〔非公開〕

ＴＲ－Ｃ－０００７

言語・画像情報統合理解の研究

伯田　晃　　高橋　友一　　小林　幸雄

AKIRA HAKATA　TOMOITI TAKAHASHI　YUKIO KOBAYASHI

１９８７．１２．２５

ＡＴＲ通信システム研究所

言語・画像情報統合理解の研究

伯田　晃
高橋友一
小林幸雄

1. まえがき

近年、誰にでも使い易く、人間本来の知的活動を支援してくれるような情報システムが切望されている。それに対して、インタフェース高度化の努力が様々なアプローチで行われており、その中には、我々人間が日常使い慣れている自然言語を使えるように、機械の知的レベルの向上を図ろうとする先端的な研究もある。そして、現在のテキスト情報を中心に扱うコンピュータシステムに於いては、この自然言語を使えるインタフェースは、利用者の持っている意図を自然に表現できる点で一つの理想形態であると考えられる。

一方、ＬＩＳ技術の急激な進歩にともなって、コンピュータの処理能力は日々格段の進歩を遂げている。そして近い将来、テキスト情報は当然のこと、イメージ情報もより効率的に取り扱えるようになることは確実であると言われている。また、その技術を適用したマルチメディアデータベースを代表とする、より高度な情報処理システムの実現も強く期待されている。その様なシステムが現実のものとなった時には、いかに自然言語によるインタフェースが優れているといえども、必ずしもそれだけでは使い易いとは考え難い。我々は、そのようなイメージ情報なども取り扱える状況に於いては、人間同士が日常よく行っているように、言葉だけではなく、そのやりとりの際に必要な画像情報をうまく用いることのできるインタフェースが、解決の道を与える一つのアプローチになると考えている。

このような複数メディアを併用するインタフェースの先駆的な研究としては、1980年に MITのRichard A. Bolt が、実際の物を指で指し示す簡単な動作と指示語を組み合わせて入力された内容を理解するインタフェースを提唱している [(1)]。我々はこの Bolt の研究成果を踏まえながら、複合電話機の操作ガイダンスシステムを実験システムに採り上げて、利用者の入力する言語・画像情報を統合して理解結果を得る為の基本技術の検討を行っている。

本稿では、基本技術の一つである、言語情報と画像情報の対応関係を示す『言語・画像情報間のリンク』に関して、実験システムでの検討を通して得られた、我々人間が特徴的なパターンを見た際に認識されるグループ情報を含むリンクを中心に報告する。

2. 言語・画像を併用するコミュニケーション

図１は、言葉と画像を併用してやりとりを行う、我々人間同士のコミュニケーションの一例である。この例は、Ａが実際に花を提示して、その花の美しさをＢに伝えている場合であり、実際の花を指差しながら「これは美しい！」などと言葉で言って会話を進めることになる。もしＡが実際の花を提示しないで、言葉だけで同じ内容を伝えようとしたら「花の種類は○○で、大きさは△△で、花びらの形は□□で、色は◇◇で、 ……」など、その花に関しての一部始終を伝えるか、または「花瓶にさしてあるバラの花は美しい」などと言って、相手が持っているであろう同じイメージを刺激することで、同じ花を想起してくれることに頼る以外に方法はない。

しかし、いくら言葉で詳しく述べることができても、または共通のイメージを想起してくれることを期待してその花について述べたとしても、ＢがＡの言っている花そのものを想起できるとは考え難い。この様なやりとりの際には、画像情報を仲介としてイメージを共通化することで、話し手にとっては伝えたい意図を非常に簡潔に表現することが可能となる。また聞き手にとっても、話し手の指差している先にある実物を実際に眼で見て、それがバラの花であることを知ることで、「目の前にあるバラの花が美しい」と話し手が主張していると理解できる。
我々は、『人間』と『機械』とのインタフェースに於いても、上で述べた様なやりとりを実現することを目指している。

