Kobe University Repository : Kernel

PDF issue: 2026-11-22

マルチモーダル情報を用いた未知物体の認識とグラウンディングに関する研究

小篠，裕子

(Degree)
博士（工学）
(Date of Degree)
2015-03-25
(Date of Publication)
2017-03-25
(Resource Type)
doctoral thesis
(Report Number)
甲第6451号
(URL)
https://hdl.handle.net/20.500.14094/D1006451

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。

博 士 論 文

マルチモーダル情報を用いた未知物体の認識とグラウンディングに関する研究

平成 27 年 1 月

神戸大学大学院 工学研究科

小篠　裕子

マルチモーダル情報を用いた未知物体の認識とグラウンディングに関する研究

要 旨

「マルチモーダル」,「認識」,「グラウンディング」という言葉に馴染みを持つ人は，まだ少ないだろう．しかし，これらの概念は，私たちの生活の中に潜んでいる．私たちは，自覚することなく，これらの概念を持って生活しているのである．第 2 章では，関連研究としてこれらの概念の紹介とその関係性について述べる．

人は，情報をより正確に知覚するため，五感などの感覚から得た情報を複数組み合わせて，処理している．このような処理をマルチモーダル処理という．人は日常生活において，マルチモーダル処理を行っていない時間はほとんどない．例えば，今この論文を読んでくださっている貴方も，ただ視覚のみから論文情報を得ているのではなく，手で触ったこの論文の質感を触覚で感じたり，ページをめくる音を聴覚で感じたりしながら，この論文に関する情報を得ているのである．近年，このマルチモーダル処理を工学技術に入れ込む研究がさかんになされている．実はスマートフォンもマルチモーダル処理を取り入れたマルチモーダル機器のひとつである．着信音や，通話などによって人の聴覚に作用し，バイブレーション機能によって人の触覚に作用し，画面に様々な情報を表示することによって人の視覚に作用したりする．このようなマルチモーダル技術の集大成ともなりうる存在こそが，サービスロボットである．特にコミュニケーション機能を有したサービスロボットは，少子高齢化が進み行く現代において，開発が急がれている．スマートフォンが私たちの生活になくてはならないものになったのと同じように，マルチモーダル処理が行えるサービスロボットは，きっと私たちの生活になくてはならないものになるだろう．

人が，あるものを見たときに，それを何とみなすのか．これこそが「認識」の基本命題である．人は日常において様々なものを認識している．例えば，今机の上にあるものが何なのか，隣に座っている人が誰なのか，今聞こえる音は何なのかなど，常に認識を行いながら生活している．意識的に認識することは勿論であるが，無意識に認識を行っていることもあるだろう．これらの認識を工学的かつ統計的アプローチで行おうとした分野こそが，パターン認識である．音声を認識する音声認識技術，画像を認識する画像認識技術などが代表格である．音声認識も，画像認識も独立して研究がなされてきた．しかし，実際の環境でこれらの技術が使われるときを想像すると，これらのうちのいずれかのみが使われるシーンは少ないだろう．例えば，ロボットが，人に音声で指示された物体を，複数ある物体の中から特定し，人に運ぶというタスクを考える．このタスクをロボットが達成するためには，指示音声を認識すること，物体を認識することが必要となる．どちらか一方

だけの技術ではこのタスクを達成することはできない．音声と画像のマルチモーダル情報を用いることによって，初めて，ロボットは，人に指示された物体を認識し，特定することができるのである．本研究では，このような物体特定タスクを扱う．

人に指示された物体を特定するというタスクを達成するには，人の指示と，物体の組み合わせを予め学習しておく必要がある．言い換えると，人の指示を，物体にひもづけておく必要がある．このひもづけのことをグラウンディングと言う．特に，物体名などの記号（シンボル）を，実世界における物体にグラウンディングすることをシンボルグラウンディングと呼ぶ．物体特定タスクは，人の発話音声と物体画像を認識して，音声がグラウンディングされている物体を複数の物体の中から選択する問題として言い換えることができる．音声と画像のマルチモーダル情報を処理することで，物体特定が可能となる．

物体特定タスクは，ロボットにとって重要なタスクである．作業としての重要度は勿論，ロボットの自律学習においても重要なタスクである．物体特定タスクを行うためには，音声や画像，音声と画像の対応関係などを学習する必要がある．ロボットに予めこれらの情報を全て学習させておくことは困難である．家庭環境内にある全ての物体に関する音声や画像を予め収集することは，大変な手間である．また，音声と物体の対応関係は，人によって異なり，かつ，同じ人であっても同じ物体を違う表現で指示してしまうこともあるため，予め学習させておくことは非常に困難である．人が人との対話などのインタラクションを通じて，学習をしていくように，ロボットも，人とのインタラクションを通じて，自立的に学習していくことが望ましい．本研究では，ロボットが物体特定タスクを通じて，自立的に学習していくことを目標とする．目標を達成するに当たっての問題点を明記し，各問題についての解決策を提案する．

人が物体を指示する際，物体の表現方法は人それぞれである．ある人は，ある物体を「コップ」と表現するかもしれないし，ある人は「赤いコップ」と表現するかもしれない．物体名で指示するときもあれば，色などの属性を用いるときもある．人による表現の違いを考慮した物体の特定手法が必要となる．そこで，第 3 章では，複数の属性に基づいた表現による物体特定手法を提案する．人が物体名のみ，もしくは属性と物体名から成る表現を用いて，音声で物体を指示すると，音声と物体画像のマルチモーダル情報を統合して，ロボットが複数物体の中から指示された物体を特定するための手法を提案する．本研究で扱った表現は，物体名のみ，色属性のみ，または，色属性と物体名を用いた表現の 3 種類である．例えば，「赤いのとって」「赤いボールとって」というような表現を用いる．複数の物体があるシーンを被験者に見せ，ロボットに指示する想定で，指示をしてもらい，実際に物体特定が可能であるのかの評価を行った．

人がある物体を指示したいとき，物体名を知らなくとも，物体の色は表現できるときが多いだろう．しかし，物体の色に関する研究はまだ少ない．従来の研究は，物体の特徴として，物体の色を用いるものが大半を占める．物体自体の色を人が思うように認識するた

めの議論は十分になされていない．そこで，第 4 章では色属性のコンピュータビジョンにおける扱い方を議論する．ロボットの視覚属性と人の視覚属性は異なる．ロボットが物体領域全域を見て色を見るのに対し，人は物体で最も顕著な部分の色を見て，その物体が何色であるか判断する．人の視覚属性を考慮したコンピュータビジョンによる色属性モデルが必要とされる．本研究の目的は，人の視覚属性を考慮することで，複数の色で構成された物体であったとしても，人がその物体を見て思う 1 色で，ロボットに物体を特定させることである．本研究では，色による物体特定のための顕著性検出手法を提案する．

人と人とのインタラクションを考えたとき，人は相手が知っているもののみを相手に指示するだけではない．相手が知らないものも指示するときがある．人は指示された物体が既知か未知かをまず判別する．未知であった場合，人はその未知物体を学習していく．未知物体の学習の第一歩として，未知物体を検出する手法を第 5 章で提案する．未知物体の検出には，マルチモーダル情報を用いる．本研究では，マルチモーダル情報を用いた認識を拡張することで，未知物体の検出を可能とする．

物体と音声の対応関係が既知であるか，未知であるのかの判定は可能となった．しかし，対応関係が未知である場合，何が未知であったから，対応関係が未知であったのかの詳細な判定まではすることができない．例えば，音声のみが未知である場合，物体のみが未知である場合，両者が未知である場合，音声と物体は既知であるが，対応関係のみが未知である場合などがある．学習を行うことを考慮すると，何を学習すべきか，つまり何が未知であったのかを知ることは重要である．第 6 章では，物体と音声の対応関係が未知であった場合に，何が未知であったのかを判定する．これにより，各場合において有効な学習方法を選ぶことができる．

第 7 章では，今後の展望と，本研究の結論を述べる．

目次

第 1 章　まえがき　1

第 2 章　マルチモーダル，認識，グラウンディングの関係性　5
2.1　マルチモーダル　5
2.1.1　マルチモーダル処理　5
2.1.2　マルチモーダル技術　6
2.2　認識　6
2.2.1　パターン認識　7
2.2.2　マルチモーダル情報を用いた認識　7
2.3　グラウンディング　8

第 3 章　マルチモーダル情報を用いた既知物体の特定　11
3.1　特定物体認識に基づく物体特定　11
3.1.1　あらまし　11
3.1.2　関連研究　12
3.1.3　タスクの設定　13
3.1.4　1 属性の表現を用いた物体特定　13
3.1.5　複数属性の表現を用いた物体特定　20
3.1.6　物体特定データセット　21
3.1.7　物体特定実験　25
3.1.8　まとめと今後の課題　27
3.2　一般物体認識に基づく物体特定　28
3.2.1　あらまし　29
3.2.2　タスクの設定　29
3.2.3　一般物体認識に基づく物体特定　30
3.2.4　物体特定実験　32
3.2.5　まとめと今後の課題　34

第 4 章 色属性による物体特定のための顕著性 35
4.1 あらまし 35
4.2 関連研究 36
4.3 問題設定 37
4.4 物体特定のための色顕著性 39
4.4.1 色名称との相関度 40
4.4.2 色顕著性を用いた物体特定 41
4.5 実験 42
4.5.1 データセット 42
4.5.2 色顕著性の考察 43
4.5.3 色顕著性を用いた物体特定実験 43
4.6 まとめと今後の課題 46

第 5 章 マルチモーダル情報を用いた未知物体識別 47
5.1 あらまし 47
5.2 タスクの設定 49
5.3 マルチモーダル情報を用いた未知物体識別手法 50
5.3.1 認識信頼度 51
5.3.2 ロジスティック回帰を用いたモダリティ統合 52
5.3.3 未知物体識別手法 54
5.3.4 複数の物体がある場合の未知物体識別 55
5.4 実験 57
5.4.1 未知物体検出手法の評価 58
5.4.2 物体認識の評価 60
5.5 まとめと今後の課題 60

第 6 章 未知物体学習のための未知モダリティ検出 61
6.1 あらまし 61
6.2 問題設定 63
6.3 未知モダリティ検出 64
6.3.1 未知モダリティ 65
6.3.2 未知モダリティ検出のための識別器 65
6.3.3 クラス分類アルゴリズム 66
6.3.4 実験 67
6.3.5 まとめと今後の課題 70

