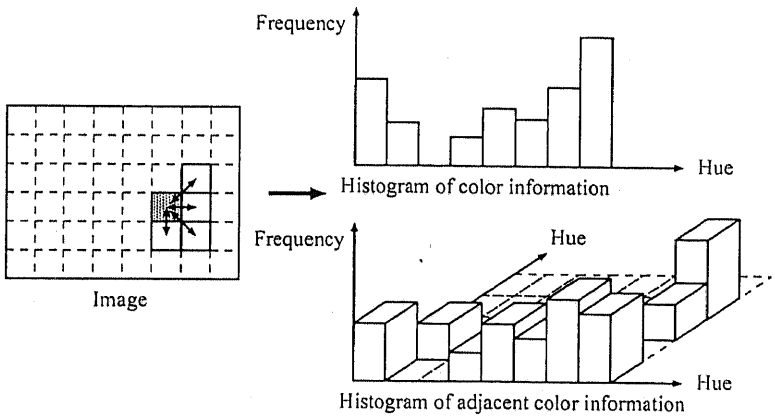


図 4.5: 特徴量の例—色相ヒストグラム

4.3.4 類似画像検索

シーンデータベースは、シーン、カット、画像というように階層的に構成されている。ここで、カットを同種の画像群からなる一つのクラスと考え、画像のシーン/カットの同定は、対象画像がどのクラスに類似するか、どの画像に類似するかにより判定を行なうことにする。

また、前述したように、シーン/カットの類似性を考える上では、オブジェクトの多様性や動き、撮影条件の違いなどを考慮しなければならない。そこで、同一カットに対する特徴量の変化を考慮するため、同一カットを構成する画像群における各特徴量の分散を利用

し、特徴量間の関係を正規化、有効特徴量の選択を行ない、画像の類似度を評価することにした。

シーン/カットの同定は図4.6に示すように、基本的には3ステップで実行する。この図における、シーン/カットの同定による画像認識の枠組については後述する。

1. 特定クラスとしてみた場合の対象画像と各クラスの距離の判定

   はじめに、対象画像のシーンデータベース内の各カットを定義する画像群との距離を求めるため、各カットにおける特徴量のクラス内平均、クラス内分散を求める。そして、これらの値を用いて、対象画像と各クラスの重心までの正規化ユークリッド距離を求めることで、各クラスとしてみた場合の対象画像とそのクラスの距離を求めることができる。これにより、第一段階の検索を行なう。ここで、クラス内分散により正規化を行なうことは、そのクラスにおける有効な特徴量に対し、重みをつけて評価していることにほかならない。クラスまでの距離の概念図を図4.7に示す。対象とクラスA、及び、クラスBまでの平面上での距離は同じであるが、クラスの分散を考慮すると、それぞれのクラスまでの距離は異なるものとなる。

2. 画像群全体からみた場合の対象画像と各クラスの距離の判定

   次に、対象画像がそのクラスとしてみた場合に近いと判断されたクラスに対して、クラス間分散を用いて、対象画像とそのクラスの重心の正規化ユークリッド距離を求め、シーンデータベース内の画像群全体からみた場合の、対象画像とそのクラスとの距離を求める。これにより、第2段階の検索を行なう。

3. 対象画像と各クラス内の画像の距離の判定

   最後に、ある程度近いと判断されクラス内の各画像に対し、クラス間分散を用いて、対象画像との正規化ユークリッド距離を求めることにより、類似画像の検索を行ない、シーン/カットの同定を行なう。

具体的計算は次のように行なう。ここで、シーンデータベースは、シーンI個、カットJ個、画像K個から構成され、i番目のシーンをM_i、シーンM_iのうちj番目のカットをM_{ij}、カットM_{ij}のうちk番目の画像をM_{ijk}であらわすものとする。また、シーンM_iは

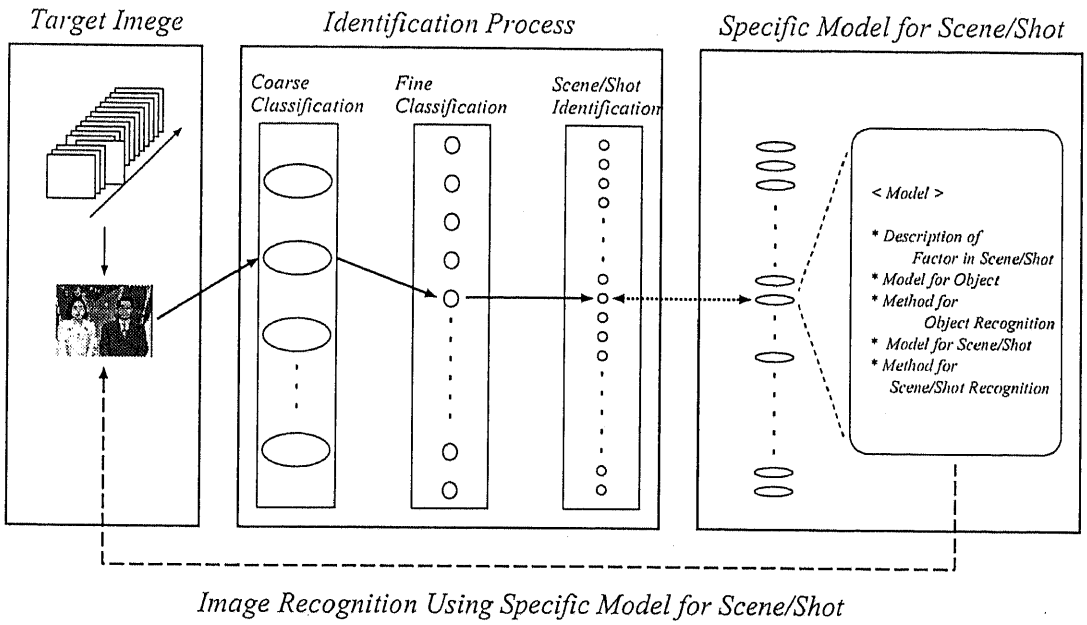


図 4.6: シーン／カットの同定の過程

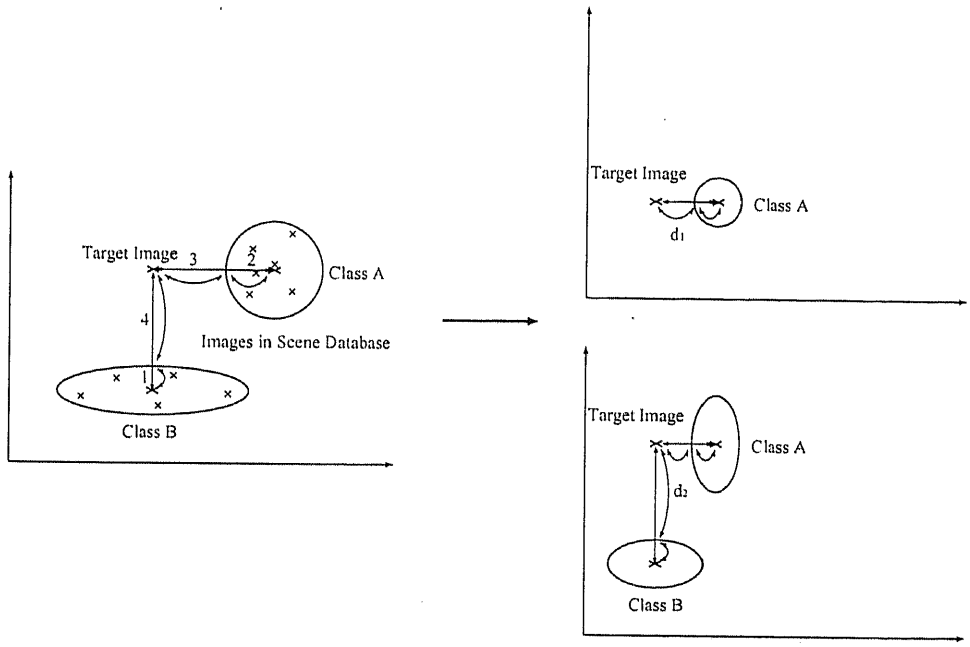


図 4.7: 距離の概念図

カット J_i個 ($J = \sum_i J_i$)、シーン M_iのカット M_{ij}は画像 K_{ij}個 ($K = \sum_i \sum_j K_{ij}$) からなるものとする。

シーンデータベース内の画像 M_{ijk}から計算される L 個の特徴量を $X^l_{ijk}, (l = 1, 2, \cdots, L)$、対象画像 Iから計算される L 個の特徴量を X^lとすると、対象画像とシーンデータベースで定義されるシーン/カットの類似は以下のように評価する。

はじめに、各カットを定義する画像群から、各特徴量のクラス内平均$\overline{X}^l_{ij}$、クラス内分散$(\sigma^l_{ij})^2$を求める。

$$\overline{X}^l_{ij} = \frac{\sum_{k=1}^{K_{ij}} X^l_{ijk}}{K_{ij}} \tag{4.10}$$

$$(\sigma^l_{ij})^2 = \frac{\sum_{k=1}^{K_{ij}} (X^l_{ijk} - \overline{X}^l_{ij})^2}{K_{ij}} \tag{4.11}$$

次に、各特徴量の全体の平均$\overline{X}^l$とクラス間分散$(\sigma_B^l)^2$を求める。

$$\overline{X}^l = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} X_{ijk}^l}{\sum_{i=1}^{I}\sum_{j=1}^{J_i} K_{ij}} \tag{4.12}$$

$$(\sigma_B^l)^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} (\overline{X}_{ij}^l - \overline{X}^l)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J_i} K_{ij}} \tag{4.13}$$

1. クラス内分散によるクラスの同定

クラス内分散を利用し、対象画像の特徴量と各クラスの平均特徴量との正規化ユークリッド距離$d_{1_{ij}}$を求め、類似するクラスの候補を限定する。$d_{1_{ij}} < th_1$を候補とする。

$$d_{1_{ij}} = \sqrt{\sum_{l=1}^{L} \frac{1}{(\sigma_{ij}^l)^2}(X^l - \overline{X}_{ij}^l)^2} \tag{4.14}$$

2. クラス間分散によるクラスの同定

クラス間分散を利用し、対象画像の特徴量と限定された各クラスの平均特徴量との正規化ユークリッド距離$d_{2_{ij}}$を求め、類似するクラスを求める。$d_{2_{ij}} < th_2$を候補とする。

$$d_{2_{ij}} = \sqrt{\sum_{l=1}^{L} \frac{1}{(\sigma_B^l)^2}(X^l - \overline{X}_{ij}^l)^2} \tag{4.15}$$

3. 類似画像の検索

最後に、クラス間分散を利用し、対象画像の特徴量と各候補クラス内の画像の特徴量との正規化ユークリッド距離$d_{3_{ijk}}$を求め、類似する画像を決定する。

$$d_{3_{ijk}} = \sqrt{\sum_{l=1}^{L} \frac{1}{(\sigma_B^l)^2}(X^l - X_{ijk}^l)^2} \tag{4.16}$$

4.3.5 シーン/カットの同定実験

提案手法のシーン/カットの同定の能力を検討するため、スポーツ画像を対象としたプロトタイプの同定システムの構築を行なった。実験に用いたシーンデータベースは、水泳、相撲、ゴルフ、サッカーなど7種類のスポーツ、各々5〜10カット、全体で約1000枚の画像から構成されている。図4.8にシーンデータベース内の画像の例を示す。また、類似画像検索結果の例を図4.9に、距離d_1及びd_2を用いた場合の正解率を表4.1に示す。ここで、正解率は任意のそのシーンの画像50枚与えたときに、それと同じカットの画像が検索されたかどうかの割合を示す。

この正解率からわかることは、色的に特徴的な水泳などは判別しやすいが、ゴルフとサッカーなどほぼ同じような色合い、画面構成のものに関しては、どちらか一方の正解率が高くなる反面、他方は低くなる傾向があることがわかる。これは、検索のための条件を緩くすることで解決される問題であり、画像的に類似しているものを同一と考える本手法の枠組のなかでは、シーンの候補を得ることができれば良いため、差し支えない。この結果から、シーンデータベースのカットの定義、クラス設定の方法が、判別率に大きな影響を持つことがわかる。

表4.1: カットの同定結果

対象	正解率 d_1	d_2
相撲	0.53	0.53
サッカー	0.68	0.27
水泳	0.65	0.65
ゴルフ	0.27	0.68
バレーボール	0.48	0.10

正解率Cは、次のように定義される。

- 正解率

$$C = \frac{\text{正しく同定された画像数}}{\text{検索に用いた画像数}} \tag{4.17}$$

図 4.8: シーンデータベース内の画像の例

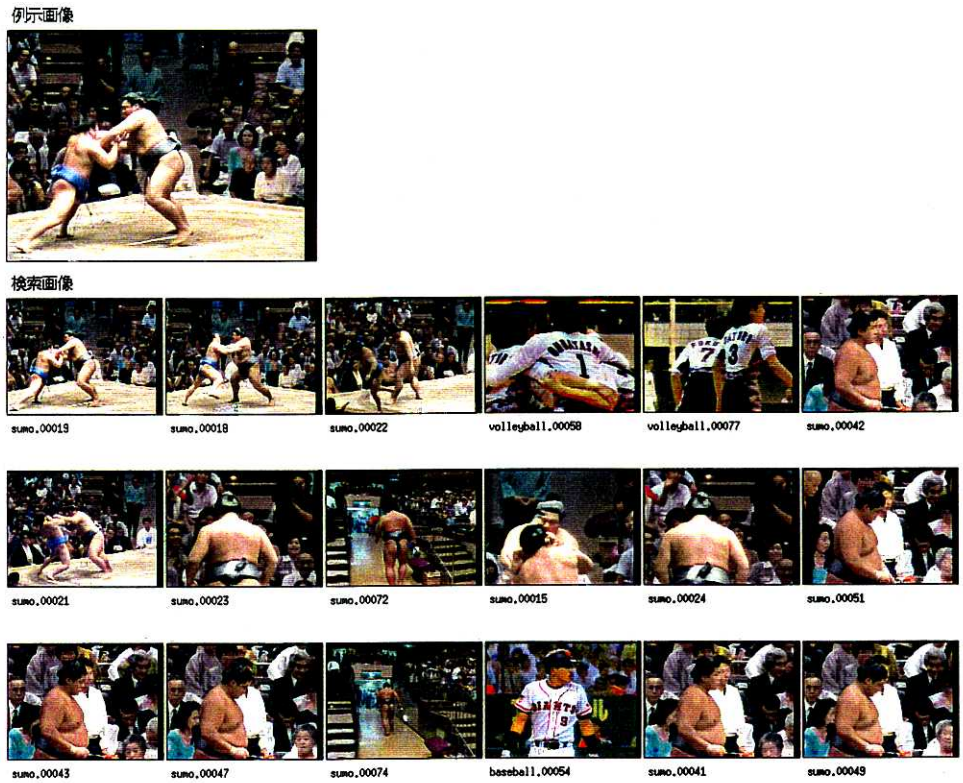


図 4.9: 類似画像検索結果

また、図 4.10に画像シーケンスにおけるシーン/カットの類似度の関係を示す。これは、10 秒おきにとった画像 400 枚に対し、シーン/カットの類似度を計算したものである。実験に用いた映像は、0 から 200 フレーム、300 から 400 フレームまでは一般のテレビ映像から、200 から 300 フレームまではサッカー映像から構成されているものである。