3. 言語・画像情報統合理解実験システム

本章では、言語・画像情報統合理解の実験システムとして具体的に検討をすすめている、複合電話機の操作ガイダンスシステムについて述べる。

3.1 実験システムへの入力

実験システムは、(1)各種サービスを享受する為に必要な電話機の操作手順　(2)電話機パネルの表示（ランプの点灯　など）の意味　について質問する際に、実際の電話機の画像を自由に使いながら自然言語で質問が行えるシステムである。
我々は検討に先立ち、上記(1)(2)を尋ねる際の実際の質問の仕方を収集した。以下にその一部抜粋を示す。

(1)電話機の操作手順についての質問

① 短縮ダイヤルのセットの仕方は？
② 外線にかけるにはどうすればよいのですか？
③ （会議ボタンを指差しながら）このボタンを押すと会議ができるのですか？
④ 右上のボタンを押すと外線にかけられるのですか？
⑤ 一番左側に縦に並んでいる上から２つ目のボタンを押すだけで相手につながるのですか？

(2)電話機パネルの表示の意味

① （伝言ランプを指差しながら）何故点滅しているのですか？
② （不在転送ボタンを指差しながら）これが点滅し始めました。
③ スピーカボタンを押すと右上のランプが点きましたが？
④ 保留ボタンが点灯していますけど何を意味しているのですか？
⑤ 不在転送と書いたボタンがチカチカしてますけど？

上の例文中の、網掛けを施した部分が実際の電話機画像が用いられて入力されている部分である。

3.2 画像情報の入力方法とその読み取り

入力に画像情報が用いられている場合には、実際の画像を参照しなければその具体的内容はわからない。その為には、画像情報の入力のされ方を知る必要がある。表1は、収集した質問を画像情報の入力のされ方の観点で分析した結果である。大きく分けて、指などを使って特定の画像を指し示して入力される場合(1)と、「右上のボタン」・「保留ボタン」などのように言葉で述べられて入力される場合(2)に分類できる。また、これらの組合せとして、指し示す動作と「これ」・「この」などの指示語が合わせて用いられて入力されることがあり、この(3)の場合が Bolt の提唱しているインタフェースに該当する。

(1)の方法で画像情報が入力された場合には、指先にある画像が入力されている画像であり、それを読み取る処理を行う。また、(2)の入力から画像情報を読み取る為には、言語情報と画像情報との対応関係を示す情報が必要で、その情報と入力された言語情報とを合わせて用いることにより画像を発見する処理を行う。この言語情報と画像情報の対応関係を示す情報を我々は『言語・画像情報間のリンク』と呼んでいる。そして、(3)の入力に対しては、上記(1)(2)で得られる各々の結果を併用して画像を読み取る処理が必要である。

本稿では、指などで指し示す方法で入力される情報の読み取りには、当研究室で現在進めている画像処理技術の研究成果を流用することを考えており、次章では(1)の入力を理解する際に用いる『言語・画像情報間のリンク』について議論を行う。

4. 言語・画像情報間のリンク

『言語・画像情報間のリンク』とは、前章でも述べたように言語情報と画像情報の対応関係を示す情報であり、具体的には電話機各部分の画像を言葉で呼ぶ時の言い方に相当する。

4.1 言語・画像情報間のリンクの二つの要素

収集した質問で得られた、電話機各部分の画像を言葉で呼ぶ時の言い方の中から数例を抜粋して以下に示す。

①「右上のボタン」　②「右上の赤いランプ」
③「保留ボタン」　④「左上のもの」

ここにあげた例文の中には、

(1)画像が何であるかを認識しているレベルの情報〔画像の呼び名〕
(2)画像が、当人の視界の中でどの様に見えているかを述べるレベルの情報

が含まれていると考えている。

我々は、(1)の情報を『概念レベルの情報』、(2)の情報を『視覚レベルの情報』と呼んで区別して扱う。上の例文では、下線を引いた部分が『概念レベルの情報』に、網かけを施した部分が『視覚レベルの情報』に該当している。

『概念レベルの情報』では、電話機を話題にしている時には、電話機表面の「四角い形状をしたもの」が『短縮ボタン』・『ボタン』などと呼ばれる。しかし一方、地図について話している時には、地図上の同じような「四角い形状をしたもの」を、『ツイン21』・『ビル』などと呼ぶことになる。このように『概念レベルの情報』は、話題にしている事柄が何なのか、その話題に対