6.4 Web 画像を用いた物体特定手法 71
6.4.1 Web 画像による画像モデルの構築 72
6.4.2 実験 72
6.4.3 まとめと今後の課題 74

第 7 章 おわりに 75

付録 79

謝辞 83

参考文献 85

奥付 95

図目次

3.1 Deterministic finite automaton of expression based on one attribute. 15
3.2 Algorithm for speech-recognition likelihood of each object. 16
3.3 Algorithm for object identification based on an attribute. 18
3.4 Algorithm for construction of the training dataset. 19
3.5 Deterministic finite automaton of expression based on two attributes. 20
3.6 Algorithm for object identification based on two attributes. 21
3.7 Examples of dataset 1. 23
3.8 Examples of dataset 2. 23
3.9 Example of segmented image. 23
3.10 Accuracy of object identification. 34

4.1 Examples of various-colored objects. 36
4.2 Examples of object identification task using color name. 38
4.3 Overview of the saliency detection of color names for object identification. 39
4.4 Relation between accuracy of the object identification and the dimension of subspace (RGB). 44
4.5 Relation between accuracy of the object identification and the dimension of subspace (color histogram). 45

5.1 Autonomous discrimination of unknown objects and their names by a robot. 48
5.2 Variation of robot behaviors. 49
5.3 Proposed system configuration diagram. 50
5.4 Joint distribution of values of speech and object confidence. 54
5.5 Cases where the input word is known. 55
5.6 Cases where the input word is unknown. 56

5.7 Matching input pairs of a speech and an image with object models. . 57
5.8 Examples of objects used in the experiment. 58
5.9 11 images of a bear taken from 11 different angles. 58
5.10 Example of information obtained by Kinect. 58

6.1 Object identification. 62
6.2 Case when speech and image are grounded. 63
6.3 Case when speech and image are not grounded. 63
6.4 Conceptual diagram of confidence distribution in 5 cases. 65
6.5 Algorithm for 5-class classification. 67
6.6 Thresholding border and confidence measure of known/unknown speech and image. Green dots denote C1, yellow dots denote C2, blue dots denote C3, and red dots denote C4. 70
6.7 Discriminant border by logistic regression and confidence measure of known/unknown speech and image. Green dots denote C1, yellow dots denote C2, blue dots denote C3, and red dots denote C4. . . . 71
6.8 Accuracy of object identification (%) 73

7.1 Effective visual field. 79

表目次

3.1 Examples of utterances in a scene. 24
3.2 Number of speech and its variations. 25
3.3 Accuracy of speech recognition(%). 25
3.4 Accuracy of object recognition (%). 25
3.5 Accuracy of object identification (%). 26
3.6 Examples of the result of object identification(%). 28

4.1 Examples of objects in dataset. 42
4.2 Saliency maps of 1 object. Colors of each object is different each other. 43
4.3 Accuracy of the object identification (%). 45

5.1 Accuracy of multiple unknown object discrimination (%). 58
5.2 Accuracy of object recognition for respective feature (%). 59

6.1 5 cases. 64
6.2 Composition of dataset. 67
6.3 Accuracy of unknown modarity detection (%). 69
6.4 Accuracy of unknown/known modarity detection by thresholding confidence measure of speech and image (%). 69
6.5 Accuracy of discrimination by logistic regression used in detection of unknown/known modarity (%). 69

7.1 Speeches of 7 participants in 15 scenes. 81

第 1 章

まえがき

少子高齢化の課題解決策のひとつとして，家庭用サービスロボットの開発が急がれている．サービスロボットの中でも，人とのコミュニケーション能力を有したコミュニケーションロボットが特に期待されている．家庭用サービスロボットが，人との言語によるコミュニケーションを通じて，家庭環境で作業する未来はそう遠くない．人は「自分」という存在を他者との関係性によって定義する社会的生物である [1]．人は一人で生きることはできず，他者と共存して初めて生きていくことができる．他者と共存する際，他者とのコミュニケーションが必要となる [2]．相手が何を言っているのかを理解し合う能力がなければ，共存して生活することができない．人とロボットの共存を考えると，人同士の共存で必要なことと同じことが必要となる．ロボットがヒト化する必要は必ずしもないが，ロボットと人が自然なコミュニケーションをとることや，ロボットが人とのコミュニケーションに対応したインタラクションを行う必要がある．

人は社会性を持ちながら，物理的世界に生きる．言語理解と同時に，言語と物理的世界とのグラウンディングを行い，コミュニケーションを成立させる [3]．よって，家庭用ロボットが人と共に生活する際，ロボットも人の言語を理解することやグラウンディングを行うことが必要となる．例えば，物体を介したコミュニケーションを行う際，その物体に関する言語情報や人の発話指示音声情報，物体の物理的観測情報等の複数のモダリティが必要となる．ここで [4, 5, 6, 7, 8, 9, 10] は人の発話音声や物体画像情報など複数のモダリティを用いることで言語とその意味を学習するという研究を行ってきた．ロボットに学習をさせる場合，事前に必要となる情報をロボットに与え，学習させるという方法もあるが，実際にロボットが家庭環境内で生活することを考えると，ロボットに学習させたいことの全てを事前に教えることは不可能であるため，ロボットが自動的にそれらの情報を学んでいくことが望ましい．

人は本などから知識を学ぶだけではなく，人との対話やインタラクションを通じて知識を学習していく．ロボットが知識を得ていく際も，人との対話やインタラクションを通じて学習していく方が自然である．近年，人との対話やインタラクションを通じてロボットが知識を学習していく手法が提案されている [11, 12, 13, 14, 15]．対話やインタラクションを通じた学習に関する研究には，幼児の学習に模した学習に関する研究と，タスクドメインの学習に関する研究とがある．幼児の学習に模した研究には [11] がある．タスクドメインの学習には，[12, 13, 14, 15] がある．本論文ではタスクドメインの学習を目指す．具体的には，人がロボットに指示をし，ロボットはそのインタラクションを通じて自然に未知物体に関する情報を学習していくことを目指す．そのインタラクションとして，机の上に複数の物体があり，ロボットは人が発話指示した物体を把持するというタスクを設定する．これを物体特定タスクと呼ぶ．タスクを達成するためには，人の発話音声とロボットがカメラでとらえた物体画像のセットを学習し，認識する必要がある．つまり，音声認識と画像認識の両方が必要となる．本研究では，音声と物体画像の共起性を考慮するため，音声情報と画像情報を統合したマルチモーダル情報を用いる．音声と共起する物体が特定対象となる．本論文では，ロボットが人との物体特定タスクを通じて，物体を学習することを目標としている．目標を達成するに当たっての問題点を明らかにし，各問題点についての解決方法を提案する．

人が物体を指示する際，物体の表現方法は人それぞれである．ある人は，ある物体を「コップ」と表現するかもしれないし，ある人は「赤いコップ」と表現するかもしれない．物体名で指示するときもあれば，色などの属性を用いるときもある．人による表現の違いを考慮した物体の特定手法が必要となる．そこで，本研究では，複数の属性に基づいた表現による物体特定手法を第 3 章で提案する．これは，人が物体名のみ，もしくは属性と物体名から成る表現を用いて，音声で物体を指示すると，音声と物体画像のマルチモーダル情報を統合して，ロボットが複数物体の中から指示された物体を特定する手法である．本研究で扱った表現は，物体名のみ，色属性のみ，または，色属性と物体名を用いた表現の 3 種類である．例えば，「赤いのとって」「赤いボールとって」というような表現を用いる．[16, 17] は物体画像を被験者に見せて，テキストでアノテーションをしてもらい，そのアノテーションと物体画像をデータセットに用いて，物体を特定する実験を行っている．[18, 19] は物体画像に予めアノテーションをしておき，それを複数の被験者に読ませて得た発話音声と物体画像をデータセットに用いて実験を行っている．しかし，このデータセットは，実際に人がロボットに指示する状況を考えると不自然である．なぜなら，人は画像ではなく実物体を見て（あるいは実物体を思い描いて），時間をかけずに瞬間的にロボットに指示するからである．本研究では複数の実物体を人に見せて，この中の 1 つの物体を音声でロボットに対して指示してもらう状況を想定する．照明条件等が同じであったとしても，物体の見えは見る位置によって違うため，人から見える物体の向きと，

ロボット（実際にはカメラ）から見える物体の向きを変えて指示する必要がある．人がロボットに指示を行うとき，必ずしも人とロボットの物体の見えが同じであるとは限らないからである．

色属性は，人が幼児期に学習する重要な概念のひとつである．人がある物体を指示したいとき，物体名を知らなくとも，物体の色は表現できるときが多いだろう．しかし，物体の色に関する研究はまだ少ない．従来の研究は，物体の特徴として，物体の色を用いるものが大半を占める．物体自体の色を人が思うように認識するための議論は十分になされていない．そこで，本研究では色属性のコンピュータビジョンにおける扱い方を議論する．色属性を用いて物体を指示する際，人は 1 色で指示を行うのが自然である．例えば，「赤で緑のものとって」という言い方よりも，「赤いのとって」という言い方をされる場合の方が多いだろう．ロボットがコンピュータビジョンを用いて，複数の色で構成された物体の色を 1 色で認識することは困難である．また，ロボットの視覚属性と人の視覚属性は異なる．ロボットが物体領域全域を見て色を見るのに対し，人は物体で最も顕著な部分の色を見て，その物体が何色であるか判断する．人の視覚属性を考慮したコンピュータビジョンによる色属性モデルが必要とされる．本研究の目的は，人の視覚属性を考慮することで，複数の色で構成された物体であったとしても，人がその物体を見て思う 1 色で，ロボットに物体を特定させることである．本論文では，色による物体特定のための顕著性検出手法を第 4 章で提案する．顕著性検出には人に指示された色との相関に基づいたマニフォールドランキングを用いている．マニフォールドランキングを用いて，[20] と同様に，グラフベースで顕著性マップを構築している．

人と人とのインタラクションを考えたとき，人は相手が知っているもののみを相手に指示するだけではない．相手が知らないものも指示するときがある．人は指示された物体が既知か未知かをまず判別する．未知であった場合，人はその未知物体を学習していく．ここでいう未知とはアウトオブボキャブラリであること，つまり対象のモデルを持っていないことを指す．同様に，既知とは対象のモデルを持っていることを指す．本研究ではインタラクションを通じた自然な学習のための第一歩として未知物体の判別手法を第 5 章で提案する．設定したタスクを達成するためには，人の発話音声とロボットがカメラでとらえた物体画像のセットを学習し，認識する必要がある．そのため，未知物体の判別には音声情報と画像情報を統合したマルチモーダル情報を用いる．本タスクを達成するには，音声認識と画像認識の両方が必要となる．このとき，音声と画像のセットには，既知音声と既知画像，未知音声と未知画像，既知音声と未知画像，未知音声と既知画像の 4 種類の組み合わせが存在する．本研究ではこの 4 種類の組み合わせが存在することに着目している．複数物体の中から発話指示された物体を選択するタスクにおいて，この 4 種類の組み合わせを判別することができれば，あるケースにおいては発話指示された物体名が未知であり，ロボットの目の前にある物体画像が未知であったとしても，指示された物体を選ぶこ