この結果から、シーンデータベースが対応していない一般の映像部分に対しては、どちらのシーン/カットの類似度も基本的には低い値となることがわかる。また、サッカー映像の部分に対しては、サッカーシーンのみ類似度が高くなり、他のシーンに関しては、他の映像部分と変わらない値となることが読みとれる。(ここでは、グラフをみやすくするため、サッカーシーンとバレーボールシーンの類似度を示しているが、他のシーンに対しても、バレーボールと同様の結果が得られている。) これより、本手法を用いることで、リアルタイムでの映像のシーン/カットの判別を実現できる可能性が示せたのではないかと考える。

4.4 シーン/カットの同定による画像認識

4.4.1 画像認識の枠組

画像のシーン/カットの同定が可能でれば、そのシーン/カットの情報を利用した画像認識が可能となる。これは、シーン/カットの同定に用いるシーンデータベース内の代表的な各カット、あるいは、各画像に対し、そのカットや画像に依存したオブジェクトの認識モデル、認識手法を関連づけることで、認識モデルの単純化、特定の認識手法の利用が可能となるためである。これにより、認識モデルの設定や認識処理が容易になり、認識処理時間の短縮や認識率の向上が期待できることになる。

これは、例えば、スポーツ画像での人間の映り方を考えてみると、図 4.11に示すように、大きく顔が映っている場合や小さく全身が映っている場合、また、ユニフォームを着ている場合や着ていない場合など様々な状態がありえるため、これらを単純に人間という枠でモデル化するのは大変困難なものとなるが、サッカーや相撲などと、シーンやカットが特定できれば人間の映り方も限定され、モデル化が行ないやすくなるということである。また、サッカーのフィールドなど、シーン/カットがわからないと認識が難しいものに関しても認識が可能となることが期待できる。

シーンデータベースの構成を図 4.12、また、画像認識の過程を図 4.6、図 4.13に示す。

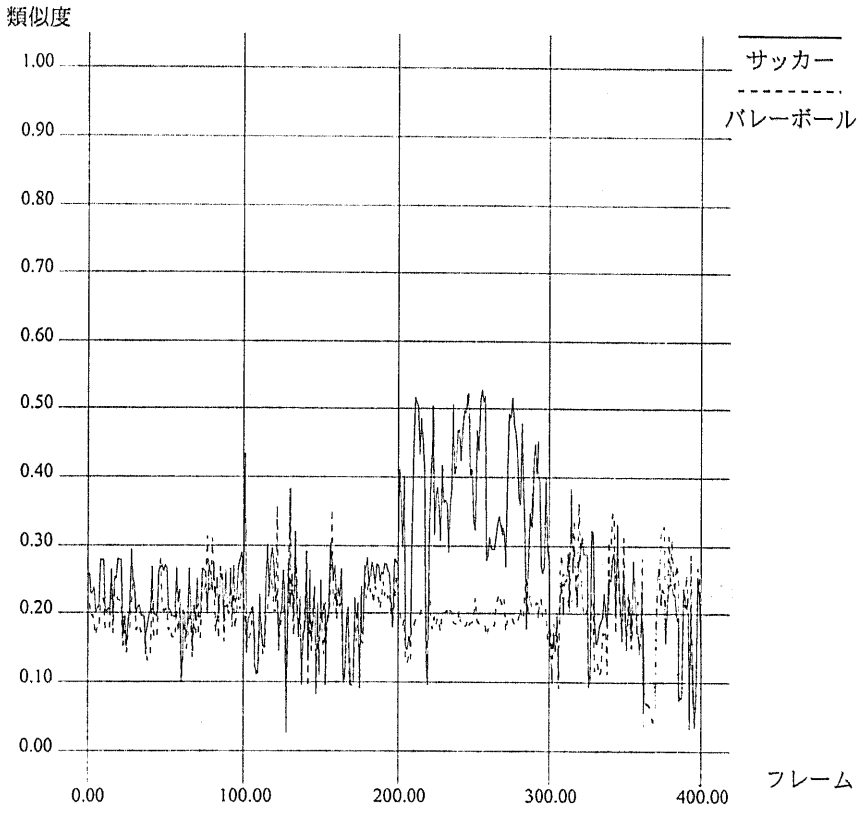


図 4.10: 画像シーケンスにおけるシーン/カットの類似度

シーン/カットの同定による画像認識は、はじめに、シーン/カットの同定を行ない、その後、その結果に応じた認識モデル、認識手法によりオブジェクト認識を実行するという形で行うことになる。

図 4.11: 人間が映っている画像の例

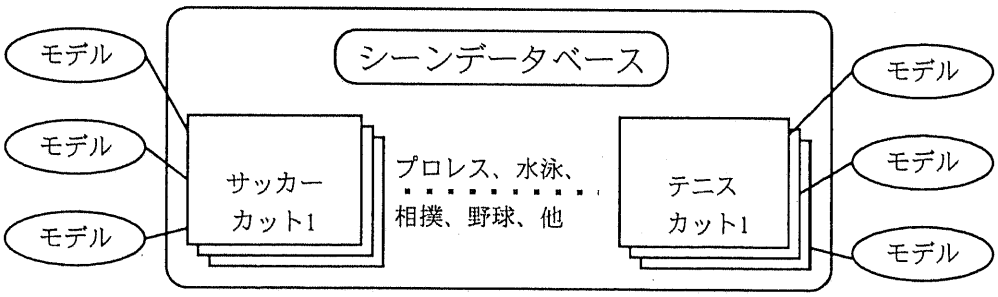


図 4.12: シーンデータベースの構成

4.4.2 認識モデル

認識モデルとしては、図 4.14に示すような状態遷移モデルを利用する。これは、第 3 章で述べた状態遷移モデルを自然画像へ適用したもので、ここで示すダイアグラムは、上半身が映っているようなサッカー選手の認識を行うモデル例である。

はじめに、色情報によるセグメンテーションを行い、そのセグメントを初期トークン「Segment」とする。次に色の評価を行うことで、次の「Green_Segment」などの状態へボ

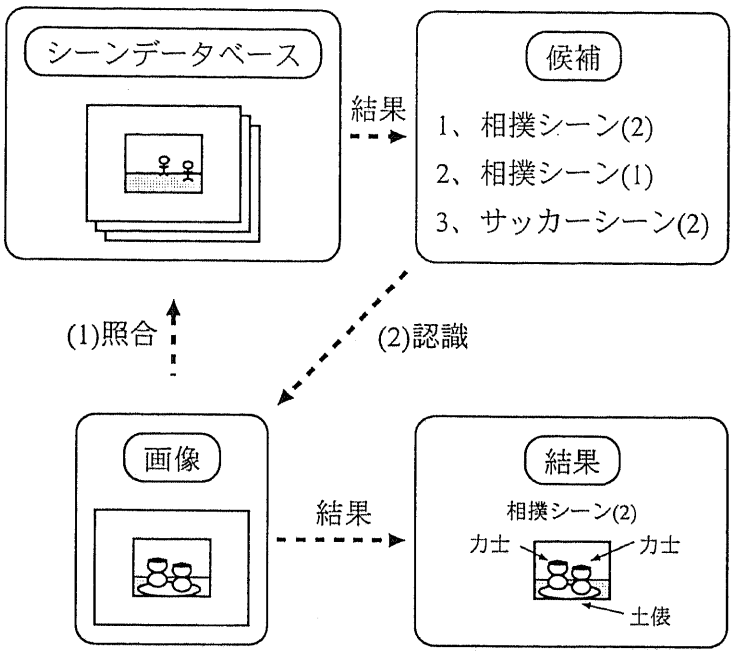


図 4.13: 画像認識の過程

トムアップ的に遷移させる。そして、その状態からは、確実な状態の遷移は不可能なので、トップダウン的にセグメントの大きさ、形などの特徴量を評価して、次の「Field」などの状態へ遷移させ、最後に「Not_Field」、「Face」、「Hair」の状態を利用して、選手の認識を行うというものである。自然画像への適用の場合、トークンはセグメントを認識基本要素とする。

4.4.3 シーン/カットの同定による画像認識の例

前述したスポーツ画像を対象としたシーン/カットの同定システムを用いて、画像認識の実験を行なった。ここでは、スポーツ画像における人間の認識の簡単な実験例を示す。類似画像検索、及び、人間の認識結果を図 4.15に示す。認識モデルとしては、前述した状態遷移モデルを利用している。

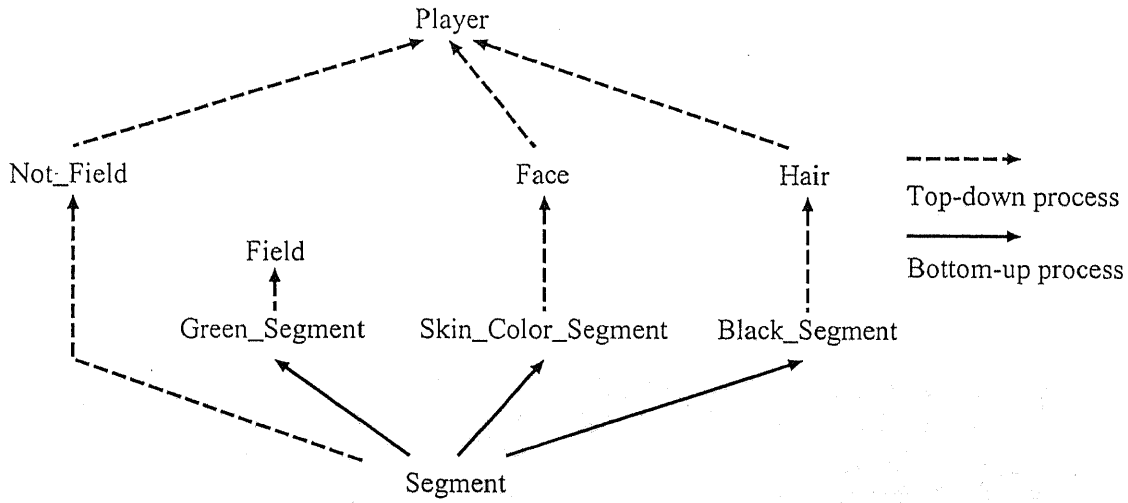


図 4.14: サッカー選手の認識ダイアグラムの例

4.5 シーンデータベースを用いた映像フィルタリング

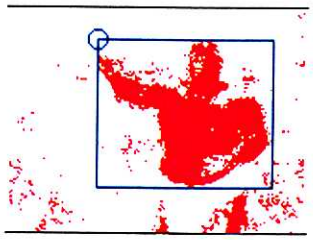
4.5.1 映像フィルタリングの枠組

映像フィルタリングとは、利用者の視点で、多くの映像の中から必要な映像を選択、抽出することである。ここでは、利用者の視点として、サッカーや水泳などといった、シーン、あるいは、カットレベルのフィルタリングを考えている。

画像のシーン/カットの同定は、シーンデータベースから対象画像に対する類似画像を検索することで行うものであった。これに対し、シーンデータベースを用いた映像フィルタリングでは、シーンデータベース内の画像から、フィルタリングしたいシーン/カットの画像を選択し、流れてくる映像に対してその選択した画像との類似性を判定することで、フィルタリングを実現する。シーンデータベースを用いた映像フィルタリングの枠組を図 4.16 に示す。

この方法では、対象映像と、フィルタリングしたいシーン/カットの画像、あるいは、カットのクラスの類似評価を行なうことでフィルタリングが実現されるため、計算量は選択した画像、あるいは、カットの数に比例することになる。

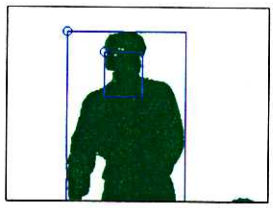
(a) Target Image



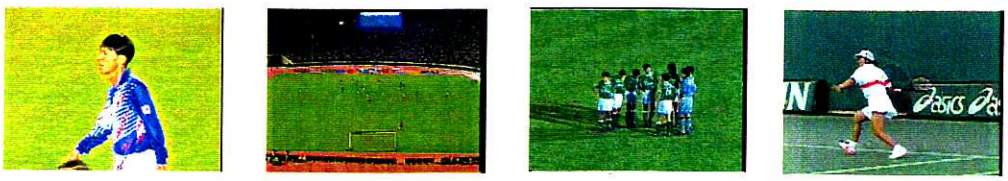
(b) Result of Recognition

(c) Result of Retrieval

(a) Target Image



(b) Result of Recognition



(c) Result of Retrieval

図 4.15: 実験例

これは、シーン/カットの同定を行ないながらフィルタリングするのに比べ、速度的にはかなり高速で実現できる。

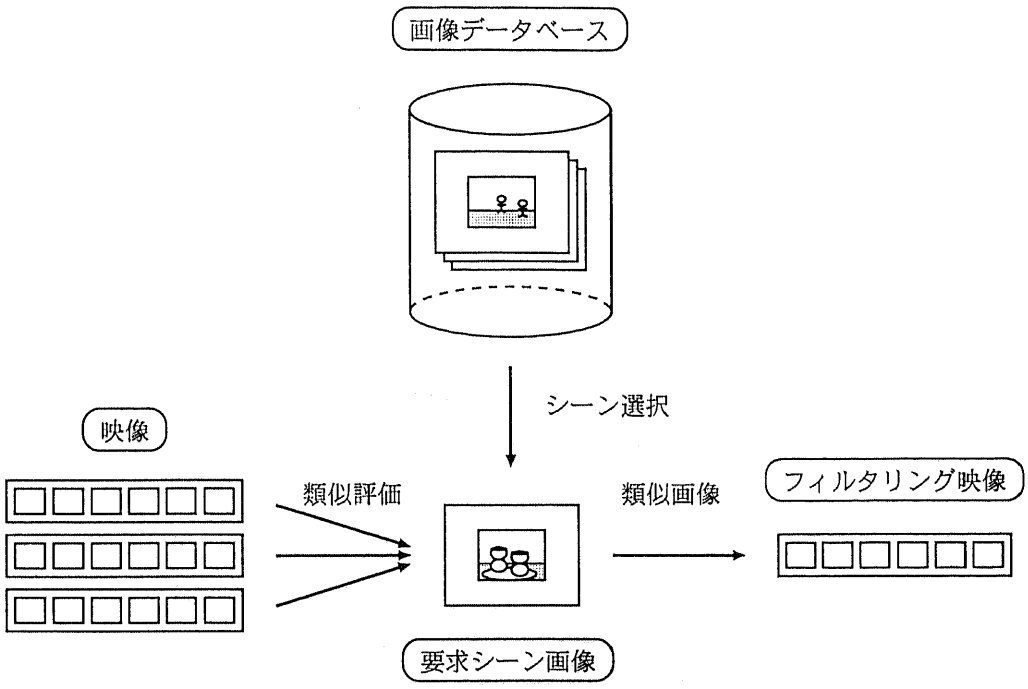


図 4.16: 映像フィルタリングの枠組

4.5.2 映像フィルタリングの例

ここでは、サッカーシーンの映像フィルタリングの実験例を示す。実験に用いたシーンデータベースは、画像のシーン/カットの同定に用いたものと同じである。フィルタリング結果の映像、及び、その評価を図 4.17、表 4.2に示す。これは、対象シーンの画像を数枚用意し、その画像までの距離を判定し、求めた結果である。実際に入力した画像は、一般の映像シーケンスであるため、対応できないシーンもあり、再現率はあまり高くならない。これはフィルタリング要求を与える画像を増やすことで解決がはかれると考えられる。また、サッカーの適合率が小さいのは、画面の似ているゴルフのシーンを取得してしまうためである。先のシーンの同定結果とともに考えると、シーンデータベースの構築が本手法