して発話者がどの程度の造形を持っているのか、また発話者の知識がどのように構造化されているかなどの要因によって、その表現は当然異なることになる。従って、話題の分野に造形の無い人は、単に抽象的な『もの』などの言葉を使って言及するような場合も考えられる。
一方、『視覚レベルの情報』に該当する代表的なものとして、画像の位置、形、大きさ、色などが述べられる。この『視覚レベルの情報』は、『概念レベルの情報』が、取り扱っている話題に依存しているのに対して、基本的には我々人間が共通して持つ認識機能に基づくものであり、対象とする分野（話題）には依存しないと考えられる。本稿では特に、この『視覚レベルの情報』について検討を加える。

4.2 実験システムに於ける特徴的なリンク

実際に収集した質問の中には、図２の複合電話機のボタン(a)を、「一番左側に縦に並んでいる上から２つ目のボタン」と言う様なものが多くあった。複合電話機のパネルには、短縮ボタン・ワンタッチダイヤルなどの各種ボタンや、電話機の状態を知らせる表示ランプが整然と配置されており、我々人間はこれら複数の画像（ボタンやランプなど）をグループ化して認識すると言われている。多くの人々が、(a)のボタンについて上のような言い方をした背景には、認識されたグループの情報を、(a)の言及の際にも使っている為だと推測される。
グループの認識については、認知心理学の分野で1920年頃からゲシュタルト心理学派を中心として研究がすすめられており、 Wertheimer が彼の研究成果を『グループ化のゲシュタルト法則』[2] としてまとめていることは有名である。
その法則の代表的なものは以下の通りである。

①近接の法則： 距離の近いもの同士はグループ化して認識される。
②類同の法則： 同じ属性（色・大きさ・形など）を持ったもの同士はグループ化して認識される。
③よい連続の法則： 滑らかに連なるもの同士はグループ化して認識される。
④閉合の法則： 閉じた図形を構成するもの同士はグループ化して認識される。

我々は、グループ情報を含む『言語・画像情報間のリンク』を検討する為に、これらグループ化の法則によってグループが認識されると言われている特徴的なパターンを用いた実験を行った。

4.3 グループ情報を含む『言語・画像情報間のリンク』に関する実験

複合電話機のパネルのパターンでは、近接の法則・類同の法則によるグループ化が特に関連していると思われる。そこで我々は、近接の法則・類同の法則を中心に、よい連続の法則・閉合の法則についても合わせて以下の実験を行った。

4.3.1 実験の方法

グループ化が生じると言われているパターン（近接：２種類、類同：２種類、よい連続：２種類、閉合：２種類、近接＋類同：６種類、類同＋よい連続：３種類、よい連続＋閉合：１種類 合計１８種類）と、グループ化が生じないと言われているパターン（２種類）を予めパーソナルコンピュータに準備した。そして６人の男性被験者に、準備したパターンをディスプレイ上に提示し、その中のやはり予め決めておいた特定の対象（合計約 100対象／人： パターンを構成す

る構成要素のことを対象と呼ぶ）を、その対象の位置関係や外観などを代表とする視覚情報に基づいて他と区別する様に言葉で言わせた。そして実験の後、各人の言及した内容を分析し、認識されたグループに関する言及の部分、対象に関する言及の部分を抽出した。
図3に、実験に用いたパターンの一部を示す。(1)はグループ化が生じない、(2)は近接の法則によりグループ化が生じる、(3)は類同の法則（色属性の違い）によりグループ化が生じる、(4)は良い連続の法則によりグループ化が生じる、(5)は閉合の法則によりグループ化が生じる、(6)ではグループ（近接）の中で再びグループ化（類同：色属性の違い）が生じるパターンである。パターン中の点線は認識されるグループを表しており、また矢印のついている対象は実際に言葉で言わせた対象を示している。