とができる．

これまでに述べてきた方法により，物体と音声の対応関係が既知であるか，未知であるのかの判定は可能となる．しかし，対応関係が未知である場合，何が未知であったから，対応関係が未知であったのかの詳細な判定まではすることができない．例えば，音声のみが未知である場合，物体のみが未知である場合，両者が未知である場合，音声と物体は既知であるが，対応関係のみが未知である場合などがある．学習を行うことを考慮すると，何を学習すべきか，つまり何が未知であったのかを知ることは重要である．本研究では，物体と音声の対応関係が未知であった場合に，何が未知であったのかを判定する手法を第 6 章で提案する．これにより，すべての場合において有効な学習方法を選ぶことができる．

第 2 章

マルチモーダル，認識，グラウンディングの関係性

2.1　マルチモーダル

2.1.1　マルチモーダル処理

マルチモーダルとは，ある情報を伝達する場合に，複数のモダリティ情報が組み合わされていることを指す．各モダリティ情報の伝達の方法はモードと呼ばれる．マルチモーダルとコミュニケーションとの関係性は深い．人は人と対話などのコミュニケーションを行う際，音声言語のみを用いるのではなく，表情やジェスチャーなど様々な情報を組み合わせることで，相互的に情報の疎通を図っている．モダリティとは，情報伝達において用いる認知的手法あるいは様式のことである [21]．マルチモーダル処理に用いられるモダリティ情報としては，視覚や聴覚など五感から得られる情報を用いることが多い [22]．

人と人のコミュニケーションにおけるマルチモーダル処理において，用いられるモダリティと，人とコンピュータ（特にロボット）のコミュニケーションで用いられるモダリティは異なる [23]．人は五感などの感覚で情報を得るのに対し，ロボットはセンサから情報を得る．例えば，人は，目から得た情報を脳に伝達して視覚情報を得るのに対し，ロボットはカメラなどのセンサから得た情報を処理することで視覚情報を得る．ロボットにとってのモダリティとは，センサから得られる情報となるため，ロボットによってモダリティが異なる．また，モダリティの定義は広く，研究によって定義が異なる．例えば，画像処理システムなどの大きな枠組みから得られる情報をモダリティと定義する場合 [8, 16] もあれば，画像の特徴量をモダリティと定義する場合 [24, 25] もある．つまり，複数の情報をもってあるひとつの情報を表しているのであれば，その場合に用いた個別の情報はモ

ダリティのひとつと言えるのである．

本研究で用いるモダリティは画像情報と音声情報の２つである．画像情報と音声情報を統合したマルチモーダル情報を用いて，マルチモーダル処理を行うことによって，認識やグラウンディングを可能とする．本研究で提案するマルチモーダル処理は，音声情報，画像情報のみならず，他のモダリティを用いた処理にも拡張可能である．例えば，触感センサから得られる触感情報や，嗅覚センサから得られる嗅覚情報なども有用ではないかと考えられる．他のモダリティを用いたマルチモーダル処理は，今後の課題のひとつである．

2.1.2　マルチモーダル技術

マルチモーダル技術とは，マルチモーダル処理を工学技術に適応した技術のことである．近年，マルチモーダル技術によって構築されたシステムが，様々な分野において開発されている．特に，人と様々な情報のやりとりを対話的に行う，マルチモーダル情報統合システムが注目されている．例として，対話に重点を置いた，マルチモーダル対話システム，人間とコンピュータとのインタフェースに重点を置いたマルチモーダルインタフェースなどがある．これらは，直接操作の基本となるグラフィックスや，言葉を表示する音声やテキストなどの複数のモダリティ情報が同時に，相互補完的に利用でき，言葉や行為のモダリティを働かせてインタラクションができるようなシステムである [26].

近年，少子高齢化問題の解決策のひとつとして，対話機能を持った家庭環境用サービスロボットの開発が求められている [27]．このようなサービスロボット (以下ロボットと呼ぶ) こそ，マルチモーダル技術の集大成といえよう．近い未来，ロボットが，人とコミュニケーションを取りながら様々な家事などを行うときが来るだろう．本研究では，ロボットが家庭環境内で，人と自然なコミュニケーションを通じて，作業や学習を行っていくことを目標とする．この際に必要な，マルチモーダル技術を提案し，その問題点についても論じる．

2.2　認識

認識の基本命題は,「ある物を何とみなすか」である [28]．ある物を観測し，観測から得られた情報を用いて，何とみなすべきであるのかの識別を行い，何とみなすのかを決定するという流れになる．人は相手にある物についての情報を伝える際，ある物について認識を行い，その認識結果を用いる．統計的なアプローチで，機械に認識を行わせるための技術がパターン認識技術と言える．パターン認識技術は，人とロボットがコミュニケーションを行う際，重要な技術のひとつとなる．

2.2.1 パターン認識

ある物を何とみなすのかは，人の主観によって決まるため，認識は主観的な処理となる．パターン認識では，人が主観的に決定したパターンを機械的に処理することとなる．パターン認識研究は大きく分けて 2 つの流れで研究が進められてきた．ひとつめは，このパターンを見つけることを目標とする研究，もうひとつは，パターンを用いて数量的に認識する方法に関する研究となる．前者は発見的な研究となり，後者は理論的研究となる傾向がある．

まず，信号や現象などのパターンが観測され，ノイズ除去など特徴抽出に向けての前処理が行われる．次に，特徴抽出が行われ，識別器によって，入力されたパターンの概念，記述などのカテゴリが出力される．識別処理の後処理として識別結果を統合し，統合した結果を用いて最終結果を出力するという流れの研究もある [18]．

代表的なパターン認識技術には，音声認識技術 [29] や画像認識技術 [30] がある．本研究では，音声認識技術と画像認識技術を用いる．

2.2.2 マルチモーダル情報を用いた認識

マルチモーダル処理を用いて，あるものや事象の認識を行う研究のことをマルチモーダル情報を用いた認識と呼ぶ．これには，マルチモーダル技術と認識技術が必要となる．マルチモーダル情報を用いた認識技術は，様々な分野において提案されている．本項では，マルチモーダル技術を用いた認識技術としての従来研究の紹介を行う．従来研究は，大きく分けて，2 種類ある．一つ目は，人に関する情報を認識する研究，二つ目は，人以外 (例えば物体) に関する情報を認識する研究である．

まず，人に関する情報を認識している研究を紹介する [31, 32, 33]．人に関する情報として，人のジェスチャーや，生体情報，感情などが例として挙げられる．Wu らは，人のジェスチャー画像 (動画) と，ジェスチャーに付随する音声を用いて，人のジェスチャー認識を行っている [31]．Ouellet らは，ロボットに人の生体認証をさせる手法を提案している [32]．生体認証には，人の顔画像と，人体画像を用いた画像認識と，人の声を用いた話者認識を用いている．Zeng らは，人の音声と画像を用いて，人の感情などの人に現れた影響を議論している論文について，レビューを行っている [33]．

次に，人以外，特に物体に関する情報を認識している研究を紹介する [34, 35]．これらの研究は，更に，大きく分けて 2 種類の研究がある．一つ目は，物体のモダリティ (画像など) と，物体を示す他のモダリティ (音声や言語など) の認識を行っている研究である．Giard らは，物体と音声の組を人はどのように認識するのかについての研究を行っている

[34]．具体的には，円形と音の組み合わせを 2 種類用意し，人にその組み合わせを見せたり，聞かせたりしたときの脳波情報を用いて認識を行っている．二つ目は，物体を複数の特徴量で表すことで，特徴量をモダリティとして捉え，複数の特徴量を用いた認識を行っている研究である．Aldoma らは，物体認識と物体の向きの認識を行うアルゴリズムを提案している [35]．物体の認識には $2D$ の画像特徴量を，物体の向きには $3D$ の画像特徴量を用いている．つまり，ひとつの物体を $2D$ と $3D$ という相異なる特徴量を用いて認識している．物体の向きには $3D$ の画像特徴量を用いている．つまり，ひとつの物体を $2D$ と $3D$ という相異なる特徴量を用いて認識している．

本研究は，人が物体を指示する声（音声）を，ロボットが認識し，認識結果から視覚的に想起した物体のモデルを用いて，目の前にある複数の物体を画像処理を用いて認識し，特定するタスクを扱う．つまり，本研究は，音声認識と物体認識を用いるマルチモーダル情報を用いた認識に関する研究とも言える．

2.3　グラウンディング

グラウンディングという概念は，Searle [36] によって提案された概念である．Searle は，人の心と脳はどのような結びつきをもっているのかの議論を通じて，人は，ある事象や現象と心の動きをひもづけており，それらのひもづけを想起することを主張している．このひもづけこそが，グラウンディングの定義である．この研究が発端となって，後に Harnad らが，文献 [3] で，シンボル (記号) のグラウンディングに関する問題の提案を行った．シンボルグラウンディング問題とは，コンピュータによる記号システム内で，シンボルがどのようにして実世界の意味と結びつけられるかという問題のことである．例えば，物体の名称を実世界の物体に結びつけるというような問題のことである．シンボルが実世界と結びついていなければ，コンピュータは意味を理解することができない．例えば「りんご」というシンボルに，「赤い」というような他のシンボルとのひもづけを行っても，実世界にある「りんご」を理解することはできない．視覚を用いて，「りんご」の形や，色，大きさなどを認識し，嗅覚を用いて，「りんご」の匂いを認識し，味覚によって味や食感を認識して初めて人と共通した「りんご」の意味理解に近づく．しかし，コンピュータの視覚や嗅覚などの属性は人のそれとは異なるため，厳密に人と同じ意味理解をコンピュータは行うことができない．人と完全に同じ意味理解を行うことはできなくとも，ある程度の意味理解を行うことができなければ，人の要求する動作をコンピュータは行うことが出来ないだろう．

Steels らは，人工知能分野におけるシンボルグラウンディングの扱い方についての議論を行っている [37]．シンボルと概念，物体の関係性についての提案を行っている．本研究は，パターン認識のアプローチを用いて，物体のシンボルグラウンディング問題の解決を

図る．本研究で扱う物体特定タスクは，物体名称や色名称などのシンボルがグラウンディングしている物体を，複数の物体の中から特定する問題であると言い換えることができる．この問題を解決するために，音声情報と画像情報のマルチモーダル情報の認識を用いる．