の能力を決定づける大きな要素であるといえる。今回のシーンデータベースは人間の主観により分類していた。これについては、多変量解析を行なうなど、検討しなければならないと思われる。

表 4.2: フィルタリング結果

対象	再現率	適合率
相撲	0.87	0.86
サッカー	0.60	0.38
ゴルフ	0.55	0.72

- 再現率 (recall ratio)

$$R = \frac{\text{検索結果に含まれる正解画像数}}{\text{正解画像数}} \tag{4.18}$$

- 適合率 (precision ratio)

$$P = \frac{\text{検索結果に含まれる正解画像数}}{\text{検索結果の画像数}} \tag{4.19}$$

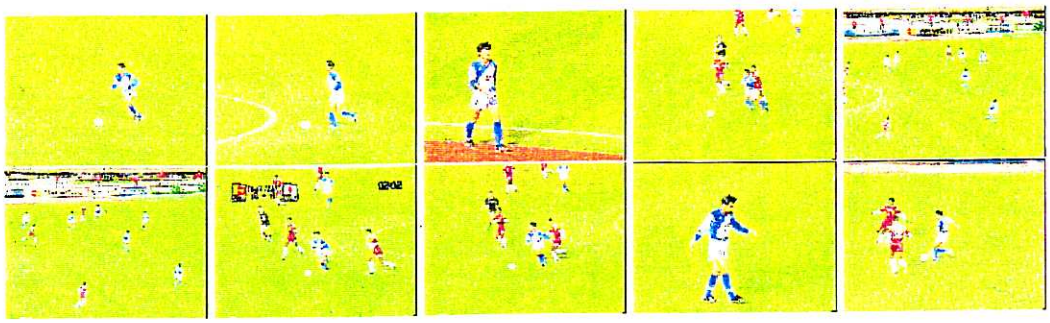
ここで、正解画像とは検索条件に対して検索されるべき画像のことである。再現率は検索されるべき画像のうちどれだけが実際に検索されたか、適合率は検索された画像のうちどれだけが検索されるべきものであったかをあらわしている。

4.6 映像のシーン/カットの同定

4.6.1 映像の特徴とシーン/カットの同定

提案した画像のシーン/カットの同定手法は、画像データベースと類似画像検索手法を用い、代表的なシーン/カットをあらわす画像群から、対象画像に類似する画像を検索することで、画像一枚一枚のシーン/カットの同定を行うものであった。このため、画像が連続的に得られるテレビやビデオなどの映像のシーン/カットの同定を行いたい場合でも、それら映像の特徴である画像シーケンスから得られる情報を利用することができない。ここでは、映像のシーン/カットの同定を実現するため、提案手法の映像への拡張について考える。

(a) Request Scenes



(b) Filtered Images

(c)Rejected Images

図 4.17: フィルタリング結果

テレビやビデオなどの映像を扱う場合、画像一枚一枚の特徴を用いるだけでなく、画像が連続的に得られるという映像の特徴を利用するべきである。シーン/カットの同定に利用できる映像の特徴としては、大きく二つを考えることができる。

- 動き
- 映像構造

動きの特徴とは、画像に映っている物体の移動やカメラワークなどによる画面上での動き情報から得られる特徴のことで、映像構造の特徴とは、シーン、カット、フレームというような映像の階層構造やそれぞれの連続性など映像の構造的な情報に着目した特徴のことである。

動きの特徴の利用においては、映像のカットにおける物体の移動やカメラワークによる画面上での動きの特徴をモデル化し、対象映像から抽出できる動きの特徴と比較、照合を行うことで、対象映像のシーン/カットの同定を実現することなどが考えられる。現在までに、オプティカルフローやブロックマッチングなどを利用することで、映像から動きの特徴を得ようとする研究が数多くなされており、得られた動きの特徴を用いて、画像内容の理解やカメラワークの認識を行う手法が検討されている。映像のシーン/カットの同定を行う場合、動きの特徴を用いることで、「ズーミングされている場面」、「物体が右から左へ移動している場面」などという観点でのシーン/カットの同定が可能となる。

一方、映像構造の特徴の利用においては、映像はシーン、カット、フレームというように階層的に構成されているため、連続して得られる複数の画像は同一のカットのものである可能性が高く、また、連続して得られる複数のカットは同一のシーンのものである可能性が高いというような観点から、対象映像の画像シーケンスの構造を評価することなどが可能となる。

映像のシーン/カットの同定において、動きは映像情報の本質に深く関連する特徴であるため重要であるが、ここでは、提案した画像のシーン/カットの同定手法により、映像を構成する画像一枚一枚のシーン/カットを同定することが可能であることに注目し、画像一枚一枚に対する同定結果と映像構造の特徴を用いることで、映像としてのシーン/カットの同定を実現することとした。

4.6.2 映像構造の利用

映像は図4.18に示すようにシーン、カット、フレームといった形で階層的に構成される。シーンは映像内容の意味的なまとまりを、カットは画面構成の同一性を、フレームは映像を構成する一枚一枚の画像をあらわし、シーンは複数のカットから、カットは複数のフレームから構成される。このような映像の構造的特徴を利用する方法としては、局所的な映像構造の利用と大局的な映像構造の利用の二つの側面から考えることができる。

局所的な映像構造の利用

映像を連続する画像の群と考える。このとき、先に述べた映像の構造から、連続して得られる画像は、同一のシーン、更にいえば、同一のカットのものである可能性が高い。特にカットチェンジの検出によりカット毎に分割された画像群の各画像は、ほぼ同一のカットの画像であると考えてよい。また、より細かくは、シーンにおける各カット、カットにおける各画像には、出現する順序の関係があるといえる。局所的な映像構造の利用とは、このような、連続して得られる画像は同一のカットの画像であり、連続して得られるカットは同一のシーンのカットであるとの構造的特徴を利用することを意味する。

大局的な映像構造の利用

シーンを同定する場合を考える。あるシーンにおけるカット群において、このカット群のなかにそのシーンとは判断しにくいカットがある場合、シーンを判断できるカットからシーンの同定を行う必要がある。例えば、キャスターの場面と現場リポートの場面が繰り返されるようなニュース番組の場合、現場リポートの場面からはニュース番組とは判定しにくいため、いくつかのキャスターの場面を利用してニュース番組と同定しなければならないというようにである。画像の連続性を重視する局所的な映像構造の利用に対し、大局的な映像構造の利用とは、画像の大局的な意味での一貫性を考慮、利用することを意味する。

4.6.3 映像のシーン/カットの同定手法

映像の構造的特徴の利用の方法としては、シーンやカットの出現確率や順序関係、そして、連続性の関係を利用したものなど、様々な形態のものが考えられるが、ここでは、提案した画像のシーン/カットの同定手法とともに、前述した映像の構造的特徴を利用することで、映像のシーン/カットの構造的な類似度を算出し、映像のシーン/カットの同定を行

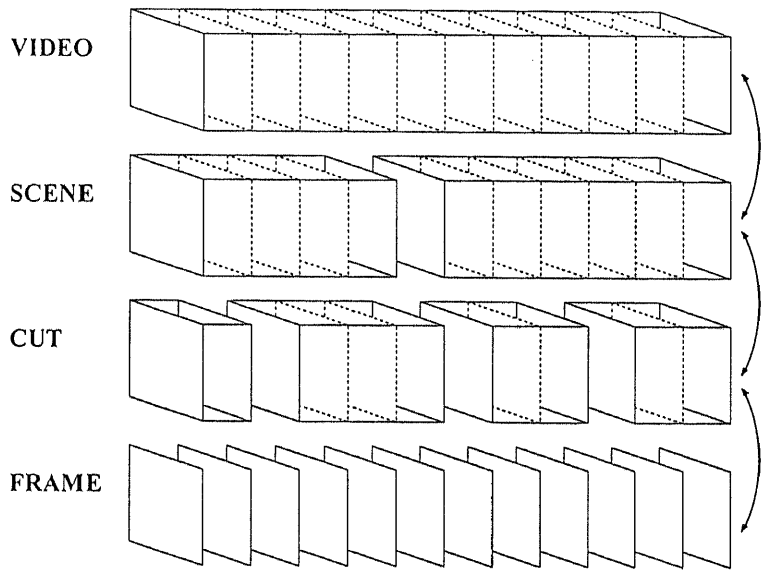


図 4.18: 映像の階層構造

う手法を提案する。

提案した画像のシーン/カットの同定手法では、類似画像検索を利用することで画像のシーンの同定を実現した。このため、同定を実行することにより、シーンデータベース内の各画像に対して、対象画像との類似度が計算されることになる。映像は連続する画像の群と考えることができるため、各画像に対しシーン/カットの同定を行えば、映像を構成する各画像に対する類似度を連続的に求めることができる。この連続的に求められる類似度を、映像の構造的特徴を用いて評価することで、映像として捉えた場合のシーン/カットの類似度を算出する。

いま、与えられた映像を構成する画像を $I(t),(t = 1,2,\cdots)$、同定を行うことで得られる、画像 $I(t)$ のシーンデータベースに蓄積されたシーン $M_i,(i = 1,2,\cdots,I)$ におけるカット $M_{ij},(j = 1,2,\cdots,J_i)$ を構成する画像 $M_{ijk},(k = 1,2,\cdots,K_{ij})$ との類似度を $S_{ijk}(I(t)),(S_{ijk} \in (0,1))$ とする。また、与えられた映像において、カットチェンジの検出により、カットチェンジが $t = T_c,(c = 1,2,\cdots)$ で検出されたとする。このとき、与えられた映像の c 番目のカット $(T_{c-1} \leq t < T_c)$ のシーンデータベースで定義されるシーン M_i

におけるカット M_{ij} との類似度 $S_{ij}, (0 \leq S_{ij}(c) \leq 1)$ を以下のように定義する。

$$S_{ijk}(I(t)) = \frac{1}{1 + \alpha d_{ijk}(I(t))} \tag{4.20}$$

$$S_{ij}(c) = \frac{\sum_{t=T_{c-1}}^{T_c} (\vee_{k=1}^{K_{ij}} S_{ijk}(I(t)))}{T_c - T_{c-1}} \tag{4.21}$$

ここで、d_{ijk}は画像の類似距離、αは類似度へ変換する際の係数である。

この類似度計算は、局所的な映像構造の評価にあたり、あるカットの映像において連続して得られる画像は、シーンデータベース内の画像のうち、そのカットをあらわす画像に多く類似するはずであるとの特徴を利用している。つまり、シーンデータベース内のカットのうち、与えられた画像群と類似する画像が多く属するカットほど、高い類似度が付与される。図4.19に映像におけるシーン/カットの類似度計算の枠組を示す。リアルタイムで映像を観測しながらカットの類似度を求める場合、現在の映像のカットは現在観測されている画像が終端であると仮定し、計算を行えば良い。

与えられた映像のシーンのシーンデータベースで定義されるシーン i との類似度 $S_i, (0 \leq S_i \leq 1)$ は、n 個のカットが得られている場合、カットの類似度の場合と同様に以下のように定義する。

$$S_i = \frac{\sum_{c=1}^{n} (\vee_{j=1}^{J_i} S_{ij}(c) \times (T_c - T_{c-1}))}{T_n} \tag{4.22}$$

リアルタイムで映像を観測しながらシーンの類似度を求める場合、カットの類似度の算出の場合と同様に、現在の映像のカットは現在観測されている画像が終端であると仮定し、計算を行えば良い。

次に、大局的な映像構造の評価について考える。ここまで述べた類似度は、連続して得られる画像がすべてシーンデータベースで定義される同一のカット、あるいは、同一のシーンに当てはまるはずであるとの仮定にも基づいていた。しかし、前述したようにニュース番組などではうまくシーン/カットが同定できない映像が混在していたり、また、シーンデータベースの不完全性のため、映像のシーンやカットが完全には定義されない場合も数多く

存在する。このような場合、先に示した類似度計算では、適切なシーン/カットに対する類似度は、低い値となってしまう。

そこで、この問題を解決するため、対象映像に対してシーンデータベースで扱えるシーン/カットの割合をあらわすカバー率 Pという値を導入する。これは、シーンデータベース内の画像が、どの程度、映像のシーンやカットを代表できているかをあらわすものである。シーン M_iにおけるカット M_{ij}のカバー率を $P_{ij},(0 \leq P_{ij} \leq 1)$、シーン M_iのカバー率を $P_i,(0 \leq P_i \leq 1)$ とすると、式 (4.21) と式 (4.22) から次式を得ることができる。

$$S_{ij}(c) = \frac{\sum_{t=T_{c-1}}^{T_c} (\vee_{k=1}^{K_{ij}} S_{ijk}(I(t)))}{(T_c - T_{c-1}) \times P_{ij}} \wedge (\vee_{t=T_{c-1}}^{T_c} (\vee_{k=1}^{K_{ij}} S_{ijk}(I(t)))) \tag{4.23}$$

$$S_i = \frac{\sum_{c=1}^{n} (\vee_{j=1}^{J_i} S_{ij}(c) \times (T_c - T_{c-1}))}{T_n \times P_i} \wedge (\vee_{c=1}^{n} (\vee_{j=1}^{J_i} S_{ij}(c))) \tag{4.24}$$