4.3.2 実験結果、および分析

(1)グループについての言及の割合

表2に、各パターンでの対象を言及する際のグループ情報言及者の人数の比率を示す。本実験を通じて、(3)を除くグループ化の生じるパターンでは、多くの被験者が対象を言及する際に、グループに関しての情報を用いることの確認が得られた。
被験者がグループ情報を用いて対象を言及するのは、グループ情報を用いることで画面の中での範囲を絞り込むことが可能となり、そのことがより表現を容易にすることにつながる為であると推測される。(3)のパターンの場合にグループ情報の言及率が低かったのは、グループ情報を用いなくても比較的容易に対象を言及できた為であろう。今後、パターンとそのパターンを言及する際のし易さなどについても検討を行っていく必要がある。
尚、表3には参考の為に、図3のパターン(2)(3)(6)の(a)～(c)の対象についての実験結果を記載しておく。

(2)対象特定の為の情報

本実験では、全ての対象は異なった位置に配置されている為、基本的には対象の存在する位置情報を正しく明示さえすればよいはずである。しかし実際には、位置情報だけで言及してくるとは必ずしも限らず、対象の持つ属性（色、形、大きさ）についての情報も合わせて用いられて述べられる。例えば、図3(3)の(a)は「上から3番目で一番右の白い四角」と呼ばれることがあり、その場合には、位置情報の他に色情報（白い）・形情報（四角）が述べられている。

(3)グループ特定の為の情報

認識されるグループもその位置が全て異なっている為、基本的にはグループの存在する位置情報を明示さえすればよいはずである。しかしこの場合にも、必ずしも位置情報だけでグループを言及してくるわけではなく、グループ化が生じた原因（近接、類同、連続、閉合）によって特徴的な情報を用いて言及してくる。例えば、近接によるグループではグループの中に存在する対象の内訳（数、属性）・接近してできあがったパターンの様子などが、類同によるグループではグループ化の原因となった対象の属性（本実験では色情報）が、連続によるグループではグループの中での対象の並び方が、閉合によるグループでは閉じた図形の呼び方（名前）がグループの位置情報に合わせて言及される傾向にある。例えば、図3(2)の（A)のグループが「左上の二つの四角からなるグループ」と言われた場合は、位置情報（左上の）の他に対象の内訳情報（二つの四角からなる）が述べられていることになる。

(4)入れ子のグループについての言及

図3(6)のパターンでは、一つのグループの中で再びグループ化（いわゆる「入れ子構造」）が生じる。この場合には、対象についての言及は、外側のグループについての言及から内側のグループについての言及へと移り、最終的に対象を絞り込むような方法で行われる。例えば、(6)の(a)が「左下のグループ「1」の赤のグループ「2」の下の右から2番目の対象「3」」（原稿では白黒のコピーになっているが、実験で用いたのは白と赤である）と言われる場合には、先ず始めに左下にあるグループが述べられ（「1」）、引き続いてその中の赤い対象ばかりのグループが述べられ（「2」）、最後に対象の言及（「3」）が行われている。

(5)グループ・対象言及の実際の表現方法

グループ・対象を言及する実際の言い方は、「右上にある」、「二つの四角からなる」、「赤い」などの連体修飾語を用いて、「グループ」・「対象」を代表とする単語を限定する言い方が基本である。そしてその表現は、アンダーラインの部分が助詞『の』に置き換わり、『右上の対象、右上のグループ、二つの四角のグループ、赤のグループ』などの簡略化された表現になる事がよくある。また、グループ・対象言及ともに、修飾されるべき「グループ」・「対象」などの単語が省略されることが多く、従って、複数の入れ子構造を持つグループを述べる際に、簡略化された表現が用いられ、かつ単語「グループ」が省略されると、『～の～の～の』と修飾語が連なった特徴的な形の表現になってしまう。例えば、図3(6)の(a)は『左下の赤の下の右から2番目』などの表現となる。

5. グループ情報を含む『言語・画像情報間のリンク』実現の課題

グループの言及を含めて入力された言語情報から、それに該当する対象を見つけ出すには、先ずグループに関する言及を使ってグループを同定し、次にそのグループの中で、対象に関する言及を使って対象の同定を行う方法が自然である。このような方法を採ろうとした場合、次の二つの課題が存在する。