第 3 章

マルチモーダル情報を用いた既知物体の特定

近い将来，各家庭にロボットがいて，人とコミュニケーションをとりながらサービスする時代が来ると考えられている [38, 39]．そのときには，人がロボットに対して「コップをとって」と音声で指示すると，ロボットはテーブルに置かれた複数の物体の中から，極自然に「コップ」を見つけて持ってきてくれるであろう．しかし，このタスクは 1) 人の指示発話を認識する，2) 認識した結果をもとに探すべき対象を視覚的に想起する，3) それに対応する物体を画像上で認識するという処理を含んでいる．しかも，4) 聞き間違いや見誤りが生じても誤り訂正を行って，誤りをできるだけ減らす仕組みも必要である．本論文では，このような人が音声で指示した物体を，複数物体の中から認識するタスクを，「物体特定タスク」と呼び，このタスクに含まれている問題を明らかにすると共に，その解決法について述べる．

物体特定の際，音声認識と物体認識が必要となる．物体認識には，インスタンスレベルでの物体認識を行う特定物体認識と，カテゴリレベルでの物体認識を行う一般物体認識がある．本章では，特定物体認識に基づく物体特定手法と，一般物体認識に基づく物体特定手法を提案する．

3.1　特定物体認識に基づく物体特定

3.1.1　あらまし

「物体特定タスク」を達成するためには，発話音声を認識し，物体画像を認識する必要がある．しかし，音声認識と画像認識が共に成功するとは限らない．[18, 19] は，このよ

うな問題を解決するため，音声認識と画像認識の結果を統合し，統合した認識結果を用いて物体特定を行っている．本研究では [18, 19] と同様に音声認識結果と画像認識結果を統合して物体特定を行う．

[18, 19, 40] は，「ボールをとって」というような，物体名称を用いて指示する「物体特定タスク」を扱っている．しかし，人が物体を特定するために発話する場合，必ずしも物体名称のみを用いるとは限らない．物体の特定が難しいときは，物体名称のみならず他の表現を用いる．例えば，「赤いのとって」「赤いボールとって」というような表現を用いる [16, 17]．ここで，「赤，緑」などを色属性と呼び，「赤い，緑の」など，色属性を表現したものを色名称と呼ぶ．「赤いの」「ボール」は 1 つの属性を含んだ表現であり，「赤いボール」は 2 つの属性を含んだ表現とみなせる．本研究では，物体名称のみならず，色名称を用いた表現で物体を特定する．また 1 つの属性を用いた表現のみならず，複数の属性を含んだ表現も扱う．

物体認識を行う際，どの属性を用いて認識を行うべきかによって，用いる特徴量や，モデルの構築方法が異なる．本研究では，ベイズ識別における識別器をモデルと定義する．例えば，色名称に関するモデルは色に関する特徴量で構築すべきと考えられるが，物体名称は色によらないため，色に関する特徴量よりも，物体の形状に関する特徴量で構築する方が望ましい．どの属性を用いるべきかは，音声認識結果や画像認識結果に依存する．本研究では，音声認識結果に基づいてどの属性を用いるべきかを判断する．

本論文では，複数属性を含んだ表現を用いて人が音声で指示すると，音声と物体画像のマルチモーダル情報を統合して，ロボットが複数物体の中から指示された物体を特定するための手法を提案する．音声と物体画像の統合情報を用いて物体特定を行う．また実験では，人が実物体を見て，声で発話指示した場合にも，ロボットが物体特定可能であるかを評価する．

3.1.2　関連研究

人が指示した物体を，ロボットが複数物体の中から特定するタスクを扱った研究には以下のようなものがある．[16, 17] では，人がテキストで指示を入力し，入力されたテキストを用いて物体を特定している．ロボットが人の発話音声を認識してテキスト化すれば，[16, 17] の手法が使用可能であるが，音声認識を誤ると物体特定が不可能になるという問題がある．[18, 19] は，音声認識結果と物体認識結果を統合することで，物体特定の精度を向上させている．音声認識と物体認識が同時に誤ることは少ないと仮定し，認識誤りを互いの認識結果で解消するという考え方に基づいた手法である．本研究では [18, 19] と同様に，音声認識結果と画像認識結果を統合して物体特定を行う．

[16, 17] は物体画像を協力者に見せて，テキストでアノテーションをしてもらい，その

アノテーションと物体画像をデータセットに用いて，物体を特定する実験を行っている．[18, 19] は物体画像に予めアノテーションをしておき，それを複数の協力者に読ませて得た発話音声と物体画像をデータセットに用いて実験を行っている．しかし，このデータセットは，実際に人がロボットに指示する状況を考えると不自然である．なぜなら，人は画像ではなく実物体を見て (あるいは実物体を思い描いて)，時間をかけずに瞬間的にロボットに指示するからである．本研究では複数の実物体を人に見せて，ロボットに物体を音声で指示してもらう状況を想定する．照明条件等が同じであったとしても，物体の見えは見る位置によって違うため，人から見える物体の向きと，ロボット (実際にはカメラ) から見える物体の向きを変えて実験を行う．人がロボットに指示を行うとき，必ずしも人とロボットの物体の見えが同じであるとは限らないからである．

3.1.3　タスクの設定

ロボットは，人が発話した音声を認識し，複数の物体画像の中から，音声に対応する物体画像を認識して物体を特定する．本研究で対象としているタスクの設定条件を以下に示す．

1. 机の上に複数の物体がある．
2. 人が指示した物体は必ず机の上にある．
3. ロボットから見た物体に遮蔽はない．
4. 人は「～ 取って」という文型で，ロボットに対象となる物体を把持するよう，音声で指示する．
5. 物体を指示する表現は，色名称のみ，物体名称のみ，もしくは色名称と物体名称の両方である．
6. 物体を指示する際の表現は，人によって異なっていてもよい．

人が物体を指示するとき，1 属性のみで物体を指示するとは限らない．「赤いボールとって」というように，複数の属性を用いて物体を指示する場合もある．しかし，指示に用いる属性数は予め分かっているわけではない．本論文では，1 属性もしくは複数の属性を用いて物体を指示にする表現を許し，それに対応可能な物体特定手法を提案する．

3.1.4　1 属性の表現を用いた物体特定

本項では，色名称もしくは物体名称 (1 属性) を用いて物体を指示することにより物体特定を行う．本研究では，物体名称のみならず，色名称のモデルも含んだモデルを構築する．

3.1.4.1　1 属性のみを用いた表現による音声認識

本研究では，「ボールとって」というような連続音声を扱い，連続音声認識を行う．物体の特定には，色名称もしくは物体名称の単語尤度のみを用いる．

人が指示に用いることができる属性は，物体名称もしくは色名称である (「ボールとって」「赤いのとって」などの表現を扱う)．音声認識は，図 3.1 にような決定性有限状態オートマトンを用いて行う．図中の「$silB$」は開始無音記号を，「W_{Color}」は「赤いの」というような色名称単語を，「W_{Name}」は「ボール」というような物体名称単語を，「$W_{Request}$」は「とって」というような指示表現単語を，「$silE$」は終了無音記号を示す．

物体特定では，属性を表す単語 (物体名称や色名称) の音声認識信頼度を用いる．音声認識信頼度はベイズの定理に基づき，単語尤度を事後確率化した値である [41]．音声認識の特徴量には，短時間スペクトル分析に基づく，メル周波数ケプストラム係数を用いた．そのΔとΔΔ，そして対数パワーのΔも用いた．これらの特徴量は，音声認識ソフトウェア，Julius [42] を用いて得た．音声認識信頼度は，以下の式によって得られる [43]．

$$C_s(s;\Lambda_i) = \frac{1}{n(s)} \log \frac{P(s;\Lambda_i)}{\max\limits_u P(s;\Lambda_u)} \tag{3.1}$$

ここで，$P(s;\Lambda_i)$ は音声尤度，Λ_i は i 番目の物体の単語 Hidden Markov Model (HMM) [44] を示す．i は，物体のインデックスであり，$i = 1, \cdots M$ である．M は全物体数である．$n(s)$ は，入力音声のフレーム数，u は任意の音素系列を示す．$\max\limits_u P(s;\Lambda_u)$ は音素タイプライタ，つまり，あらゆる音素系列を許すような言語モデルを用いないときの音声認識の結果の尤度を示す．この言語モデルは音素系列の制約を仮定しないので，$\max\limits_u P(s;\Lambda_u)$ は，入力音声が与えられたときの尤度の最大値とみなすことができる．よって，$C_s(s;\Lambda_i)$ は，Λ_i がどれだけ入力音声と似ている単語かを表す.[*1]と等しい．

本研究では，人によって物体を指示する表現が異なることを許している．各物体に対して複数の表現を実験により収集して，対応リストを作成しておき，リストを用いて物体に対して音声認識を行う．例えば，ある物体を指示する際，「コップとって」と発話する人や，「緑のコップとって」と表現する人がいた場合，その物体には，「コップ」と「緑」と

[*1] $C_s(s;\Lambda_i)$ が，Λ_i の確率と，入力音声が与えられたとき音素系列の確率比と等しい．これは，以下のベイズの定理が成り立つからである．

$$P(s;\Lambda_i) = \frac{P(\Lambda_i;s)P(\Lambda_i)}{P(s)}$$

本研究では，$P(\Lambda_i)$ は定数であると仮定するため，$C_s(s;\Lambda_i)$ は，

$$\frac{1}{n(s)} \log \frac{P(\Lambda_i;s)}{\max\limits_u P(\Lambda_u;s)}.$$

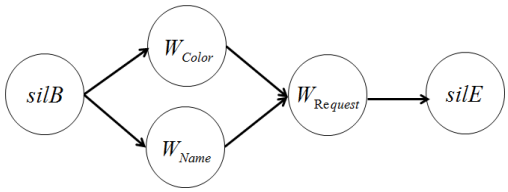


図 3.1 Deterministic finite automaton of expression based on one attribute.