これは、カバー率 Pの割合でうまく同定ができるはずであるということをあらわしている。ここで、注意が必要なのは、算出される類似度からみると、計算式に簡易的にカバー率 Pを導入したため、シーンの類似度においては、カットの類似度が低い場合と同定できるカットが少ない場合の区別が、また、カットの類似度においては、画像の類似度が低い場合と同定できる画像が少ない場合の区別がつかない。このため、カバー率 Pの設定値によっては類似していないにもかかわらず、類似度が高くなる可能性がある。そこで、ここでは、カットや画像の類似度の最大値を上限とすることで、不必要な類似度の上昇を防ぐようにしている。

4.6.4 映像のシーン/カットの同定の実験

サッカー映像におけるシーン/カットの同定の実験を行った。

図 4.20に入力フレーム数と算出されるシーン/カットの類似度の関係を示す。これは、1秒おきにとったサッカー画像 400 枚に対し、シーンの類似度を計算したものである。

この結果から、先に示した画像のシーン/カットの同定の場合と同様に、サッカーとゴルフの画像は類似しているため、両者の類似度が高くなってしまっていることがわかる。これにより、より正確な認識を行うためには、第 2 章で示したような所望のシーンを記述で

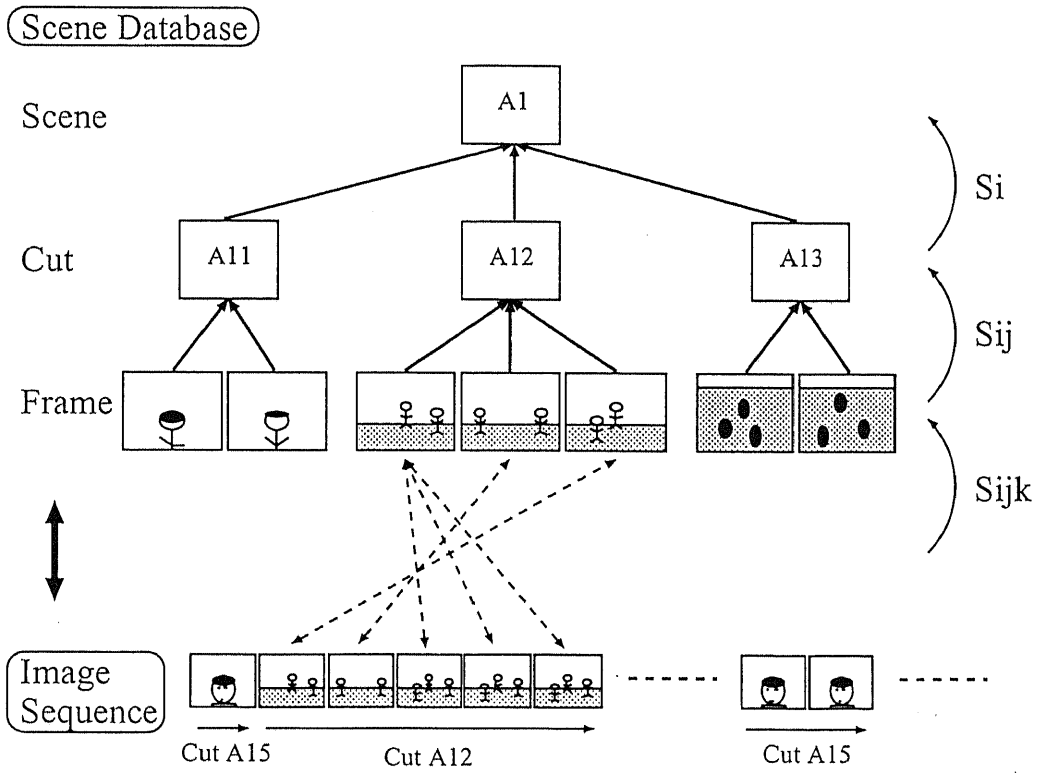


図 4.19: 映像におけるシーン/カットの類似度計算の枠組

きるVSDLなど他の手法との相互利用を検討する必要があるということになる。

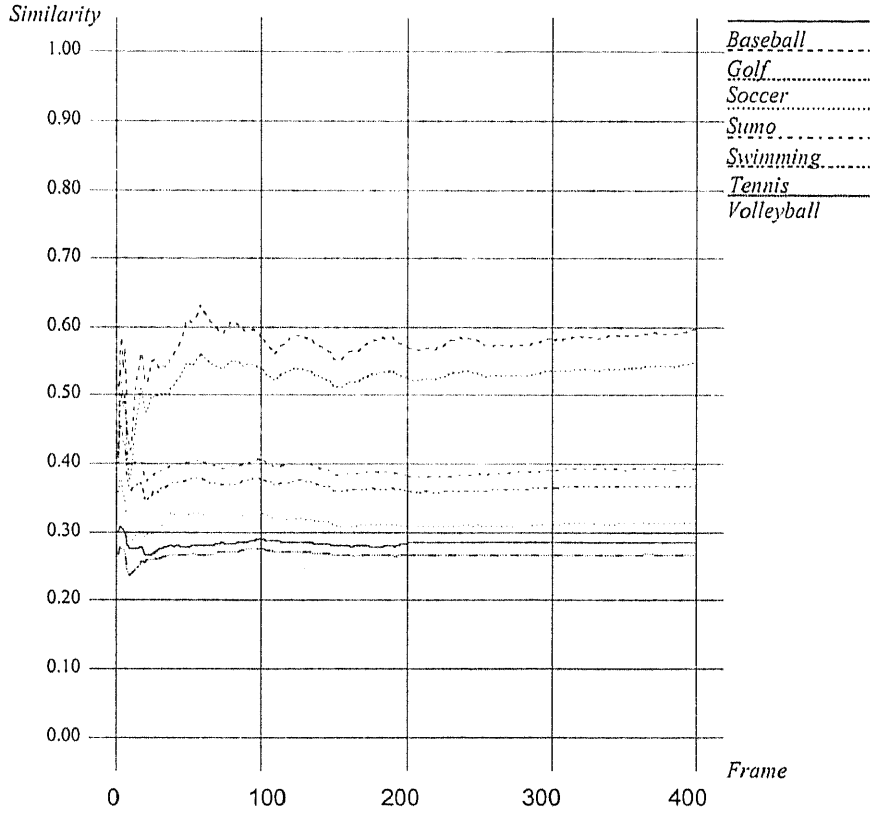


図 4.20: 入力フレーム数とシーン/カットの類似度

4.7 本手法のテレビ映像への適用の可能性

本手法はテレビ映像など多様な画像への適用を目指して提案したものである。スポーツ画像に対する実験においては、特徴量の選択を含め、検索手法が十分でない面もあり、十分な結果が得られているとはいいがたい。しかし、シーンデータベースの構成方法や検索手法を更に検討、改良を加えることで、能力の向上を期待することができると思われる。

ここで、本手法のテレビ映像への適用について考えてみる。テレビ番組は、ニュース、ドラマ、映画、スポーツ、音楽などの番組から構成されている。本手法は、画像の画面構成

がほぼ定まっており、同類の番組において繰り返し類似の画像が得られる場合に適用可能である。このため、テレビ番組へ適用する場合、前述のような条件をみたすものでなければ対応できない。ニュースの場合、定まったキャスターの場面がキーポイントとなり同定できる可能性がある。ドラマや映画では、対象とする映像が幅広く適用が困難と思われるが、番組のオープニングやエンディングがいつも同じであることに注目すれば、番組の同定も可能であるといえる。スポーツについては、ほぼ画面構成が定まっており、背景が類似しているため同定できる可能性が高い。このように考えると、テレビ映像に対して、困難な面もあるが十分対応できる面もあるといえる。

ところで、テレビ番組の大きな問題は番組改編が3ヶ月単位で行なわれることにある。このため、シーンデータベースが対応できなくなる可能性があるが、画像を集めればシーンデータベースを構成できるため、この番組改編時にシーンデータベースを更新、再構成すれば対応できるといえる。

このように考えていくと、検索手法の高度化により、テレビ映像への対応は十分意識できる手法であることがわかる。

4.8 まとめ

本章においては、画像・映像のシーン/カットの認識を実現するため、画像データベースと類似画像検索技術を用いたシーン/カットの同定手法を提案した。

画像・映像情報の増加に伴い、利用者の視点での画像や映像の選択や利用を可能とする画像認識技術への要求が高まっている。また、画像認識技術で多様な画像を扱う場合、その画像が何を示しているかを知ることは重要である。これは、その画像が何を示しているのかわからなければ、その画像からどのような情報を取得できるのかがわからない上に、どのような手段により必要な情報を獲得すればよいのかもわからないためである。このような要求に対して、画像が何を示しているのかの認識を実現するシーン/カットの同定手法は有効となる。

提案した手法の特徴は、シーン/カットの認識のために、画像に映っている物体などを直接あらわす知識を必要としないことにある。一般に、画像のシーン/カットは、画像に映っている物体や、それらの組み合わせや配置で表現される。このため、画像のシーンを認識

するためには、多くの知識が必要となる。特に、多種類の画像を扱おうとする場合、あらわれる可能性のある多くの物体のモデルを用意しなけらばならないなど、必要とする知識は膨大なものとなる。本手法の適用においては、オブジェクト認識のための知識が必要なく、実例であるシーン/カットを代表する画像そのものを用意すればよい。このため、テレビのような多様な映像を扱う場合でも、必要なシーン/カットの画像を集めるだけでよいため、容易に適用できることになる。

映像のシーン/カットの特徴を記述し、映像フィルタリングを行なうものとしては、VSDL(Video Scene DescriptionLaunguage)[59] を用いたシステムなどが開発されている。これは、利用者が必要に応じてシーン/カットを定義できるので、利用者の意図にそった形でシーン/カットをモデル化できる。一方、本手法では、画像を集め、画像から自動的に抽出できる特徴量を利用してシーン/カットをモデル化する。このため、利用者の意図をどこまで反映できるかは検討を要する。また、VSDL では人手によりモデルを構築するため、広範囲な種類のシーン/カットを扱うことは困難であるが、本手法では画像データベースを構築するだけで良いため、対象とするシーン/カットの拡張が容易である。ただし、あまり、多くの種類を対象とすると、シーン/カットの特徴がうまく抽出、利用できず、識別能力の低下を招く。これは、例外的な画像をさけ、代表的な画像を集めることが必要であることを意味している。

本手法では、類似画像検索技術を用いておりシーン/カットの同定を実現している。類似画像検索技術については、現在までに多くのものが検討されシステムの開発が行なわれているが、いまだ、類似性の評価基準、特徴量の選択基準など多くの問題が残されている。このため、これらの課題を解決していくことが、本手法の実現性を決定づける大きな要因の一つになると考えられる。

第 5 章

撮像距離に依存しない認識を可能とする階層型距離モデル

5.1 概要

画像認識は認識対象に関する知識をもとに行われる。このため、認識対象をモデル化する場合、その画像への映り方が重要となる。これは、認識対象の映り方により、対象を認識するために利用できる特徴が異なるためである。したがって、対象の映り方の違いによる特徴の変化を考慮したモデル化手法が必要かつ重要となる。

現在まで、様相をモデル化するなど、対象の映る向きの違いについて検討したものはあったが、距離方向、つまり、特徴の変化を伴う対象の映る大きさの違いについてはほとんど考えられていなかった。これは、今までの画像認識においては、特定の画像、特定の対象の認識を意識しており、対象の映る大きさの違いによる特徴の変化をさほど意識する必要がなかったためである。しかし、テレビ映像などを扱う場合、撮像距離、カメラワークなどの関係から認識対象は様々な大きさで映っていることが考えられ、これらは特徴の変化を伴うことから、対象を同一のモデルや認識手法で扱うのは困難である。そこで、ここでは、このような特徴の変化を伴う対象を統一のモデルで表現する一つのアプローチとして、階層型距離モデル [8]-[10] の考え方を提案する。

階層型距離モデルとは、近景や中景、遠景の場合で認識に用いる特徴が変わることに注目し、これら特徴の違いをうまく階層的にモデル化することで、対象の映る大きさに依存しない認識を試みようとするものである。具体的には、テレビ画像の認識で特に重要となる人間の認識を例にあげ、人間の映る大きさを階層的にモデル化し、そのモデルを統合的に利用することで、映る大きさの変化に柔軟に対応する手法の検討を行なった。

本章では、撮像距離による対象の映り方の違いを表現できる階層型距離モデルについて述べる。

5.2 階層型距離モデルによる認識の枠組

5.2.1 階層型距離モデル

一般の画像において、認識対象は、カメラパラメータの違いなどにより、様々な大きさで映っている可能性がある。認識対象が大きく映っている場合には、その対象の局所的特徴が、また、小さく映っている場合には、その対象の全体的特徴や周囲状況が、その認識対象を表す特徴となる。このため、対象を認識するためのモデルは、認識対象の映る大きさ

別に用意する必要がある。例えば、画像に映る人間の大きさは、カメラの焦点距離を固定とした場合、距離の二乗に反比例する。このため、カメラの近くにいる人間は大きく細部まで映るため、顔を構成する目や口の検出により人間であることを認識できる。一方、遠くにいる人間は小さく細部まで映らないため、場との関係や人間の全体像や動きの検出により人間であることを認識する必要がある。そこで、これら特徴の違いをモデル化するため、階層型距離モデルの考え方を提案する。階層型距離モデルとは、認識対象をあらわす特徴の違いを画像に映る大きさ別に記述し、これらを一つの階層関係で捉えることで、映る大きさに依存しない認識を実現しようとするものである。この階層の数は、対象画像、認識対象の特徴により決定すればよい。ここでは、人間の認識を例に取り上げ、近景、中景、遠景の場合の三階層でモデル化することにした。ここで、各階層のモデルの総称を距離モデル、一つ一つのモデルを近景モデル、中景モデル、遠景モデルと呼ぶことにする。人間の場合の各距離モデルが対象とする画像を図 5.1、各距離モデルのモデル化の関係を図 5.2 に示す。各階層でモデル化するべき特徴は以下のようになる。