(1)グループ・対象を言及する情報の発見

入力に、「グループ」・「対象」を意味する直接的な単語が出てくる場合（図3(6)の(a)を「左下にあるグループの赤い対象からなるグループの下の右から2番目の対象」と言われる）には、当然それぞれを限定する情報も明確である。この場合には、先ず第一番目のグループの情報が「グループの位置＝左下にある」であるので、それを使ってグループを見つけ出す。そして更に、そのグループの中から第二番目のグループの情報として「グループの構成要素＝赤い対象からなる」を持つグループを見つけ出す。そして、見つかったそのグループの中から対象の情報として「対象の位置＝下の右から2番目」を持つ対象を捜し出すことになる。

しかし、直接的な単語が出てこない場合（例えば、「左下の赤の下の右から2番目」と言われる）には、言語解析の段階ではグループ・対象についての情報を明確に特定することができない。つまり、 ①グループについての情報：「グループの位置＝左下にある」 かつ 「グループの構成要素＝赤い対象からなる」、対象についての情報：「対象の位置＝下の右から2番目」 と解釈すべきなのか、あるいはグループが入れ子構造となっており、 ②外側のグループについての情報：「グループの位置＝左下にある」、内側のグループについての情報：「グループの構成

要素＝赤い対象からなる」、対象についての情報：「対象の位置＝下の右から２番目」と解釈すべきなのか、この段階では判断がつかない。
このようなシステムでは、言語解析処理と、実際にグループ・対象を見つけ出す同定処理とを各々単独で考えることは不適当であり、各々の処理結果の良し悪しをフィードバックし、互いに協力し合って処理を進める必要がある。

(2)グループ・対象の同定処理

前節でも述べた様に、原理的には位置情報だけでグループ・対象を特定することは可能であるにもかかわらず、必ずしも位置情報だけで特定するとは限らない。このことは言い換えれば、入力される位置情報は、必ずしも単一の対象に絞り込めるほど厳密なものでないということである。位置情報の他に与えられる、色・形・大きさなどの情報をも含めてこそ同定が可能である場合も存在する。どの様なパターンでは位置情報だけが述べられるのか、どの様なパターンではその他の情報もあわせて述べるのか　などの基礎実験のデータを基に、色・形・大きさ情報などを総合的に評価する同定処理を考える必要がある。

6. まとめ

将来のマルチメディア環境において、利用者が自分の持っている意図を容易に表現可能とすることを目指したインタフェースの一つとして、言語情報に加えて画像情報を併用できる複合メディア指向のインタフェースの検討を行った。
そのようなインターフェースを実現するためには、言葉で述べられて入力される画像を発見するための『言語・画像情報間のリンク』が必要であり、具体的に検討を進めている実験システムに於いては、このリンク情報に人間がパターンを見た時に認識するグループの情報を含めて考えることの必要性が、基本的な実験からも明らかになった。このグループ情報を含んだ『言語・画像情報間のリンク』は、人間本来の持つ画像認識機能を基本とするもので、本実験システムの中に限定されるものではないと考えている。
今後、視覚情報に該当する言葉と画像情報の対応関係をうまく取るためにの画像記述方法、およびそれら情報を取り扱う同定処理を実現するための演算・推論方式などについて、具体的な実験プロトタイプシステムを構築しつつ検討を進める予定である。

【参考文献】

(1) Bolt, Richard A, "'Put-That-There': Voice and Gesture at the Graphic Interface." Computer Graphics 14(3) 262-270. (1980).
(2) Beardslee, D.C., and M. Wertheimer(Eds.), Readings in perception. Princeton, N.J.: Van Nostrand, (1958).
(3) 伯田、小林、山下　「マンマシン・インタフェースに於ける言語・視覚情報の統合化に関する一考察」、昭和62年信学総合全大、1445
(4) 伯田、高橋、小林　「言語・画像情報を統合化するユーザインタフェースの一考察〔グループ化の考えを取り入れた言語・画像情報間のリンク〕、通信学会AI研究会、AI87－28