いう表現を対応付けるようなリストを作成する．物体特定の際は，全表現に対して，認識を行う．物体に対して，表現が競合する場合は，最も高かった尤度を採用する．例えば，ある物体に「コップ」「マグカップ」「緑の」という表現が対応付けられている場合，色名称単語 (W_{Color}) を表す「緑の」は 1 つであるが，物体名称単語 (W_{Name}) を表す「コップ」「マグカップ」は，2 つある場合を考える．このように，物体名称が競合する場合は，尤度が高かった方をその物体の物体名称に対する尤度として採用する．具体的な 処理の流れを以下に示す．今，表現 $e(W_{Name}$ や $W_{Color})$ の音声モデルを，$\Lambda_e(e = 1, \cdots, E)$ とする．E は全表現数である．e 番目の表現の音声モデル $\Lambda_e(e = 1, \cdots, E)$ に対する入力音声 s の音声認識信頼度を $C_s(s; \Lambda_e)$ とする．

次に，表現と各物体との対応リストを用いて，物体ごとの音声認識信頼度を求める．同じ物体に対して異なる表現がある場合には，それらの中で最も大きい音声認識信頼度をその物体に対する音声認識信頼度とする．各物体に対する音声認識信頼度の求め方のアルゴリズムを図 3.2 に示す．物体 i を示す色名称，物体名称をそれぞれ m_c，m_n とする．$C_s(s; \Lambda_{m_c})$ を，物体 i の色名称に関する音声認識信頼度とし，$C_s(s; \Lambda_{m_n})$ を，物体 i の物体名称に関する音声認識信頼度とする．各物体ごとの音声認識信頼度 $C_s(s; \Lambda_i)$ は，$C_s(s; \Lambda_i) = \{C_s(s; \Lambda_{m_c}), C_s(s; \Lambda_{m_n})\}$ と定義する．$C_s(s; \Lambda_i) = C_s(s; \Lambda_{m_c})$ のときは，$C_s(s; \Lambda_{m_n}) = 0$，$C_s(s; \Lambda_i) = C_s(s; \Lambda_{m_n})$ のときは，$C_s(s; \Lambda_{m_c}) = 0$ とする．

3.1.4.2 物体認識

1 枚の画像に物体が 1 つ写っている画像 (物体画像と呼ぶ) を多数用意する．ロボットにはこの画像が複数枚入力され，物体認識を行う．入力画像を $v_n(n = 1, 2, \cdots, N)$ とし，m 番目の物体の画像モデルを $o_m(m = 1, 2, \cdots, M)$ とし，画像尤度を $P(v_n; o_m)$ とする．画像尤度は事後確率化したものを画像信頼度として用いる．信頼度は，事前確率を等確率として，以下のように事後確率として求める．

$$C_v(v_n; o_m) = \frac{P(v_n; o_m)}{\sum_m P(v_n; o_m)}. \tag{3.2}$$

画像モデル $o_m(m = 1, 2, \cdots, M)$ は，音声モデル $\Lambda_m(m = 1, 2, \cdots, M)$ と対応関係にあり，色名称と物体名称のモデルが含まれる．色名称に関する画像モデル (色名称モデルと呼

Input: $s,\ \Lambda_e(e = 1, \cdots, E),$ Correspondence table of expression and object.
Output: $C_s(s; \Lambda_m)(m = 1, \cdots, M)$

for m_c, m_n in range$(0, M)$:
    $C_s(s; \Lambda_{m_c}) \leftarrow 0$
    $C_s(s; \Lambda_{m_n}) \leftarrow 0$

for e in range$(1, E)$:
    if e-th expression is based on color name:
        Get k indices of objects corresponded to e-th expression $(m_1, \cdots m_k)$
        from the correspondence table.
        for m_c in $(m_1, \cdots m_k)$:
            if $C_s(s; \Lambda_{m_c}) = 0$
                $C_s(s; \Lambda_{m_c}) \leftarrow C_s(s; \Lambda_e)$
            else:
                if $C_s(s; \Lambda_{m_c}) \leq C_s(s; \Lambda_e)$
                    $C_s(s; \Lambda_{m_c}) \leftarrow C_s(s; \Lambda_e)$
                 else:
                    continue:
    else :
        Get k indices of objects corresponded to e-th expression $(m_1, \cdots m_k)$
        from the correspondence table.
        for m_n in $(m_1, \cdots m_k)$:
            if $C_s(s; \Lambda_{m_n}) = 0$
                $C_s(s; \Lambda_{m_n}) \leftarrow C_s(s; \Lambda_e)$
            else:
                if $C_s(s; \Lambda_{m_n}) \leq C_s(s; \Lambda_e)$
                    $C_s(s; \Lambda_{m_n}) \leftarrow C_s(s; \Lambda_e)$
                else:
                    continue:

図 3.2　Algorithm for speech-recognition likelihood of each object.

ぶ) を o_{m_c}，物体名称に関する画像モデル (物体名称モデルと呼ぶ) を o_{m_n} とすると，$o_m \in \{o_{m_c}, o_{m_n}\}$．各物体に対する画像信頼度は，$C_v(v_n; o_m) = \{C_v(v_n; o_{m_c}), C_v(v_n; o_{m_n})\}$ と定義する．各物体に対する画像信頼度は音声信頼度と同様，同じ物体に対して複数の表現が対応付けられている場合は，最も尤度が高かった尤度を採用する．

3.1.4.3 1 属性を含んだ表現によるマルチモーダル情報を用いた物体特定

物体認識では，全物体画像モデルに対して，入力された物体画像の尤度を算出し，最大尤度の物体画像モデルを認識結果としている．一方，物体の特定では，一般的に音声認識結果候補に対応する物体画像モデルを用いて，入力された複数の物体画像の尤度を算出する．このうち，最大の尤度を持つ物体画像を認識結果としている．

本研究では，音声認識結果の信頼度と画像認識結果の信頼度をロジスティック回帰により統合し，統合した結果を用いて物体特定を行っている．2 つの認識結果を統合することで，物体特定精度が高くなることを実験により確かめる．j 番目の物体の i 番目のモデルに対する音声認識結果から得られた信頼度を $C_s(s; \Lambda_{i_c})$，$C_s(s; \Lambda_{i_n})$，画像尤度から得られた信頼度を $C_v(v_j; o_{i_c})$，$C_v(v_j; o_{i_n})$ とすると，次の 2 つのベクトル $\mathbf{C_c}_j^i$，$\mathbf{C_n}_j^i$ が求められる．

$$\mathbf{C_c}_j^{i\,\mathrm{T}} = (1, C_s(s; \Lambda_{i_c}), C_v(v_j; o_{i_c})). \tag{3.3}$$

$$\mathbf{C_n}_j^{i\,\mathrm{T}} = (1, C_s(s; \Lambda_{i_n}), C_v(v_j; o_{i_n})). \tag{3.4}$$

ロジスティック回帰は以下の式で表される．

$$F(\mathbf{C_c}_j^i) = \frac{1}{1 + \exp\{-\alpha_{\mathbf{c}}{}^{\mathrm{T}} \mathbf{C_c}_j^i\}}. \tag{3.5}$$

$$F(\mathbf{C_n}_j^i) = \frac{1}{1 + \exp\{-\alpha_{\mathbf{n}}{}^{\mathrm{T}} \mathbf{C_n}_j^i\}}. \tag{3.6}$$

ここで，$\alpha_{\mathbf{c}}{}^{\mathrm{T}} = (\alpha_{c_0}, \alpha_{c_1}, \alpha_{c_2})$，$\alpha_{\mathbf{n}}{}^{\mathrm{T}} = (\alpha_{n_0}, \alpha_{n_1}, \alpha_{n_2})$ とする．物体特定の結果，特定された入力画像を $\hat{n}$，特定された物体モデルを $\hat{m}$ とする．ここで，物体モデルとは，音声モデルと物体画像モデルの組と定義する．図 3.3 に 1 属性のみを用いた場合の物体特定アルゴリズム示す．入力された 1 音声に対する全音声モデルによる音声認識結果の信頼度と，入力された複数物体画像に対する全物体モデルによる画像認識結果の信頼度の組が入力となる．i 番目の音声モデルによる音声信頼度と，j 番目の物体画像に対する i 番目の物体画像モデルによる画像信頼度をロジスティック回帰で統合し，その値を $f[i][j]$ とする．全ての候補に対して，$f[i][j]$ を算出し，最大となるときの j を物体特定の結果，特定された入力画像 $\hat{n}$ とし，i を特定された物体モデル $\hat{m}$ とする．

Input: ${\mathbf{C_c}}_j^i$, ${\mathbf{C_n}}_j^i$ $(j = 1, \cdots, N, i = 1, \cdots, M)$
Output: $\hat{n}, \hat{m}$

for j in range$(1, N)$:
　for i in range$(1, M)$:
　　if $C_s(s; \Lambda_{i_n}) = 0$:
　　$f[i][j] \leftarrow F({\mathbf{C_c}}_j^i)$
　　else:
　　$f[i][j] \leftarrow F({\mathbf{C_n}}_j^i)$

$$\hat{n}, \hat{m} = \underset{1 \le j \le N, 1 \le i \le M}{\operatorname{argmax}} f[i][j]$$

図 3.3　Algorithm for object identification based on an attribute.

3.1.4.4　ロジスティック回帰の学習

次に, ロジスティック回帰のパラメータ $({\alpha_{\mathbf{c}}}^{\mathrm{T}} = (\alpha_{c_0}, \alpha_{c_1}, \alpha_{c_2}), {\alpha_{\mathbf{n}}}^{\mathrm{T}} = (\alpha_{n_0}, \alpha_{n_1}, \alpha_{n_2}))$ の学習方法について述べる．$\alpha_{\mathbf{c}}$ と $\alpha_{\mathbf{n}}$ の学習方法は同じである．まず例として，$\alpha_{\mathbf{c}}$ の学習方法を述べる.

s^j を入力音声, v^j を入力物体画像とする．ここで s^j, v^j は, j 番目の学習サンプルであるとする．ロジスティック回帰の係数は, 音声と画像の信頼度の組 $\{C_s(s_j; \Lambda_{i_c}), C_v(v_j; o_{i_c})\}$ と教師信号 $d_{i,j} = \{0, 1\}$ によって学習される．M モデル，K サンプルの学習データセットを T とすると,

$$\begin{aligned} \mathrm{T}^{M \times K} = \{ & C_s(s_j; \Lambda_{i_c}), C_v(v_j; o_{i_c}), d_{i,j} \\ & (i = 1, \cdots, M \ , j = 1, \cdots, K)\} \end{aligned} \tag{3.7}$$

音声と画像の信頼度の組を音声と画像の依存関係を考慮して分類すると，以下の 4 つのパターンに分類できる.