- 各距離モデルにおけるモデル化特徴 (人間の場合)
  - 近景モデル → 人間頭部のモデル
  - 中景モデル → 人間全身のモデル
  - 遠景モデル → 場における人間のモデル

5.2.2 階層型距離モデルの構成と階層間の協調

階層型距離モデルにおいては、認識対象はいくつかの距離モデルにより階層的にモデル化される。これら各距離モデルは、映る大きさ別に認識対象の特徴をモデル化したもの、つまり、同一対象の異なった面からみた場合の特徴をモデル化したものである。対象の映る大きさに依存しない認識を実現するためには、これら各距離モデルであらわされる特徴を相互に有効に活用することが必要となる。これは、人間の認識の場合、人間が全身で映っている場合でも、中景モデルだけでなく、場との関係をあらわす遠景モデルや頭部の詳細をあらわす近景モデルの適用が可能であり、より多くの特徴が抽出できれば、それだけ人間としての認識度をあげることが可能であるということである。このような各距離モデル

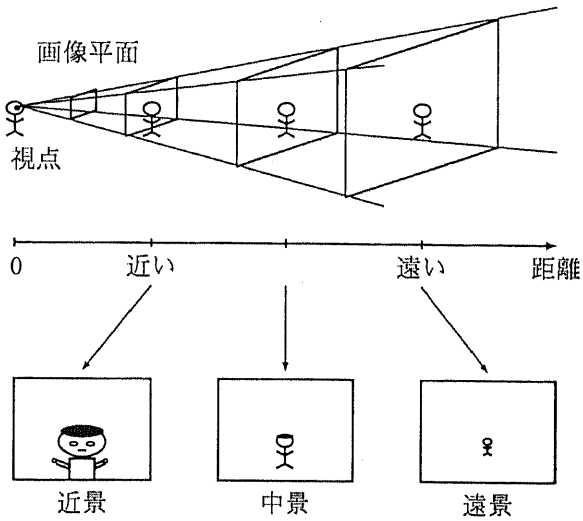


図 5.1: 各距離モデルの対象画像

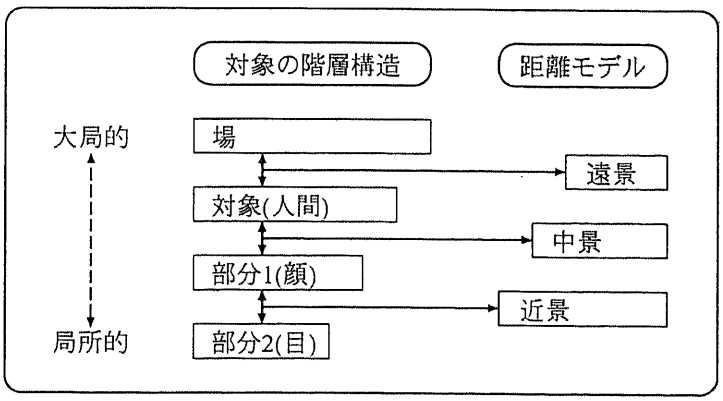


図 5.2: 距離モデルの階層関係

を相互に利用する手段として、階層型距離モデルでは、以下に示す二つの観点から階層間の協調を実現する。

- 各距離モデルの相互駆動

  ある距離モデルにおいて、何らかのオブジェクトが認識されたとする。この場合、他の距離モデルにおいてそのオブジェクトの認識が行えないような場合でも、その距離モデルにおいて、同じオブジェクトを表しているとの観点から、評価することは可能である。例えば、近景モデルにより、人間の頭部が抽出されたとすると、その領域に対して、人間全身のモデルである中景モデルをトップダウンに適用することで、頭部だけでなくその周りに人間の体らしいものがあるかどうかを判定することができる。これは、近景モデルの結果から中景モデルを駆動させることを意味する。これは、各距離モデルにおいて、同一オブジェクトを関連づけることで実現する。

- 各距離モデルによる認識結果の統合

  各距離モデルの相互駆動により、あるいは、独立に適用することにより、各距離モデルによる認識結果が同一領域に対して求まることになる。これらの結果をうまく統合すれば、映る大きさによらない認識が可能となるはずである。ここでは、各距離モデルにおける認識結果を確信度という形で表し、それら確信度の統合をはかることで実現する。

したがって、階層型距離モデルによる認識の枠組は図 5.3に示すようなものとなる。これは、各距離モデルは独自に認識を実行するが、必要に応じて各距離モデル間でモデルの駆動を制御し、最終認識結果は、各距離モデルによる認識結果の統合をはかることで求めることをあらわしている。このときの階層型距離モデルの構成を図 5.4に示す。図中の各階層間のリンクはモデルの駆動関係をあらわす。

5.2.3 距離モデルの構成

各距離モデルは大きく二つのモデルで構成される。一つは、認識を実行する認識モデル、もう一方は、他の距離モデルからのリンクにより駆動される評価モデルである。認識モデルは、各距離モデルにおいて利用できる特徴から独立に認識を実現するものであり、評価

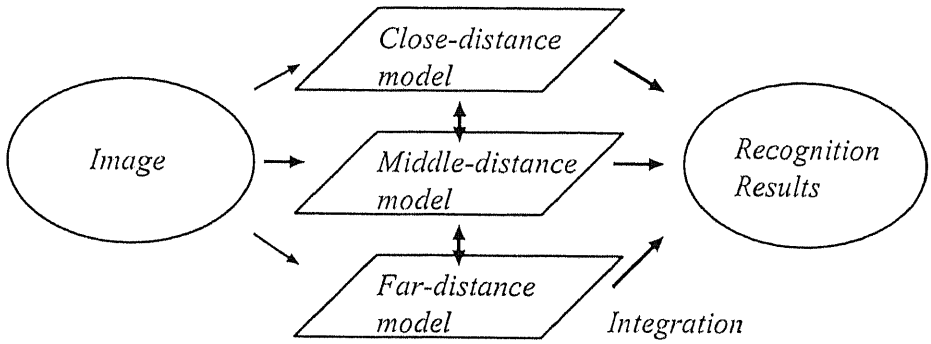


図 5.3: 階層型距離モデルによる認識の枠組

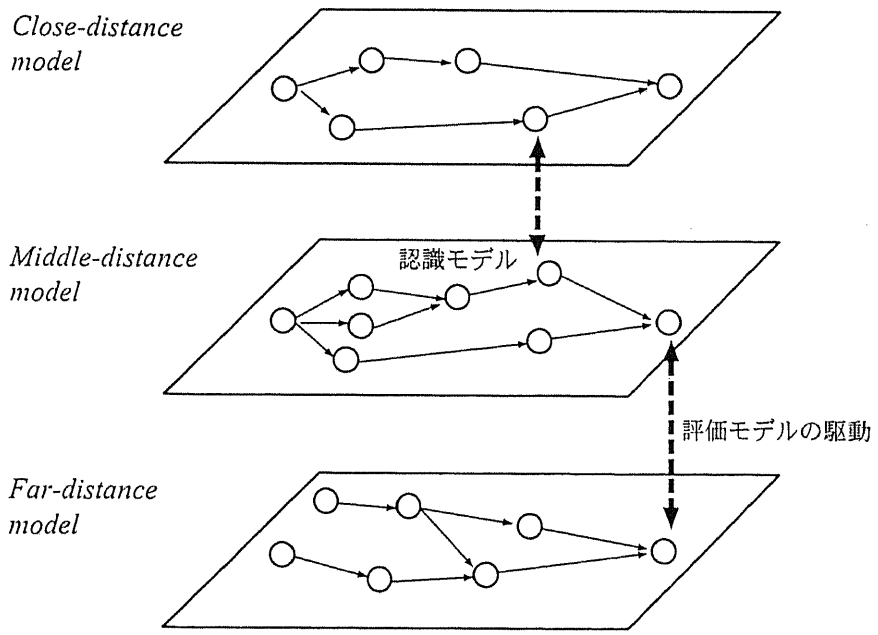


図 5.4: 階層型距離モデルの構成

モデルは、他の距離モデルで認識された領域に対し、その距離モデルで利用できる特徴からみた場合の評価を実現するものである。図 5.5に示す。また、各距離モデルは基本的に独立した状態遷移モデルの枠組で構成され、認識時には条件判断部におけるパラメータを評価することで、モデルに対する満足度が同時に算出される構成をとる。これにより、最終的な距離モデル同士の認識結果の確信度をとおしての統合が可能となる。

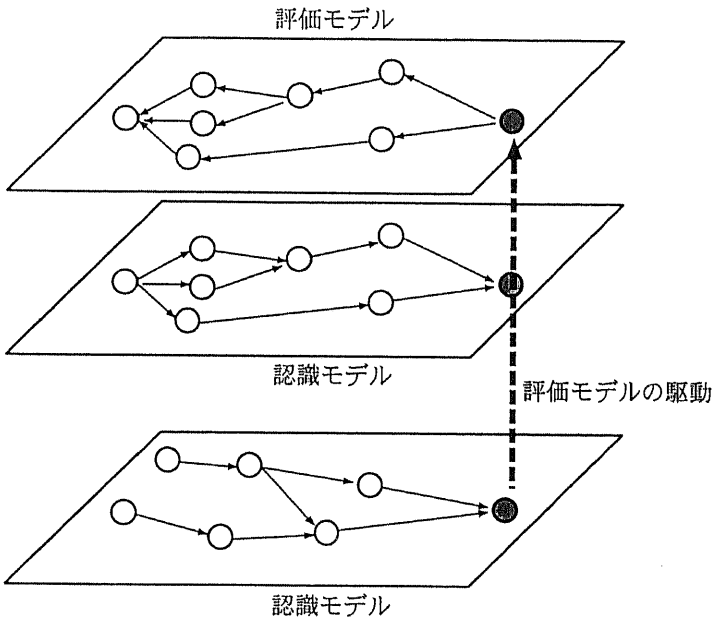


図 5.5: 認識モデルと評価モデル

ここで、人間の認識の場合の簡単なモデルの例を示す。

- 近景モデル

  人間の特徴的な肌色を利用して、人間の頭部候補領域を抽出する。エッジ分布から、目、口領域の抽出を行ない、各候補領域の確信度を求める。

- 中景モデル

  シーンが決定されていると仮定し、サッカーの場合、フィールド領域を分離し、人間候補の領域を抽出する。また、肌色を利用して頭部領域、手領域を抽出する。場にお

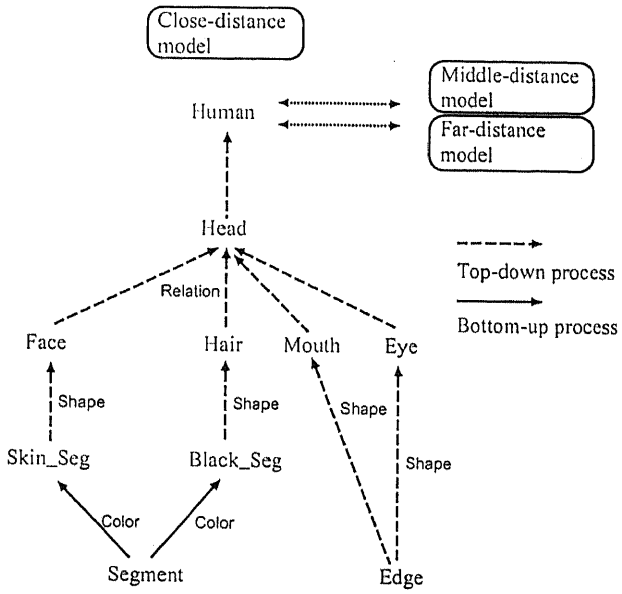


図 5.6: 近景モデルの例

ける位置、及び、形状が縦長であることなどから評価を行い、各候補領域の確信度を求める。映像の場合には、外接長方形の変化を利用する。

- 遠景モデル

  中景の場合と同様、人間候補の領域を抽出する。場における位置から、各候補領域の確信度を求める。映像の場合には、画面内での動き情報を利用する。

5.2.4 階層型距離モデルの特徴

階層型距離モデルの特徴としては、次のような特徴がある。

- 認識モデルの個別化

  階層型距離モデルでは、近景、中景、遠景の各モデルを階層的にあらわすことで、映る大きさ別に抽出すべき特徴を明示的に与えることが可能である。これにより、階層的な関係の中で特徴の評価を行なうのみでなく、候補領域の抽出の方法も別々に記述できるため、柔軟な領域抽出、認識が可能となる。これは、各モデルの独立性を高くし、人間にとって定義しやすいものとなる。例えば、人間の抽出、領域特定を行なう

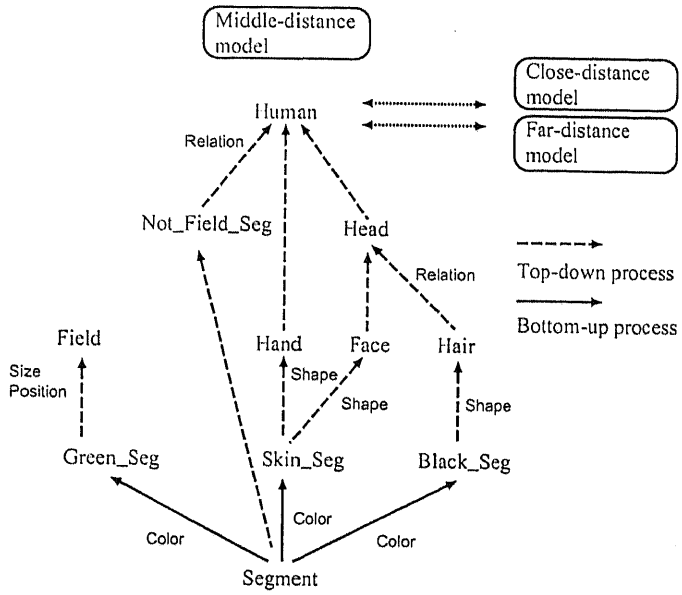


図 5.7: 中景モデルの例

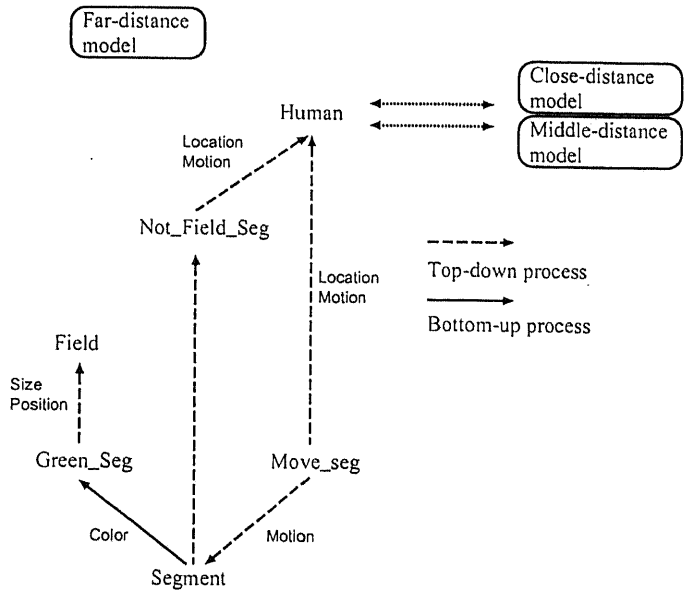


図 5.8: 遠景モデルの例

場合を考えてみると、色情報を中心に行なう場合や、動き情報から行なう場合、あるいは、細部の構造的関係から行なう場合などがあり得る。これらは、人間の画面上への映る大きさにより決定される部分であり、各距離モデルに明示的に記述できることは重要である。