表1. 画像情報の入力方法

入力方法	例　　文
(1)指などで指し示す （画像情報単体）	①（点灯している伝言ランプを指で指し示しながら）何を意味しているのですか？ ②（外線ボタンを指で指し示しながら）押すとどうなりますか？
(2)言葉で述べる （言語情報単体）	①外線発信は、右上のボタンを押せばよいのですか？ ②点灯しているランプは、何を意味しているのですか？ ③保留サービスは左側にかたまっているボタンの右から2番目を押すのですか？ ④左側の上から3番目のランプはなぜ点滅しているのですか？
(3)言葉で述べると同時に、指などで指し示す （言語情報＋画像情報）	①（”＃”ボタンを指で指し示しながら）このボタンを押すと外線発信できるのですか？ ②（点灯している伝言ランプを指示しながら）これは何を意味しているのですか？

（注）　アンダーラインを引いた部分が、画像情報として入力されている部分

表2. グループ情報言及者の比率

パターン番号	対象番号	グループ情報言及者数	パターン番号	対象番号	グループ情報言及者数
(1)	(a)	0人	(4)	(a)	2人
	(b)	0		(b)	3
	(c)	0		(c)	5
(2)	(a)	6	(5)	(a)	2
	(b)	5		(b)	4
	(c)	6		(c)	3
(3)	(a)	1	(6)	(a)	6
	(b)	1		(b)	6
	(c)	0		(c)	6

（注）被験者は総数6名

表3. 実験で得られた実際の言い方（抜粋）

パターン番号	対象番号	実験で得られた実際の言い方
(2)	(a)	①左上の二つの四角からなるグループの右側のもの　②左上の右側
	(b)	①右上の二つ並んだ四角形の右側　②一番右端の上
	(c)	①下の二つの四角の右側　②下の方の右側
(3)	(a)	①上から3番目で一番右の白い四角　②白の一番右上
	(b)	①上から2列目の左から2番目　②白のうちの下の左から2番目
	(c)	①一番右上　②1列目の右端の四角
(6)	(a)	①左下のグループの赤のグループの下の右から2番目の四角 ②左下の16個のかたまりの一番下の列の右から2番目
	(b)	①左側の4つのかたまりの内の白の右側 ②左上にある4つの四角形の右上にある白い四角形
	(c)	①右側の16個のかたまりの上から3列目の左から2番目 ②右上の四角群の3列目の左から2番目の四角