1. s^j と v^j が共に物体 i の音声と画像である場合
2. s^j と v^j が共に物体 i の音声と画像でない場合
3. s^j は物体 i の音声であるが，v^j は物体 i の画像でない場合
4. s^j は物体 i の音声ではないが，v^j は物体 i の画像である場合

教師信号 $d_{i,j}$ は次のように指定する．パターン (1) の場合は $d_{i,j} = 1$ に，パターン (2),

Input: s^j, v^j $(j = 1, \cdots, K)$
Output: T

for j in range$(1, K)$:
  for i in range$(1, M)$:
    if s^j and v^j are speech and image of object i, respectively:
      T $\leftarrow C_s(s_j; \Lambda_{i_c}), C_v(v_j; o_{i_c})$, $d_{i,j} = 1$: Pattern 1
      for l in range(1, M):
        if $l \neq i$:
          T $\leftarrow$ add $\{\{(C_s(s_j; \Lambda_{i_c}), C_v(v_j; o_{l_c}))\}, d_{i,j} = 0\}$: Pattern 3
          T $\leftarrow$ add $\{\{(C_s(s_j; \Lambda_{l_c}), C_v(v_j; o_{i_c}))\}, d_{i,j} = 0\}$: Pattern 4
          T $\leftarrow$ add $\{\{(C_s(s_j; \Lambda_{l_c}), C_v(v_j; o_{l_c}))\}, d_{i,j} = 0\}$: Pattern 2
    else:
      continue:

図 3.4 Algorithm for construction of the training dataset.

(3)，(4) の場合は，$d_{i,j} = 0$ とする．

学習データセットの構築アルゴリズムを図 3.4 に示す．

次にこのロジスティック回帰の学習データセットを用いてロジスティック回帰のパラメータを推定する手法について述べる．学習データセットの対数尤度は以下のように表される．

$$P(\mathbf{d}|\alpha_{\mathbf{c}}) = \sum_{j=1}^{K}\sum_{i=1}^{M}\{d_{i,j}{\alpha_{\mathbf{c}}}^{\mathrm{T}}{\mathbf{C}_{\mathbf{c}}}_j^i - \log(1 + \exp({\alpha_{\mathbf{c}}}^{\mathrm{T}}{\mathbf{C}_{\mathbf{c}}}_j^i))\}. \tag{3.8}$$

ここで，${\mathbf{C}_{\mathbf{c}}}_j^i$ は，j 番目の物体の i 番目のモデルに対する信頼度であり $(1 \leq j \leq K,\ 1 \leq i \leq M)$，$\mathbf{d} = (d_{1,1}, \cdots, d_{M,K})$ である．ロジスティック回帰のパラメータ ${\alpha_{\mathbf{c}}}^{\mathrm{T}} = (\alpha_{c_0}, \alpha_{c_1}, \alpha_{c_2})$ はフィッシャースコアリングアルゴリズムを用いた最尤推定によって求める [45, 46]．

${\alpha_{\mathbf{n}}}^{\mathrm{T}}$ も，同様にして，$\{C_s(s_j; \Lambda_{i_n}), C_v(v_j; o_{i_n})\}$ と教師信号 $d_{i,j} = \{0, 1\}$ によって学習する．

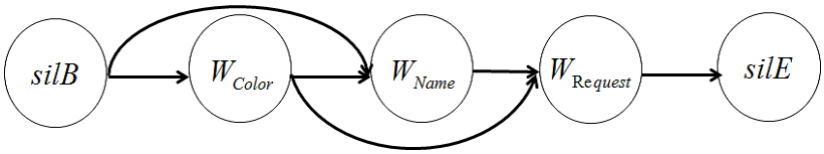


図 3.5　Deterministic finite automaton of expression based on two attributes.

3.1.5　複数属性の表現を用いた物体特定

3.1.4 項では，人が物体を指示する表現には 1 属性のみ含まれるという制約をおいていた．本章ではこの制約を緩和し，人は，1 属性または複数属性を含む表現を用いて物体特定を行う．

3.1.5.1　複数属性の表現による音声認識

物体特定時に入力される音声は，1 属性を含んだ表現 (物体名称，または，色名称)，もしくは，複数属性を含んだ表現 (色名称と物体名称を組み合わせた表現) を用いた指示発話になる．つまり，「ボールとって」，「赤いのとって」，「赤いボールとって」という音声が入力される．「ボール，赤いのとって」という表現も考えられるが，名詞，形容詞という順は不自然である．本研究では，複数属性を含んだ表現である場合，形容詞，名詞という流れのみを許すものとする．すなわち，「ボール，赤いのとって」は許さず，「赤いボールとって」のみを許す．入力される音声は，1 属性を用いたものであるか，複数属性を用いたものであるか，分からないため，提案手法では図 3.5 のような決定性有限状態オートマトンを用いて音声認識を行う．図中の「$silB$」，「W_{Color}」，「W_{Name}」，「$W_{Request}$」，「$silE$」は，図 3.1 と同様である．図 3.5 を用いて，各表現に対する音声認識を行った後に，3.1.4 項と同様にして，各物体に対する音声信頼度を算出する．

3.1.5.2　複数属性を含んだ表現を用いたマルチモーダル情報統合による物体特定

物体特定に用いる音声認識は，3.1.4.1 の方法で行い，物体認識手法は，3.1.4.2 と同様の方法で行う．3.1.4 項では，1 属性のみしか扱わなかったため，入力される音声は，色名称のみを用いたものか，物体名称のみを用いたもののみであった．本項では，色名称と物体名称の両方を用いた音声にも対応可能となる物体特定手法について述べる．

図 3.6 に複数属性を用いた場合の物体特定アルゴリズムを示す．音声認識の結果候補は，色名称のみを用いた場合，物体名称のみを用いた場合，両方を用いた場合に分けられる．i 番目のモデルによる音声信頼度と，j 番目の物体画像に対する i 番目のモデルによる画像信頼度をロジスティック回帰で統合し，その値を $f[i][j]$ とする．音声認識結果の候

Input: $\mathbf{C_c}^i_j$, $\mathbf{C_n}^i_j$ $(j = 1, \cdots, N, i = 1, \cdots, M)$
Output: $\hat{n}, \hat{m}$

for j in range$(1, N)$:
  for i in range$(1, M)$:
    if $C_s(s; \Lambda_{i_n}) = 0$:
      $f[i][j] \leftarrow F(\mathbf{C_c}^i_j)$
    else if $C_s(s; \Lambda_{i_c}) = 0$:
      $f[i][j] \leftarrow F(\mathbf{C_n}^i_j)$
    else:
      $f[i][j] \leftarrow F(\mathbf{C_c}^i_j) + F(\mathbf{C_n}^i_j)$
$\hat{n}, \hat{m} = \underset{1 \leq j \leq N, 1 \leq i \leq M}{\operatorname{argmax}} f[i][j]$

図 3.6 Algorithm for object identification based on two attributes.

補が，色名称のみ，もしくは物体名称のみを用いた表現である場合 (1 属性) は，3.1.4 項と同じ方法となる．すなわち，音声認識結果の候補が i 番目の物体モデルであり，かつ，色名称のみを用いている場合は，$f[i][j] = F(\mathbf{C_c}^i_j)$ となり，物体名称のみを用いている場合は，$f[i][j] = F(\mathbf{C_n}^i_j)$ となる．音声認識結果候補が両名称を用いている場合は，色名称のみを用いた表現による物体特定と，物体名称のみを用いた表現による物体特定を同時に行っていることになる．そこで，両名称を用いている場合は，色名称を用いている場合の物体特定結果と，物体名称を用いている場合の物体特定結果を次式で統合する．

$$f[i][j] = F(\mathbf{C_c}^i_j) + F(\mathbf{C_n}^i_j) \tag{3.9}$$

全ての候補に対して，$f[i][j]$ を算出し，最大となるときの j が物体特定の結果，特定された入力画像 $\hat{n}$，i が特定された物体モデル $\hat{m}$ となる．

3.1.6 物体特定データセット

本研究で扱うタスクに対する提案手法の有効性を実験的に示すために，データセットを構築する．データセットは人の発話指示データセットと，実験に用いる物体の物体画像データセットから構成される．

3.1.6.1　発話指示データセット

45 物体の実物体を用意し，7 名 (女性 3 名，男性 4 名) の協力者に発話を依頼した．45 物体の中から重複なく 3 物体を選び，机の上に置いた．全部で 15 通りになる (3 物体 ×15 組=45 物体)．机の上に置かれた 3 物体の中から 1 物体を選び，その物体を指示するための発話をしてもらった．つまり，1 組 (3 画像) につき，各物体について 1 発話ずつ，合計 3 種類の発話をしてもらった．協力者には以下の 2 点を守って，発話してもらった．1 点目は，「〜 とって」という定型で，声で指示すること．2 点目は対象の物体をロボットがとれるように，色名称，物体名称を用いて指示すること．指示は，色名称のみ，物体名称のみ，もしくは色名称と物体名称の両方を含んだ表現で発話すること．

本研究では，人が物体を見て，すぐにロボットに指示するタスクを扱うため，人が瞬時に物体を識別可能である有効視野内に複数の物体を置いた [47]．具体的な有効視野の算出方法を付録 A に示す．机の上に 3 物体を $30cm$ 間隔で配置し，中心の物体の重心位置を視野中心として，各被験者の目の高さ，机の高さを考慮して，被験者ごとに机までの距離を算出し，実験を行った．

以下の手順で発話指示データを収集した．

1. 協力者ごとに算出した位置に椅子を移動させ，協力者に座ってもらう．
2. 協力者に目隠ししてもらう．
3. 机の上の指定位置に 3 物体をランダムな向きで並べる．
4. 協力者に目隠しをとってもらう．
5. 指示してほしい対象の物体を操作者がポインタで示し，協力者はポインタで指された物体をロボットに指示すると想定して，「〜 とって」という形で発話してもらう．発話指示は，ポインタで物体を示されてから，3 秒以内に答え始めてもらう．
6. 指示していない物体がなくなるまで，5. を繰り返す．
7. 3 物体を指示し終わったら，2. に戻り，次の 3 物体の組で，指示実験を行う．以上を 15 組繰り返す．

ここで，操作者が物体をポインタで指示する場合，3 物体をランダムな順番でポインタで示すこととした．

3.1.6.2　物体画像データセット

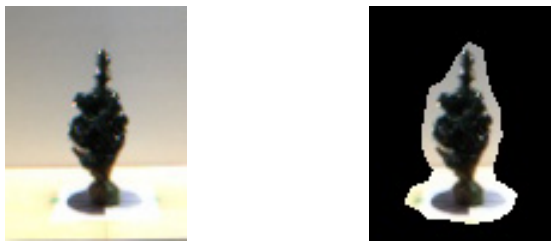
発話指示データセットの収録と同じ環境で，45 物体の各物体について回転台で 36 度ずつ回転させて各物体につき 10 枚，物体画像を撮影した．発話指示データセット構築時に視野中心とした机の中心に物体を置き，回転させた．これらの物体画像を画像データセット 1 とする．また，発話指示データセット構築時に用いた 15 組 (各 3 物体) の物体を，各



図 3.7　Examples of dataset 1.



図 3.8　Examples of dataset 2.



RGB image　　Object region

図 3.9　Example of segmented image.