- 階層間の協調による評価

  各階層は基本的には独立に認識を受け持つが、各距離モデルにおいて、同じオブジェクトをあらわす部分同士を関連づけることにより、ある距離モデルで認識された領域に対し、他の距離モデルでの評価を行なうことが可能である。これにより、すべての距離モデルが対象領域を切り出す能力を持っていなくても、一つのモデルで切り出すことが可能ならば、その切り出された領域に対して、他のモデルを適用することが可能となる。これは、人間の領域を切り出すのに肌色情報だけを利用する場合、遠景では肌色情報だけからでは切り出しが無理であるが、他のモデルで抽出された領域に対しては、遠景モデルの内容で評価が可能であるということである。

5.3 認識の枠組

認識は、図 5.3に示す枠組に従って、以下の手順で実行する。

はじめに、それぞれの距離モデルでの認識において、各知識の満足度を証拠として考え、{人間である、人間ではない} の各確率を求める。次に、この確率を図 5.9に示すようなモデル適合度を考え、各モデルの認識能力を考慮した確率へ変換する。

ここで、モデル適合度とは、各モデルの認識の有効度を表すもので、人間が大きく映っている場合には近景モデルが適合しなければならない、人間が小さく映っている場合でも遠景モデルのみで確実な人間抽出はできないということを示すものである。これは、近景モデルは、対象の局所的特徴をモデル化するため、一般に映る大きさに依存せず、対象を確実に認識するために必要な知識となり、遠景モデルは、対象の場との関係をモデル化するもので、これのみでは確実な認識は困難であるということである。

モデル適合度を考慮した新しい確率は、{人間である、人間ではない} の各確率に、適合度を乗じることで求める。そして、変換後の確率に対し Dempster の結合法則を適用することで、最終結果を求める。

1. 各距離モデルにより認識を行い、各モデルの満足度を求める。

2. ある距離モデルにおいて、他のモデルが駆動される場合、その候補領域に対して、駆動されたモデルがどれくらい適するかを計算する。

3. このときの評価値は、駆動されたモデルの満足度として保持する。

4. 同一領域に対する各距離モデルの満足度を、図5.9に示すうなモデル適合度を用いて、映る大きさに対するモデル認識能力を考慮した確信度へ変換する。

   ここで、各距離モデルによる満足度をm_1, m_2, m_3、モデル適合度を$\mu_{m1}, \mu_{m2}, \mu_{m3}$とする。$(0 \leq m \leq 1, 0 \leq \mu \leq 1)$

$$m_i' = \mu_i \cdot m_i \qquad (i = 1, 2, 3) \tag{5.1}$$

5. 変換後の確信度を Dempster の結合則により統合し、最終結果を求める。

$$m_{12} = m_1' + m_2' - m_1' m_2' \tag{5.2}$$

$$m_{123} = m_{12} + m_3' - m_{12} m_3' \tag{5.3}$$

5.4 認識例

先に述べたモデルにより人間の認識を行なった結果の例を図5.10に示す。一つは、人間がある程度大きく映っているため、顔の領域やパーツがはっきり評価、認識されている例であり、もう一つは、全身が映っておりフィールドとの関係や外接長方形などの情報から評価している例である。また、3つ目のものは、人間が小さく全身映っているもので、動きのパラメータを評価して認識を行なっている例である。この結果では、階層型距離モデルの効果がわかりにくいが、簡単な例ながらも、人間の認識が実現できていることはわかる。人間の映る大きさが連続的に変わっても、この方法では十分対応できる。ただし、本モデル化が有効に働くためには、近景モデルでは、顔の向きの変化に対するモデルを、ま

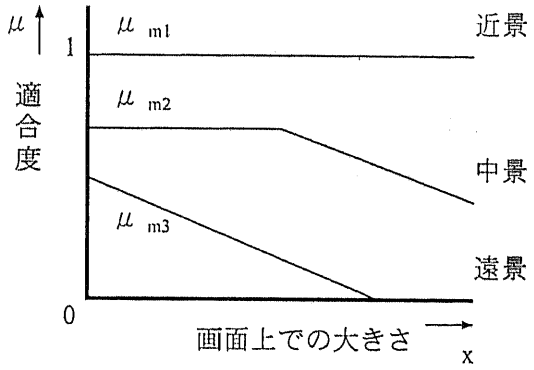


図 5.9: 映る大きさとモデル適合度

た、遠景モデルでは、多くの場に対応できるモデルを作成できるかが重要なポイントとなるといえる。

5.5 同一場面の一連の画像群の認識

階層型距離モデルによる認識は、認識対象の映る大きさ別に設定した複数のモデルそれぞれの認識結果の統合により実現される。同一場面の一連の画像群を認識するような場合でも、次の画像の認識にそれまでの画像の認識結果から獲得できる情報を利用することができない。そこで、階層型距離モデルによる同一場面の一連の画像群の認識の方法について、各距離モデル間での認識知識の共有との観点から、検討を行なった。

同一場面の一連の画像群の認識では、これらの画像群の認識に共通に利用できる認識知識の獲得と利用ができることが望ましい。そこで、各距離モデルを、認識対象を認識するための一般的知識と、同一場面の一連の画像群の認識において利用する補助的知識の二つの知識により記述することにする。

- 一般的知識

  人間の形状や色などの固定的な認識知識で、利用手順とともに記述する。

- 補助的知識

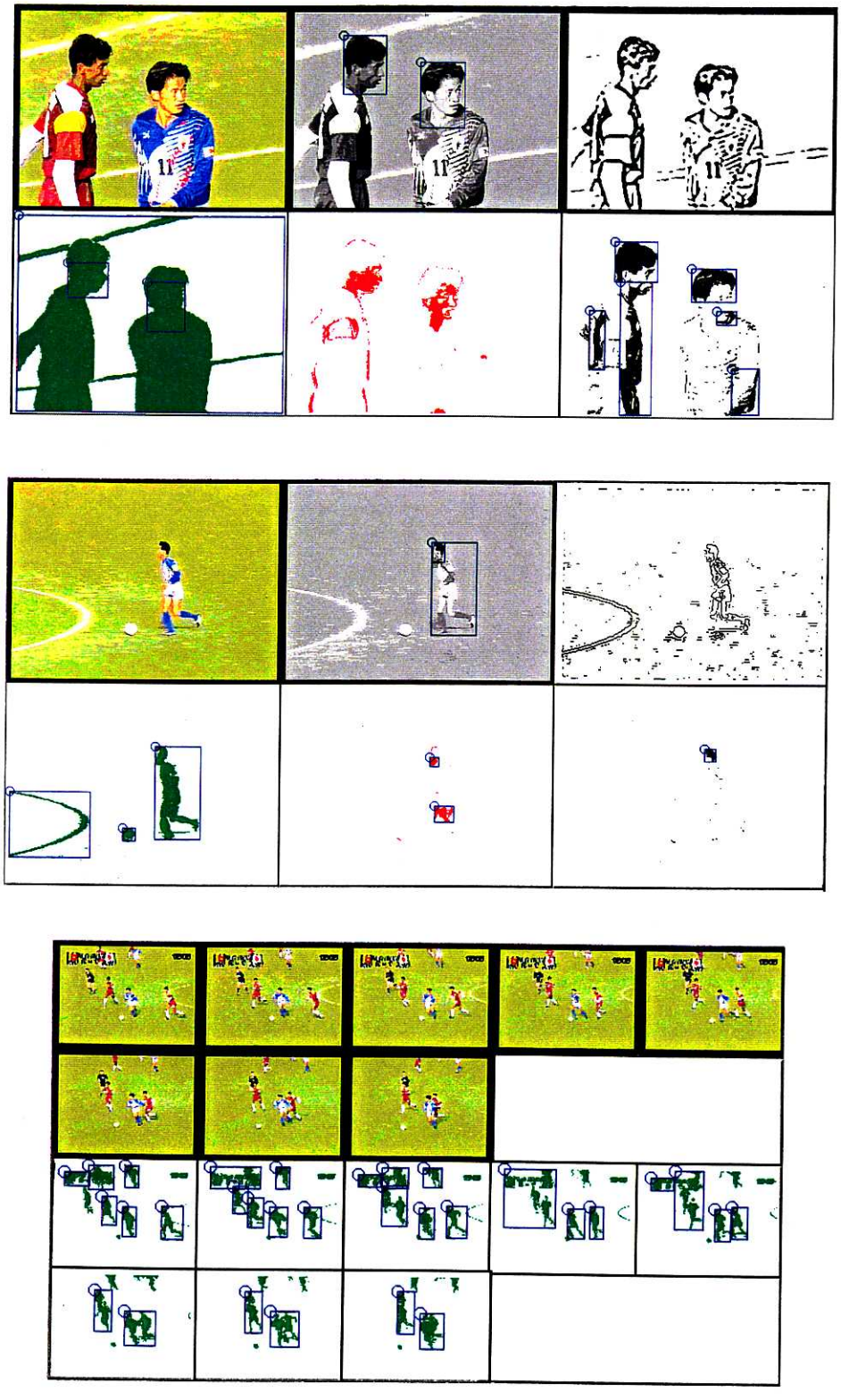


図 5.10: 認識結果の例

服装の色など認識時に画像から獲得する一時的な認識知識で、獲得、利用手順とともに記述する。

補助的知識は、認識対象が認識できた画像から獲得手順に従って獲得し、各距離モデルで共用することで、次の画像の認識に利用する。認識結果から獲得できる情報は、認識の高度化、シーン理解の観点から、次に示すような補助的認識知識や認識対象の対応づけなどでの利用が考えられる。

1. 獲得情報による対応づけ

   例えば、図5.11に示すように、近景画像で人間の認識ができたときに、それと同時に服装の特徴などを獲得し、次に続く中景画像の認識において、その服装の特徴を利用できれば、より確かな人間の認識が可能となる。また、この服装の特徴を用いることで、近景画像の人間と中景画像の人間の対応づけを行なうことも可能となる。

2. ズーミングによる対応づけ

   例えば、図5.12に示すように、ズーミングされた画像において人間領域を追跡することで、近景画像の人間と遠景画像の人間の対応づけが可能となる。これにより、近景画像で得られた詳細情報を遠景画像へ対応づけることが可能となる。

ここで、図5.13に簡単な実験例を示す。これは、前画像から服装の色情報を獲得し、次画像でどの程度、利用できるか示したものである。

5.6 カットモデルの利用

テレビ画像などの作為的な画像では、演出意図やカメラ位置の制約などから、認識対象の映り方や位置関係などカットの構図がほぼ決まっている場合が多い。このため、カットの構図から、場や認識対象の認識、評価を行なうことが可能となる。そこで、シーンの状況モデルとして、場の知識などを含めたカットの構図に関する知識を用いた認識について考えてみる。

カットモデルとは、画面上での各距離モデルにおける人間の配置関係、及び、背景などの周囲環境を記述するモデルである。このモデルは、距離モデルの上位に位置し、各距離モ

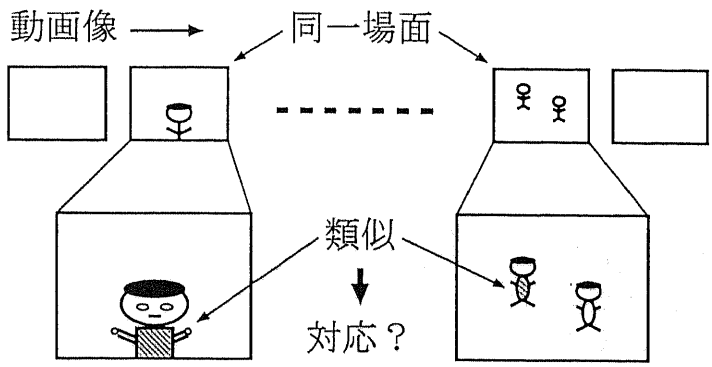


図 5.11: 獲得情報による対応づけ

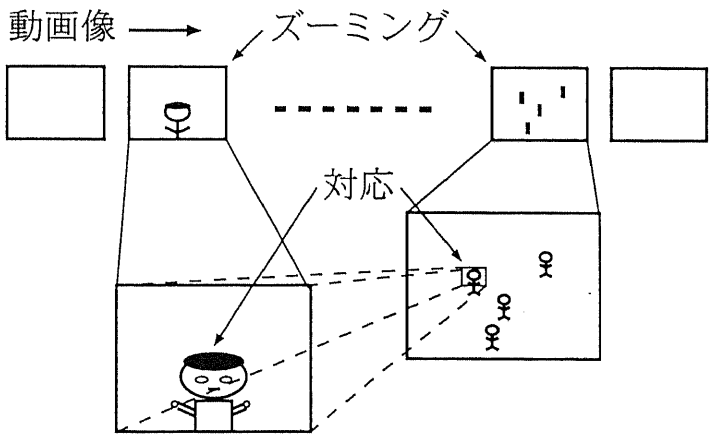


図 5.12: ズーミングによる対応づけ

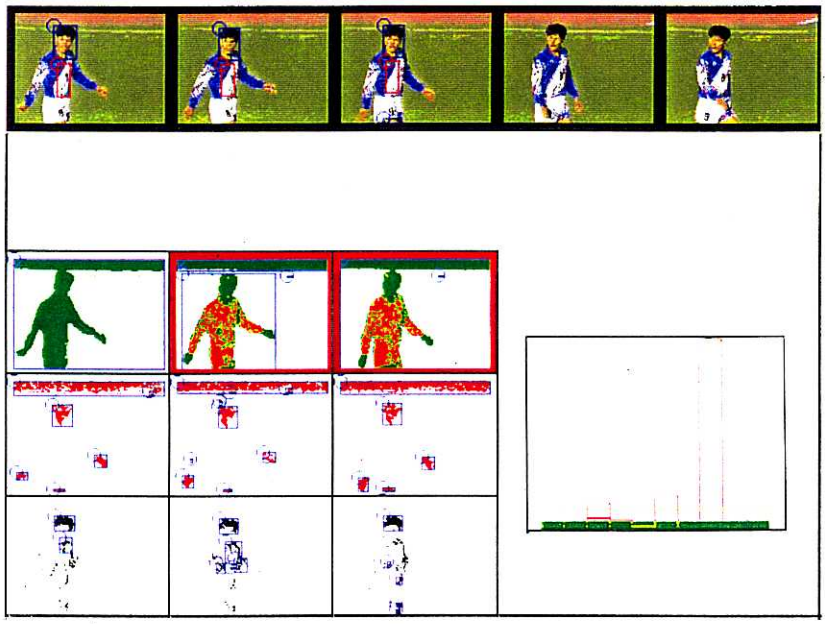


図 5.13: 獲得情報の利用例

デルの処理結果の統合管理を行なう。サッカーシーンの場合、人間の配置関係とは「フィールド上に選手は存在する」、「画面の前面下ほど近景の人物がいる可能性が高い」「フィールドの周囲の観客席に群集は存在する」などの知識で記述する。図5.14に示すようなカットモデルを定義する。このモデルの考え方は、第4章で述べた画像データベースを用いた画像認識に準ずるものである。