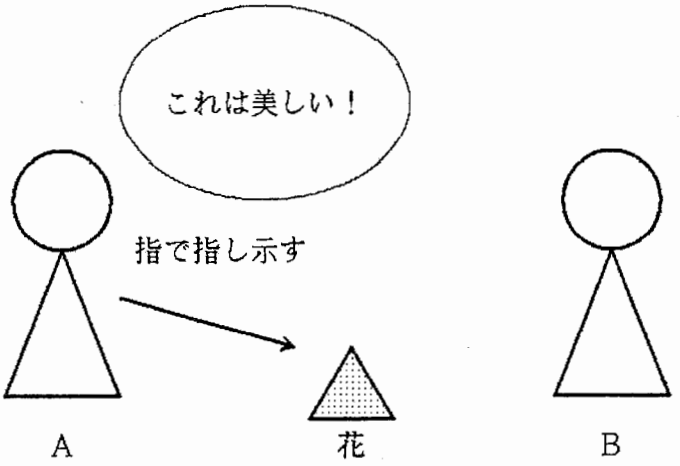


図1.　言語・画像情報を用いたコミュニケーション

(a)　「一番左側に縦に並んでいる上から2つ目のボタン」

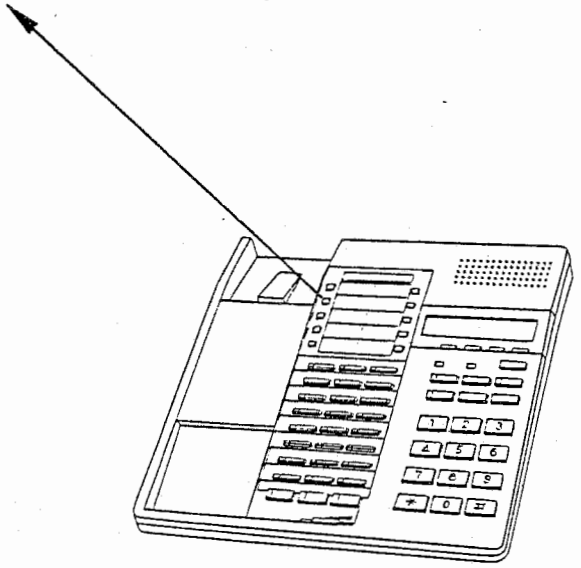
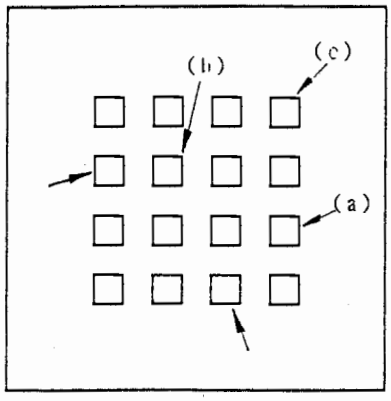
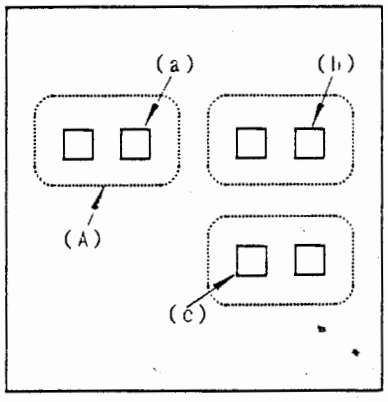


図2.　複合電話機の画像

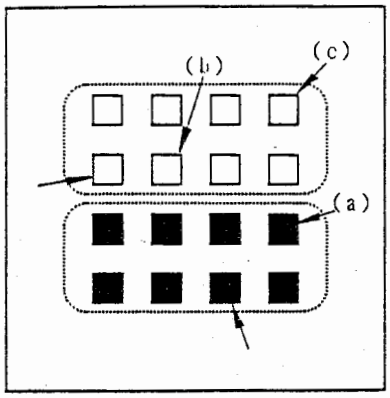
(1)



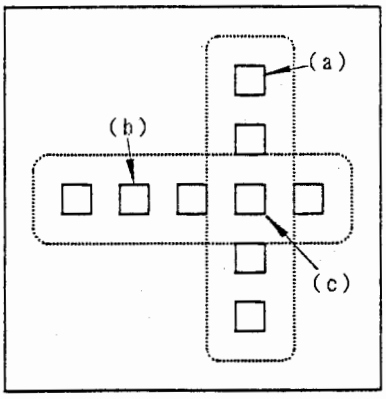
(2)



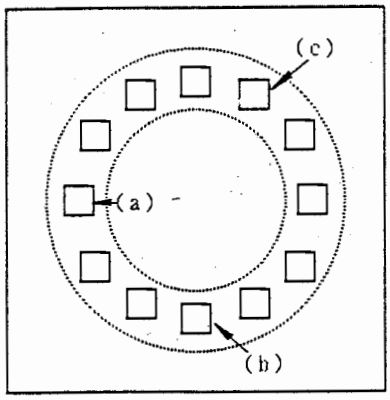
(3)



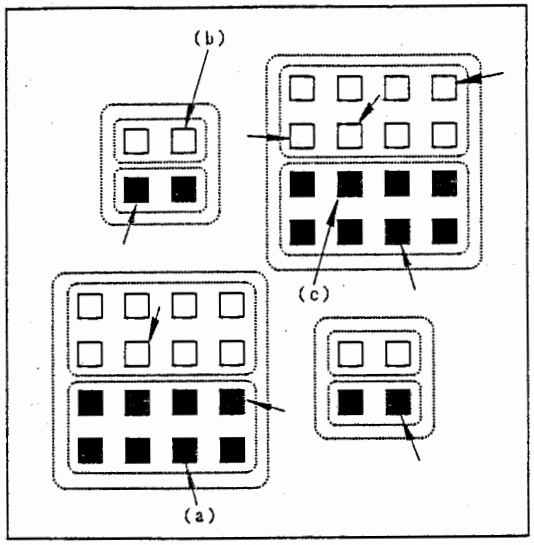
(4)



(5)



(6)



(注) 点線： 認識されるグループ
矢印： 言葉で言わせた対象

図3. 実験に用いたパターン