組 1 枚ずつ合計 15 枚撮影した．3 物体を置くときの物体の向きはランダムに設定した．これらの物体画像データを画像データセット 2 とする．図 3.7 にデータセット 1 の物体画像例を，図 3.8 にデータセット 2 のシーン画像例を示す．

カメラ位置は高さ 119cm(協力者の目の高さの平均と同じ), 物体までの距離は 112cm(協力者の距離の平均と同じ) に設定した．撮影には，Point Grey 社 [48] の近距離撮影用複眼カメラ Bumblebee2(BB2-03S2) を使用した．複眼カメラの視差より，デプス画像を求め，デプス画像を閾値処理することで得られたマスクを用いて，物体領域を抽出した．図 3.9 に撮影した物体画像例と領域抽出された物体画像例を示す．

3.1.6.3　データセット評価

本研究で構築したデータセットの評価を行う．発話指示データセットの評価として，発話指示結果の解析と発話の音声認識実験を行う．物体画像データセットの評価として，画像認識実験を行う．

まず，実際に協力者によって発話された内容について評価する．次に，各発話の音声認

表 3.1　Examples of utterances in a scene.

Example of a scene	Participant	Object 1	Object 2	Object 3
	1	Bring me a yellow one.	Bring me a bucket.	Bring me a red book.
	2	Bring me a yellow one.	Bring me a bucket.	Bring me a book.
	3	Bring me a shampoo.	Bring me a red bucket.	Bring me a book.
	4	Bring me a pomp bottle.	Bring me a red bucket.	Bring me a book.
	5	Bring me a yellow one.	Bring me a bucket.	Bring me a red book.
	6	Bring me a yellow dispenser.	Bring me bucket.	Bring me a red book.
	7	Bring me a shampoo.	Bring me a bucket.	Bring me a book.

識精度を評価する．

実際に行われた発話の例を表 3.1 に示す．表 3.1 中の発話は英語表記されているが，実際には日本語で 7 人の協力者に発話してもらった．表 3.1 は，協力者に見せた 3 物体 ×15 組のうち，ある 1 組 (表 3.1 左の写真) についての結果である．協力者によって用いる属性が異なっており，たとえ同じ属性を用いていたとしても，表現が異なる場合がある．残りの 14 シーンを含む，全シーンにおける全協力者の発話を付録 B に示す．発話指示データセット中の全発話数，発話の種類の数，種類ごとの発話数，発話された表現のバリエーション数を表 3.2 に示す．実験に用いた 45 物体は，全て相異なる物体であるため，本来は物体名のみで指示可能である．よって，物体名称を用いた発話数が一番多かった．次に多かったのは両名称を用いた指示であった．これは，協力者が物体名称に自信がなかった場合，色名称と物体名称の両方を用いたためと考えられる．

色名称のみを含む表現で用いられた色名称と，両属性を含む表現で用いられた色名称を合わせた場合の表現のバリエーション数は 19 であった．発話された色名称の例として，「黄色いの，黄色の，赤い，白い，緑の」などがあった．物体名称のみを含む表現で用いられた物体名称と，両名称を含む表現で用いられた物体名称を合わせた場合の表現のバリエーション数は，合計 70 であった．発話された物体名称の例として，「帽子，ちりとり，なべ，写真立て，ぬいぐるみ，マグカップ」などが挙げられる．色名称と物体名称の表現のバリエーション数を比べると，物体名称のバリエーション数の方がはるかに大きいことが分かる．

次に発話 (音声) の音声認識精度を評価した．音声認識は，音声認識ソフトウェア Julius [42] を用いて，図 3.5 に示した決定性有限状態オートマトンに基づいた不特定話者音声認識を行った．音響モデルには，Julius ディクテーションキット [49] に含まれている，不特定話者の PTM トライフォン HMM(Hidden Markov Model) を用いた．特徴ベクトルとして，メル周波数ケプストラム係数 (MFCC)12 次元とその Δ，対数パワーの合計 25 次元のベクトルを用いた．

発話の種類ごとに，発話の音声認識精度を表 3.3 に示す．物体名称のみを用いた場合の認識精度が一番高い．音声認識の誤り例を示す．(正)「ピンクのおけ」→ (誤)「ピンクの

表 3.2　Number of speech and its variations.

	Object name	Color name	Both names	Total
Number of speeches	224	37	54	315
Number of variations	65	9	41	115

表 3.3　Accuracy of speech recognition(%).

Object name	Color name	Both names	Average
96.9	89.2	92.6	92.9

表 3.4　Accuracy of object recognition (%).

Feature	Dataset 1	Dataset 2
Fourier	78.8	88.8
Lab	76.0	66.6

なべ」，(正)「ティッシュ」→ (誤)「クッション」，(正)「ぞう」→ (誤)「じょうろ」.

3.1.6.2 で構築した，データセット 1 とデータセット 2 を評価するため，物体認識実験を行った．画像から物体領域を切り出した画像を用いて，45 物体認識を行った (実験に用いた物体数は 45 物体)．データセット 1 の実験では，データセット 1 の中でテストデータと学習データを選び，leave-one-out-crossvaridation により物体ごとの物体認識精度を評価した．データセット 2 の実験では，テストデータにデータセット 2 の画像を，学習データにはデータセット 1 の画像を用いて行った．データセット 2 の 1 枚の画像には 3 物体が写っている．評価では 1 枚の画像から 3 物体の領域を抽出し，物体ごとに物体認識精度を評価した．物体認識は 2 種類 (輪郭のフーリエ級数 (8 次元) [50]，Lab [51] の平均値 (3 次元)) の特徴量を用いて実験を行った．物体ごとに正規分布で物体画像モデルを構築し，画像尤度を求めた．表 3.4 に物体認識結果を示す.

3.1.7　物体特定実験

物体特定実験では，発話された属性は全て既知であり，音声と画像はオープン条件で評価した．ここで，属性が既知である状態とは，属性のモデルが学習されている状態を指す．机の上にある 3 物体から，音声で指示された物体を正しく特定できるかの評価を行った ($N = 3$).

実験は 2 種類のデータセットを用いて，2 種類の実験を行った．データセット 1 のみを

表 3.5　Accuracy of object identification (%).

Name	Color name	Object name	Both names
Experiment 1	81.8	91.8	92.5
Experiment 2	83.6	89.4	92.3

用いて行う実験 (実験 1) と，データセット 2 をテストデータに，データセット 1 を学習データに用いる実験 (実験 2) を行った．

実験 1 では，各物体につき 3 枚の物体画像をテストデータに，それ以外のデータ 7 枚は物体画像モデルとロジスティック回帰の学習データに用いた．実験 2 では，物体特定実験 1 回につき，画像のテストデータから，3 物体 (15 シーン中の，あるシーンの 3 物体) 各 1 枚の物体画像を選択し，その 3 物体のうちの 1 物体を指示する発話音声を 7 音声 (協力者 7 人) から 1 音声を選択し，音声と対応する物体画像を正しく選べたか否かを評価した．実験 1 の物体特定実験では，15 シーン $\times 3^3$ 通り (1 シーンにおける画像データ 3 枚の選び方) $\times$21 発話 (各シーンでの発話数：7 協力者 $\times$3 物体)=8505 回実験を行った．実験 2 の物体特定実験では，15 シーン $\times$21 発話=315 回，実験を行った．

音声モデルは，前項同様 Julius のディクテーションキットに付属している不特定話者 PTM トライフォン HMM を用いた．色名称の物体画像モデルは，物体領域の Lab の平均値 (3 次元) を特徴量とし，物体名称の物体画像モデルは，物体の輪郭のフーリエ級数 8 次元を特徴量として，各属性，各物体毎に正規分布で構築した．物体画像モデルの学習には，テストデータ以外の画像を各物体につき 5 枚ずつランダムに選択して用いた．ロジスティック回帰の学習は，テストデータに選ばなかった発話と，テストデータにも物体画像モデルの学習データにも用いなかった画像 (各物体につき 2 枚ずつの物体画像) を用いて行った．

表 3.5 に物体特定の結果を示す．色名称のみを用いて発話指示した場合と，物体名称のみを用いて発話指示した場合と，色名称と物体名称の両方を用いた発話指示をした場合の物体特定精度を示している．実験結果より，1 属性のみを用いた物体特定精度よりも，両属性を用いた物体特定精度の方が高くなっていることが分かる．これにより，ロボットに物体を特定させる際，物体に関する属性情報を多く与えた方が特定精度が高くなることが分かる．また，色名称のみを用いた場合と，物体名称のみを用いた場合を比べると，物体名称のみを用いた場合の方が精度が高くなっていることが分かる．この理由は，データセット構築に用いた 45 物体のうち，物体自体は相異なるが，互いによく似た色名称を持つ物体が存在していることが原因であると考えられる．

表 3.6 に，実験 2 における各シーンごとの物体特定結果の例を示す．シーン画像中に写

る 3 物体を，左から順に物体 1，物体 2，物体 3 とする．シーン画像を，上から順に，シーン 1，2，···，5 とする．空白 (–) は，その属性での指示を協力者が行わなかったケースである．シーン 1 は，全物体において複数属性を用いた方が精度が高かった場合である．これは，3 物体の色も，形状も互いに似通ってなかったことが理由として考えられる．他のシーンは，必ずしも複数属性を用いる方がよかったわけではなかった例である．シーン 2 は物体 1 のみが明らかに他の 2 物体と異なり，物体 2 と 3 の色が似通っているシーンである．物体 2 と 3 を指示する場合，色での指示のみでは困難であると協力者が考え，色のみでの指示を行わなかったことが分かる．物体 1 と他の 2 物体を比べると，色も形状も異なる物体 1 の精度が高かった．シーン 3 は 3 物体の形状が似てはいるが，色が相異なるケースである．色のみでも指示可能であるが，物体 3 の精度は 0% となってしまった．多くの場合，物体 3 を物体 2 と間違っていた．人が目で見ると色が異なっているように見えるが，コンピュータ処理では間違えてしまうことが分かる．シーン 4 は，全てほぼ青と白の 2 色で構成されている 3 物体のシーンである．人により，色の指示発話が異なっていた．そのため，物体 1 を色のみで指示した場合は 0% になってしまった．シーン 5 は，物体 2 のみが明らかに色が異なり，他の 2 物体の色が似ているケースである．物体 2 は色のみで指示した場合が最も精度が高かった．物体 1 を色のみで指示したケースは精度が 0% になってしまった．多くの場合，物体 1 を 3 と間違えていた．物体 3 は，両属性を用いたものが一番精度が高かった．

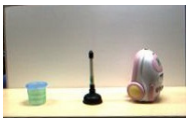
実験により，複数属性を用いる方が，物体特定の精度が高くなることが分かった．特に，指示対象と似通った属性を持つ物体が他にもある場合に，有効であることが分かった．

3.1.8　まとめと今後の課題

本章では，色名称と物体名称の 2 属性を用いた物体特定手法を提案した．音声認識と画像認識の結果をロジスティック回帰で統合し，統合した値を用いて物体特定を行った．また，ある物体をどのように表現するかは人によって異なることに着目し，人によって表現が異なる場合でも，物体特定ができるような枠組みを提案した．実験により，1 属性を含む表現で指示した場合よりも，複数属性を含む表現を用いた方が物体特定精度が高くなることが確かめられた．

本研究ではロボットに学習させていない表現による指示には対応できていない．今後は，未学習の指示においても対応可能な手法を研究する必要がある．また，本研究では，設定したタスクに対する提案手法の有効性を示すためにデータセットを構築した．構築したデータセットでは，人が実物体を見て，発話指示を行っている．人が物体画像を見て発話指示する場合との比較も行う必要がある．色名称や物体名称に基づいて物体認識を行う際，最も適した画像特徴量や，認識手法について研究することも今後必要な課題のひとつ

表 3.6　Examples of the result of object identification(%).