1. カットの同定するための特徴量

   「まとまったグリーン領域が存在」

2. 場のモデル

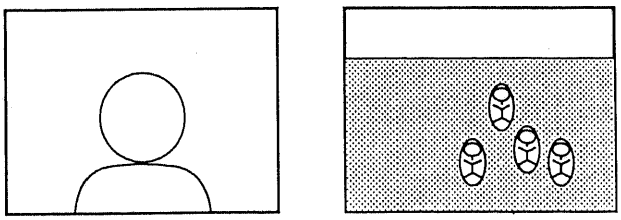
   「まとまったグリーン領域はフィールド」

3. 認識対象の映る大きさ、位置、数の関係

   「遠景の人間が複数存在」

4. 場と認識対象の関係

   「フィールド上に人間は存在」



(a) カット1 (b) カット2

図 5.14: カットモデルの例

認識は、次に示すプロセスにしたがって行なう。はじめに、画像を直接評価し、カットの同定を試みる。カットの同定ができる場合には、そのカットモデルにより場などのラベル付けを行なう。次に、ラベル付けの結果を用いて、距離モデルによる認識を実行する。こ

れにより、認識対象の候補領域を求める。最後に、求めた候補領域とカットモデルとの整合性をチェックし、必要ならば候補領域の再探索を行ない、最終的な認識結果を求める。

具体的には、カット2のような画像の場合、最初に、まとまったグリーン領域を評価することで、モデルのカット2を同定し、フィールドのラベル付けを行なう。次に、距離モデルにより人間の認識を試みる。ここで、フィールド領域の情報から遠景側の人間の認識が容易になる。そして、最後に、候補領域とモデルとの整合性をチェックし、候補領域の評価値を調整する。基本的に、抽出された候補領域の評価は、すべて各領域に対する評価値の調整により行なう。

5.7 まとめ

本章では、階層的に認識対象をモデル化することで、認識対象の映る大きさに依存しない認識を試みる階層型距離モデルの考え方を提案した。

現在までに、認識対象の映る大きさの変化に対応するためのモデル化手法としては、モデルをパラメータで定義するなどの方法も検討されてきたが、これはパラメータにより対象をあらわすモデルの大きさを調整するものであり、特徴が変化するもの、つまり、異なる特徴を利用しようとするものではなかった。

これに対し、階層型距離モデルは、認識対象を認識するための特徴を対象の映る大きさ別にモデル化し、それらを関連づけることで、映る大きさに依存しない認識を実現しようとするものである。このため、映る大きさの違いによる特徴の違いをうまく利用することができる。本モデルの枠組は、認識対象を認識できる技術があることが前提となっており、複雑な認識対象の認識を実現するものではないが、ここで示した人間の認識の場合のように、様々な観点からの情報を利用することができ、より柔軟な認識が実現できる可能性がある。現在の構成では、階層間での関係が密ではないが、より高度な認識への適用を考える上では、より密な階層間での協調を検討する必要があるといえる。

第 6 章

結論

本論文では，利用者の要求に応じた高度な映像情報処理の実現との観点から、いくつかの具体的な映像情報処理事例を取り上げ、それぞれの目的や対象に応じたモデル化、及び、これらモデル化による画像認識手法の提案を行なった。

近年、テレビ放送の多チャンネル化や、ビデオデッキ、ビデオカメラなどの普及に伴い、簡単に多くの映像情報を獲得、利用できる環境が整いつつある。しかし、このような環境の出現の一方で、映像情報の増加、氾濫により、利用者がこれら多くの映像情報に埋没し、映像情報を有効に活用するのが困難な状況にもなってきている。今後ますます増加すると考えられる映像情報の海の中で、利用者がその情報の海を有効に活用していくためには、多くの情報の中から利用者が必要とする情報の選択、獲得を実現する画像認識技術の高度化が重要なカギを握ることになる。

従来の画像認識においては、工業分野での利用を前提に、特定の画像から、特定の幾何学的な物体を抽出、認識する技術の開発が中心的課題であった。しかし、マルチメディア環境における映像情報処理という立場からは、画像認識の目的は、物体の認識というよりはむしろ画像のシーンの認識、意味的内容の理解であり、利用者の視点で様々な画像を様々な形で扱える技術の開発が望まれている。このため、従来の技術とは別に、対象や目的に応じた画像認識技術を検討する必要がでてきた。これは、現在の画像認識技術や知識情報処理のレベルが人間の能力を模倣するにはほど遠く、高度な認識を実現するためには、対象の世界を限定、画像認識に要求される能力を特化する必要があるためである。そこで、本論文においては、以下に示す三つの観点から、目的や対象に応じたモデル化手法、これらモデル化による画像認識手法について検討を行なった。

はじめに、様々な対象、様々な目的に利用可能な状態遷移モデルの枠組、及び、状態遷移モデルを定義する状態遷移ルールの作成支援システムについて述べた。このモデルでは、認識対象のモデルを認識手順を表すルールという形で与えることができるため、対象の変更に柔軟に対応でき、また、ボトムアップ解析とトップダウン解析の両解析が実行できるため、効率的な認識の実行が可能である。このモデルの大きな特徴としては、これら汎用性とともに、認識過程を状態の遷移という形で表現することで、認識の結果を状態という形で保持できる点にある。このため、最終的な目標まで認識が実行できなくても、認識できたところまでの結果を利用することができる。これは、データベースシステムなどでの利用に有効に働く。第3章では、地図図面の認識を例に示したが、ルールの作成の仕方に

より、画像や映像のシーンの認識や、画像からの物体の認識など様々に利用でき、これらを組み合わせることで、物体からシーンにいたるまで一つの状態遷移の枠組の中で表現、モデル化することが可能である。ただし、状態遷移モデルでは、認識対象をルールの形で記述できなければならないため、ルールで定義することが困難な対象には対応できない。ところで、認識のためのルールの作成に際しては、全体の枠組は人間が容易に作成できるが、条件判断部のしきい値などの設定が困難となる場合も少なくないことが実験を通して明らかとなった。このため、条件判断に関するルールの作成を容易にするルール作成支援システムの構築を行った。このシステムでは、グラフィカルユーザインターフェースを用いて、平易な対話によりルールの生成を行うことが可能である。ルールの生成は、対話により得た情報を一般化、ルール化する帰納推論による学習アルゴリズムを用いることで実現し、地図図面の認識のためのルールの生成実験を行ない、数回程度の対話で十分実用的なルールの生成が可能であることを確認することができた。本論文で述べたルール作成支援システムでは、条件判断部を持つボトムアップルールの生成を対象としていたが、複雑な認識対象を扱う場合、対象が持つ構造がより複雑になるため、トップダウンルールの作成もより困難になると考えられる。このため、トップダウンルールの生成を含めた、より一般化した形でのルール作成支援について検討していく必要があるといえる。

次に、映像フィルタリングや画像認識への利用が可能な、画像データベースを用いた画像・映像のシーン/カットの同定手法の提案を行なった。これは、代表的なシーン/カットの画像をデータベース化しておき、対象画像との類似画像を検索することで、シーン/カットの同定を実現しようとするものである。一般に画像のシーン/カットの認識は、画像に映っている物体やそれら物体の配置関係などの情報から行われるべきであるが、物体の認識にはシーン/カットの情報を利用したい場合が少なくない。また、マルチメディア環境においては、多種多様な画像・映像を扱えることが重要となるが、それら多くの画像・映像のシーン/カットを定義したり、多くの物体のモデルを作成することは非常に困難である。その上、物体の認識を行なってから、シーン/カットの認識を行なうのでは、処理時間の観点からも問題が多い。これに対し提案手法は、判別能力に関して検討の余地があるが、シーン/カットに関する知識を実例である画像そのもので与えることができるため、容易に利用、拡張が可能で、多種多様な画像・映像に簡単に対応させることができ、また、物体の認識なしに直接、画像・映像のシーン/カットの同定を行なうことができるため、シーン カットの

情報を用いた物体の認識を実現できるなど優れた特徴を有する。ただし、画像データベースと類似画像検索技術を用いるという観点から、シーン/ カットを同定するのに必要とされる画像データベースの規模や、類似画像の定義など、検討すべき課題が多いことも忘れてはならない。

そして最後に、同じ認識対象であっても、撮影時のカメラとの距離やカメラパラメータの違いにより、様々な映り方をする点に注目し、特に映る大きさの違いに柔軟に対応するための階層型距離モデルの考え方を示した。これは、近景や中景、遠景など映る大きさにより認識に用いる特徴が変わることに着目し、これら特徴の違いをうまくモデル化することで、映る大きさに依存しない認識対象のモデル化を試みるものである。階層型距離モデルでは、映る大きさによる特徴の違いを、映る大きさ別に階層的に表現し、実際の認識は、これら複数のモデルの認識結果の統合をはかることで実現する。映る大きさ別に階層的に表現することで、認識時に対象の映る大きさによる利用特徴の選択や各モデルの信頼度の加味が容易になり、また、知識が整理され見通しのよいモデル化が可能となる。一方、映像など連続画像への応用として、各階層間での情報の共有との観点からの検討も行なった。連続画像においては、同一のシーンの近景の画像や遠景の画像があらわれる。このため、各画像の認識結果から、次の画像を認識するための情報を獲得し、利用することが可能となる。例えば、人間を認識する場合、服装の色などは前もってわからないため、定常的には認識に利用できないが、同一場面では同じ人間は同じ服装をしていると考えてよいので、一度認識できれば次の画像を認識するときから服装の情報を利用することが可能であるというようにである。これは、近景画像と近景画像の間でも可能であるが、近景画像から得た特徴を遠景画像の認識に利用することなどが可能であるということである。本論文では、サッカー映像の選手の認識を例に実験を行ない、その可能性を示した。

今回、マルチメディア環境をにらんだ映像情報処理において重要な課題となる、いくつかの具体的目的や対象を想定し、それらに応じたモデル化手法、画像認識手法について提案、検討を行なった。これらは、それぞれ、シーンの表現、多様な映り方に対する物体の表現、認識知識を含めた対象の表現を可能とし、意味的、物理的、知識的な画像構成の中で階層的に位置付けることができる。このため、組み合わせて利用することで、よりグローバルな認識システムを実現することが可能である。

画像認識では、人間が可能な程度に、得ることのできるすべての情報を獲得できること

が理想であるが、その情報量の多さや、画像認識技術の限界、そして、計算機資源の制限などから、必要な情報のみを効果的に獲得できる技術が重要となっている。これはある意味で、ロボットビジョンなどにおいて、完全な情報を獲得できなくても、必要な情報のみ得られれば良いという目的指向型ビジョンの考え方に基づくものともいえる。今回示した三つのモデル化手法は、このような観点から提案を行なったもので、様々な画像・映像情報を扱うモデル化手法の一例にすぎないが、それぞれ、大きな特徴を有しており、今後のモデル化と画像認識の枠組の一方向性を示せたのではないかと考えている。

今後、画像・映像情報の高度な利用を考えていく上では、本論文で検討したような画像・映像情報のみから必要な情報の獲得を実現しようとするだけでなく、現在の技術水準を考え、文字放送など映像にリンクした形で付加される様々な情報を活用して画像・映像情報のより高度な処理を行なうことも必要となると思われる。今後の画像・映像情報の処理技術の向上により、より快適なマルチメディア環境が実現されることに期待したい。

謝 辞

本研究において、研究題目の選定や研究方針の決定など、様々な面において懇切丁寧なご指導をいただきました東京大学生産技術研究所、坂内正夫教授に心より感謝いたします。坂内正夫先生には、研究の進行具合を暖かく見守っていただいただけでなく、進路の面においても重要なご助言をいただき、感謝の念を忘れえません。

先輩であり、現在、学術情報センターにおいて助手をされている佐藤真一先生には、本研究に対し、多大なるご意見、ご助力をいただき、深く感謝いたします。

現在、東京商船大学において助教授をされている全炳東先生、東京大学生産技術研究所において講師をされている舘村純一先生、東京大学生産技術研究所において助手をされている柳沼良知先生には、研究、および、研究室生活において、多くのご助言をいただき、大いに感謝いたします。

先輩である、魯偉氏、GongYiHong 氏、林英明氏、山根淳氏には、研究のみならず、研究室生活にあたり、いろいろと重要なご助言をいただきました。心より感謝いたします。

また、研究室の生活環境やマシンの管理など、研究環境を整えていただいた、東京大学生産技術研究所の佐藤秀文部技官、東京大学大学院博士課程の相良毅氏、修士課程の谷田部智之氏に深く感謝いたします。

そして、研究室においてお世話になりました、WuWei 氏、佐藤隆氏、西角直樹氏、李春暁氏、汪平涛氏、劉佩林氏、影山誠氏、木村琢也氏、矢野尾一男氏、佐野純平氏、山野繁樹氏、勝股鉄弥氏、和泉直樹氏、河村貴弘氏、田原光穂氏、大場敏文氏、柳田岳洋氏、高崎浩二氏、奥橋俊之氏、樋渡政洋氏、小野敦史氏、高羽洋樹氏、松本文子氏、石山信郎氏、その他東京大学生産技術研究所坂内研究室の皆様に深く感謝いたします。

最後に、今後の人生を方向づける上で重要な指針を示していただきました武蔵工業大学、山田新一教授、および、藤川英司教授に深く感謝の意をあらわします。

1996 年 12 月 20 日

参考文献

発表文献（学会誌論文）

[1] S.Satoh, H.Mo and M.Sakauchi: "Drawing Understanding System Incorporating Rule Generation Support with Man-Machine Interactions", *IEICE Transactions on Information and Systems*, Vol.E77-D, No.7, pp.735-742, 1994.

[2] S.Satoh, H.Mo and M.Sakauchi: "An Efficient Extraction Method for Closed Loops Using a Graph Search Technique", *IEICE Transactions on Fundamentals*, Vol.E78-A, No.5, pp.583-586, 1995.

発表文献（国際会議論文）

[3] S.Satoh, H.Mo and M.Sakauchi: "Drawing Image Understanding System with Capability of Rule Learning", *Proceedings of the Second International Conference of Document Analysis and Recognition*, pp.119-124, 1993.

[4] S.Satoh, H.Mo and M.Sakauchi: "Robust Line Drawing Understanding Incorporating Efficient Closed Symbols Extraction", *Proceedings of IAPR Workshop on Machine Vision Applications '94 (MVA'94)*, pp.107-110, 1994.

[5] H.Mo, S.Satoh and M.Sakauchi: "A Study of Image Recognition Using Similarity Retrieval", *Proceedings of the First International Conference on Visual Information Systems (Visual'96)*, pp.136-141, 1996.

[6] H.Mo, S.Satoh and M.Sakauchi: "A New Type of Video Scene Classification System Based on Typical Model Database", *Proceedings of IAPR Workshop on Machine Vision Applications '96 (MVA'96)*, pp.329-332, 1996.

[7] W.Wu, H.Mo and M.Sakauchi: "An Image Retrieving System Based on User-Specified Recognition Model", *Proceedings of IAPR Workshop on Machine Vision Applications '96 (MVA'96)*, pp.506-509, 1996.

発表文献（大会論文）

[8] 孟, 佐藤, 坂内: "多層のモデルを用いたスポーツシーンからの人間の抽出", 情報処理学会第 47 回全国大会講演論文集, 7L-1, 1993.

[9] 孟, 佐藤, 坂内: "複数モデルの協調による人物の抽出の検討", 情報処理学会第 48 回全国大会講演論文集, 4M-2, 1993.

[10] 孟, 佐藤, 坂内: "シーンの状況モデルを用いた画像認識の検討", 情報処理学会第 49 回全国大会講演論文集, 5F-10, 1994.

[11] 佐藤, 孟, 坂内: "線図形理解のための効率の良い閉ループシンボル抽出手法", 1994 年電子情報通信学会秋期大会講演論文集, D-294, 1994.

[12] 孟, 佐藤, 坂内: "シーンデータベースを用いた画像認識の検討", 情報処理学会第 51 回全国大会講演論文集, 5Q-1, 1995.

発表文献（その他）

[13] 佐藤, 孟, 坂内: "図面理解システムのための人間機械協調を用いたルール作成支援手法", 学術情報センター紀要, 第 7 号, pp.165-178, 1995.

参考文献

[14] 佐藤: "画像データベースにおけるモデル形成に関する研究", 東京大学博士論文（情報工学）, 1992.

[15] S.Satoh, Y.Ohsawa and M.Sakauchi: "Drawing Image Understanding Framework Using State Transition Models", *Proc. of 10th International Conference on Pattern Recognition*, pp.491-495, 1990.

[16] 佐藤, 大沢, 坂内: "対象の多様性に対応しうる図面理解システムの一提案", 情報処理学会論文誌, Vol.33, No.9, pp.1092-1102, 1992.

[17] 坂内, 佐藤: "画像データベースにおけるモデル形成", 電子情報通信学会論文誌, J74-D-I, No.8, pp.455-466, 1991.

[18] M.Sakauchi: "Database Vision and Image Retrieval", *IEEE Multimedia*, Vol.1, No.1, pp.79-81, 1994.

[19] 坂内: "画像検索技術", 電子情報通信学会誌, Vol.71, No.9, pp.911-914, 1988.

[20] J.Y.Aloinomos: "Purposive and Qualitative Active Vision", Proc. of 3rd IEEE International Conference on Computer Vision, pp.346-360, 1990.

[21] 池内: "タスクオリエンティドビジョン", 電子情報通信学会誌, Vol.J74, No.4, pp.360-365, 1991.

[22] K.W.Bowyer and C.R.Dyer: "Aspect Graphs: An Introduction and Survey of Recent Results", International Journal of Imaging Systems and Technology, Vol.2, pp.315-328, 1990.

[23] R.A.Brooks: "Model-Based Three-Dimensional Interpretations of Two-Dimensional Images", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.PAMI-5, No.2, pp.140-150, 1983.

[24] 山本, 加藤, 佐藤, 井口, "機能モデルを用いた3次元物体認識", 電子情報通信学会論文誌, Vol.J74-D-II, No.5, pp.601-609, 1991.

[25] L.Stark and K.Bowyer: "Achieving Generalized Object Recognition through Reasoning about Association of Function to Structure", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.PAMI-13, No.10, pp.1097-1104, 1991.

[26] L.Stark and K.Bowyer: "Function-Based Generic Recognition for Multiple Object Categories", *CVGIP: Image Understanding*, Vol.59, No.1, pp.1-21, 1994.

[27] T.M.Strat and M.A.Fischler: "Context-Based Vision: Recognizing Objects Using Information from Both 2-D and 3-D Imagery", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.PAMI-13, No.10, pp.1050-1065, 1991.

[28] D.M.Mckeown: "Rule-based interpretation of aerial imagery", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.PAMI-7, No.5, pp.570-585, 1985.

[29] S.H.Joseph and T.P.Pridmore: "Knowledge-Directed Interpretation of Mechanical Engineering Drawings", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.PAMI-14, No.9, pp.928-940, 1992.

[30] R.M.Bolle: "A Complete and Extendable Approach to Visual Recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.PAMI-14, No.5, pp.534-548, 1992.

[31] R.Bergevin and M.D.Levine: "Generic Object Recognition: Bulding and Matching Coarse Description from Line Drawings", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.PAMI-15, No.1, pp.19-36, 1993.

[32] 松原, 坂上, 山本, 山岸: "画像学習システム MIRACLE-IV における機能的特徴と視覚的特徴の対応付け", 情報処理学会論文誌, Vol.31, No.9, pp.1302-1311, 1990.

[33] 松山, 尾崎: "LLVE:トップダウン・セグメンテーションのための画像処理エキスパートシステム", 情報処理学会論文誌, Vol.27, No.2, pp.191-204, 1986.

[34] T.Matsuyama and V.S.Hwang: *SIGMA - A Knowledge-Based Aerial Image Understanding System*, Plenum Publishing, 1990.

[35] 松山, 村山, 伊藤: "画像解析における戦略の表現", 情報処理学会論文誌, Vol.29, No.2, pp.169-177, 1988.

[36] B.K.Natarajan: *Machine Learning - A Theoretical Approach*, Morgan Kaufmann, 1991.

[37] E.Y.Shapiro: "Inductive Inference of Theories from Facts", *Tech. Rep. TR192, Yale University, Dept. of Computer Science*, 1981.

[38] W.Lu, Y.Ohsawa and M.Sakauchi: "A Database Capture System for Mechanical Drawings Using an Efficient Multi-dimensional Graphical Data Structure", *Proc. of 9th International Conference on Pattern Recognition*, pp.266-269, 1988.

[39] 高木, 下田: 画像解析ハンドブック, 東京大学出版会, 1991.

[40] D.Marr: *Vision*, W.H.Freeman, 1982.

[41] 長坂, 宮武, 上田: "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", 電子情報通信学会論文誌, Vol.J79-D-II, No.4, pp.531-537, 1996.

[42] J.Yamane, M.Sakauchi: "A Construction of A New Image Database System which Realizes Fully Automated Image Keyword Extraction", *IEICE Transactions on Information and Systems*, Vol.E76-D, No.10, pp.1211-1233, 1993.

[43] 小野, 天野, 斗谷, 佐藤, 坂内: "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", 電子情報通信学会論文誌, Vol.J79-D-II, No.4, pp.476-483, 1996.

[44] Y.Yaginuma and M.Sakauchi: "Multi-Purpose Interface for Still/Moving Image Retrieval", *Proc. of SPIE Applications of Digital Image Processing XVII*, Vol.2298, pp.260-267, 1994.

[45] 長坂, 田中: "カラービデオ映像における自動索引付け法と物体探索法", 情報処理学会論文誌, Vol.33, No.4, pp.543-550, 1992.

[46] 美濃, 岡崎, 坂井: " 対象物の属性特徴による画像検索法－風景画像中の山を例として－", 情報処理学会論文誌, Vol.32, No.4, pp.513-522, 1991.

[47] 栗田, 下垣, 加藤: "主観的類似度に適応した画像検索", 情報処理学会論文誌, Vol.31, No.2, pp.227-237, 1990.

[48] 栗田, 加藤, 福田, 坂倉: "印象語による絵画データベースの検索", 情報処理学会論文誌, Vol.33, No.11, pp.1373-1383,1992.

[49] 加藤, 下垣, 藤村: "画像対話型商標・意匠データベース TRADEMARK", 電子情報通信学会論文誌, Vol.J72-D-II, No.4, pp.535-544, 1989.

[50] T.Kato, T.Kurita, N.Otsu and K.Hirata: "A Sketch Retrieval Method for Full Color Image Database", *Proc. of 11th International Conference on Pattern Recognition*, pp.530-533, 1992.

[51] 椋木, 美濃, 池田: " 対象物スケッチによる風景画像検索とインデックスの自動生成", 電子情報通信学会論文誌, Vol.J79-D-II, No.6, pp.1025-1033, 1996.

[52] 黒川, 洪: "形状情報を用いた画像の類似検索システム", 情報処理学会論文誌, Vol.32, No.6, 1991.

[53] 高橋, 島, 岸野: "位置情報を手がかりとする画像検索法", 情報処理学会論文誌, Vol.31,No.11 1990

[54] 柴田, 井上: "画像データベースの連想検索方式", 電子情報通信学会論文誌, Vol.J73-D-II, No.4, 1990.

[55] 有村, 萩田: "統計的画像認識における射影追跡に基づく画像選別", 電子情報通信学会論文誌, Vol.J78-D-II, No.6, pp.913-921, 1995.

[56] J.N.David Hibler et al.: "A System for Content-based Storage and Retrieval in An Image Database", *Proc. of SPIE Int. Soc. Opt. Eng.*, Vol.1662, 1992.

[57] S.K.Chang, C.W.Yan: "An Intelligent Image Database System", *IEEE Trans. on Software Engineering*, Vol.14, No.5, pp.681-688, 1988.

[58] M.Flickner, H.Sawhney, W.Niblack, J.Ashley, Q.Huang, B.Dom, M.Gorkani, J.Hafner, D.Lee, D.Petkovic, D.Steele and P.Yanker: "Query by Image and Video Content: The QBIC System", IEEE Computer, Vol.28, No.9, pp.23-32, 1995.

[59] Y.Gong and M.Sakauchi: "A Method for Color Moving Image Classification using the Color and Motion Features of Moving Images", *Proc. of ICARCV'92*, 1992.

[60] 君山, 清末, 大庭: "動画像の自動記述方法の検討", 電子情報通信学会技術研究報告, IE91-110, 1991.

[61] 佐藤, 坂内: "ライブハイパメディアにおける映像情報の獲得", 電子情報通信学会論文誌, Vol.J79-D-II, No.4, pp.559-567, 1996.

[62] 小林重信: "事例ベース推論の現状と展望", 人工知能学会誌, Vol.7, No.4, pp.559-566, 1992.

[63] 鎧沢, 宮田: "人物像の画素数と識別可能情報の関係", 電子情報通信学会技術研究報告, IE86-95, pp.31-37, 1986.

[64] 田辺: "カメラから人物までの撮像距離を考慮した顔画像認識について", 第20回画像工学コンファレンス, pp.193-196, 1989.

[65] 皆川, 川嶋, 青木: "スポーツ中継における動画像解析～サッカー中継のシーン理解～", 画像の認識・理解シンポジウム (MIRU'92), Vol.1, pp.161-167, 1992.

[66] A.Tsukamoto, C.Lee and S.Tsuji: "Detection and Tracking of Human Face with Synthesized Templates", *Proc. of Asian Conference on Computer Vision '93 (ACCV'93)*, pp.183-186, 1993.

[67] S.Akamatsu, T.Sasaki, H.Fukamachi and Y.Suenaga: "Target Image Extraction for Face Recognition Using The Sub-Space Classification Method", *Proc. of IAPR Workshop on Machine Vision Applications '92 (MVA'92)*, pp.465-468, 1992.

[68] Y.Sumi and Y.OHTA: "Human Face Analysis Based on Distributed 2D Appearance Models", *Proc. of IAPR Workshop on Machine Vision Applications '92 (MVA'92)*, pp.469-472, 1992.

[69] 安居院, 長尾, 中嶋: "静止濃淡情景画像からの顔領域の抽出", 電子情報通信学会論文誌, Vol.J74-D-II, No.11, pp.1625-1627, 1991.

[70] T.Watanabe S.nakagawa and Y.Kuno: "A Method of Human Recognition Using Silhouette Images", *Proc. of Asian Conference on Computer Vision '93 (ACCV'93)*, pp.71-74, 1993.

[71] R.Brunelli and T.Poggio: "Face Recognition: Features versus Templates", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.15, No.10, pp.1042-1052, 1993.

[72] 嶌田, 小池, 伴野, 石井: "顔の向きによらない人物識別のための辞書構成法", 電子情報通信学会論文誌, Vol.J78-D-II, No.11, pp.1639-1649, 1995.

[73] M.Herman and T.Kanade: "Incremental Reconstruction 3D Scenes from Multiple, Complex Images", *Artificial Intelligence*, Vol.30, pp.289-341, 1986.

[74] 大田: "トップダウン的画像理解におけるあいまい推論の枠組みについて", 画像の認識・理解シンポジウム (MIRU'92), Vol.1, pp.1-8, 1992.

[75] 松山, 栗田: "Dempster-Shafer の確率モデルに基づくパターン分類", 電子情報通信学会論文誌, Vol.J76-D-II, No.4, pp.843-853, 1993.

[76] G.Shafer: "A Mathematical Theory of Evidence", Princeton Univ. Press, 1976.