Scene index	Scene	Name	Object 1	Object 2	Object 3
1		Color name	**100.0**	95.0	–
		Object name	93.3	65.0	93.3
		Both names	**100.0**	**100.0**	**100.0**
2		Color name	**100.0**	–	–
		Object name	**100.0**	**96.0**	90.0
		Both names	**100.0**	75.0	**100.0**
3		Color name	**100.0**	–	0.0
		Object name	90.0	88.0	**93.0**
		Both names	50.0	**100.0**	83.3
4		Color name	0.0	–	–
		Object name	**65.7**	**80.0**	78.0
		Both names	–	70.0	**100.0**
5		Color name	0	**100.0**	–
		Object name	**70.0**	94.0	98.0
		Both names	–	98.5	**100.0**

である．本研究では，属性として色名称と物体名称のみしか扱っていないが，提案手法の枠組みは，他の属性にも拡張可能である．大きさや，テクスチャ，位置関係などの属性を今後与えていく必要がある．発話指示のフォーマットや，物体に遮蔽がないという制約も緩和し，さらに自然な発話の中でロボットが物体を特定できるよう，本研究を発展させていきたい．

3.2　一般物体認識に基づく物体特定

前節では，特定物体認識に基づく物体特定を行っていた．本節では，一般物体認識に基づく物体特定手法を提案する．

3.2.1　あらまし

特定物体認識とは，インスタンスレベルでの物体認識のことである．ある特定の物体と同じ物体を特定し，認識結果とする．特定物体認識を行うためには，全ての物体を予め学習させておくことが必要となる．例えば，家庭環境内において特定物体認識を行うためには，家庭環境内にある全ての物体を予め撮影し，物体ごとにモデルを構築しておく必要がある．しかし，予め全ての物体を撮影しておくことは困難である．

本節では，一般物体認識に基づいた物体特定法を提案する．一般物体認識は物体カテゴリに基づいて認識を行う方法である [52]．物体のカテゴリについてあらかじめ学習させておけば，特定物体の物体画像をあらかじめ学習させていなかったとしても，物体特定が可能となり，汎用性が高くなる．一般物体認識手法は数多く提案されているが [53, 54, 55]，本研究では，現在 state of the art であるディープニューラルネットワークを一般物体認識に用いる．

また，物体特定に用いる候補の絞り込み手法も併せて提案する．候補の絞り込みを行うことで，物体特定精度を向上させることができる．

3.2.2　タスクの設定

前節では，複数属性を用いた指示にも対応した物体特定タスクを扱っていたが，本節では，物体名称のみを用いた指示における物体特定タスクを扱う．しかし，本手法は前節同様，複数属性に基づく物体特定手法に拡張可能である，この拡張は今後の予定とする．

以下に本研究で扱うタスクの設定を示す．

- ロボットの目の前に複数の物体がある．
- 人がロボットに，「⟨ 物体カテゴリ名 ⟩ 取って」と指示する．
- 指示された物体は必ず目の前にあり，同じ物体名で指し示すことができる物体は他に存在しない．
- 物体特定とマルチモーダル物体認識の両方が成功した場合に，タスク達成とする．ここで物体特定は，指示された物体が目の前の物体のどれなのかを推定することであり，マルチモーダル物体認識とは，ロボットがあらかじめ保持している物体モデル (物体画像モデルと物体カテゴリ名の音声モデル) のどれに一致するのかを推定することである．
- 指示された物体は既知である，つまり画像モデルを保持している．

このタスクでは，指示された物体を特定できたかどうか，さらに特定された物体の認識結

果 (音声認識結果と物体認識結果) が正しいかどうかの 2 つの評価を含んでいる．

3.2.3　一般物体認識に基づく物体特定

特定物体認識に基づく物体特定手法と，一般物体認識に基づく物体特定手法との違いは 2 点ある．1 つ目は，前者が特定物体認識を用いたのに対し，後者は一般物体認識を用いる点である．2 つ目は，前者は，全音声認識結果候補を用いて物体特定を行っていたのに対し，後者は音声認識結果候補の絞り込みを行い，絞り込んだ結果を用いて物体特定を行う点である．物体特定には，音声認識結果と物体認識結果の両方が用いられる．音声認識が間違いやすいモデルと，物体認識が間違いやすいモデルは異なる．例えば，音声認識は，音が似ている物体名称と誤りやすいが，物体認識は見えが似ている物体名称と間違いやすい．本研究では，両認識結果の違いを考慮し，候補の絞り込みを行うことで物体特定の精度向上をはかる．

一般物体認識に基づく物体特定手法の流れは以下のようになる．

1. 発話 s を音声認識する．
2. 目の前にある複数の物体 $v_n(n = 1, 2, \cdots, N)$ を画像認識する．
3. 音声認識の結果を用いて，候補の絞り込みを行う．
4. 両認識結果を統合して，物体特定を行い，認識結果とともに出力する．

3.2.3.1　一般物体認識

入力された物体画像と全く同じ物体の画像を見つけ出す特定物体認識に対して，入力された物体画像がどの物体カテゴリに属するのかを答えるのが一般物体認識である．一般物体認識においては，カテゴリに共通の性質を学習しておくため，認識すべき物体全てを事前に学習する必要がない．一般物体認識では，従来から，人手により特徴量や識別器が設計されてきた．代表的な例として，画像から SIFT [56] 特徴量を抽出した後，LLC [53] を用いてそれらを量子化し，得られた特徴ベクトルを用いて，SVM [57] によって画像認識を行う手法が存在する．しかし，近年飛躍的な進歩を遂げたディープニューラルネットワークを用いることで，特徴量や識別器の設計を全て自動的にネットワーク重みとして学習でき，カテゴリに共通した情報を抽出することが可能となってきた．本研究では，ディープニューラルネットワークの一種である Convolutional Neural Network (CNN) [55] によって一般物体認識を行う．

m 番目の物体カテゴリの物体画像モデル o_m に対して，入力画像 v の信頼度 $C_v(v; o_m)$

は，以下の式によって得られる．

$$C_v(v; o_m) = \frac{y^m(v)}{\max\limits_{u_b \neq m} y^{u_b}(v)}. \tag{3.10}$$

ここで，$y^m(v)$ は，v を入力としたときの CNN の出力値のうち，m 番目の物体カテゴリに対応する値，u_b は m 番目以外で最も信頼度の高い物体番号を示す．画像認識出力の最大値で割ることにより，信頼度としての正規化を行っている．画像認識の結果 $\hat{m}$ は以下の式によって得られる．

$$\hat{m} = \underset{m}{\operatorname{argmax}}\, C_v(v; o_m). \tag{3.11}$$

3.2.3.2　一般物体認識に基づく物体特定手法

音声認識と一般物体認識を用いて，マルチモーダル物体認識を行う．本研究におけるマルチモーダル物体認識とは，対応関係にある入力音声 s と入力画像 v が，どの物体カテゴリモデル (音声モデルと物体画像モデルの組) に対応するかを認識することである．「音声と画像が対応関係にある」とは，例えば，音声が「カップ」で，画像も「カップ」である場合のことを指す．マルチモーダル物体認識には，音声認識結果と前項で得られた一般物体認識結果の信頼度を統合した値を用いる．ここで，音声認識結果から得られた信頼度と一般物体認識結果から得られた信頼度の組を $\mathbf{C}^{\mathrm{T}} = (1, C_s(s; \Lambda_m), C_v(v; o_m))$ とする．両信頼度の統合方法には，3.1 節同様に以下のロジスティック回帰を用いる [58].

$$F(\mathbf{C}) = \frac{1}{1 + \exp\{-\alpha^{\mathrm{T}}\mathbf{C}\}} \tag{3.12}$$

音声 s と画像 v のマルチモーダル物体認識結果 $\hat{m}$ は，以下の式によって得られる．

$$\hat{m} = \underset{m \in A}{\operatorname{argmax}}\, F(\mathbf{C}). \tag{3.13}$$

物体特定では，1 つの入力音声 s に対して，複数の物体画像 $v_n (n = 1, 2, \cdots, N)$ が入力される．音声で指示された物体を特定することは，音声と各物体画像を組み合わせた複数の組の中から，対応関係にある組を見つける問題と捉えることができる．$\{s, v_1\}$ が対応関係にある組とすると，他の組 $\{s, v_2\}$，$\cdots$，$\{s, v_N\}$ は対応関係にない組となる．音声 s と画像 $v_n (n = 1, 2, \cdots, N)$ の組 $\{s, v_n\}$ に対するマルチモーダル物体認識結果 $\hat{m_n}$ は式 (3.13) と同様にして得られ，それらの結果を用いて以下の式により物体特定結果が得られる．すなわち，全モデル，全組に対して，ロジスティック回帰の値が最も高くなる組を求めることで，物体特定が達成される．

$$\hat{n} = \underset{n}{\operatorname{argmax}}\, \hat{m_n}. \tag{3.14}$$

3.2.4　物体特定実験

一般物体認識に基づく物体特定手法の評価を行う．

3.2.4.1　実験条件

音声認識，画像認識:　Julius [42] を用いて不特定話者孤立単語音声認識を行った．単語辞書は，ImageNet Large Scale Visual Recognition Challenge 2013 (ILSVRC2013) [59] の Classification and Localization (CLS-LOC) データセットで用いられる 1000 カテゴリの中から，artifact ノードより下層にあるカテゴリ 514 個のカテゴリ名を用いて，物体名のみで構成した．同じカテゴリに属する物体でも，様々な色や種類，見え方の画像が含まれている．ある音声の認識につき，514 カテゴリに対する信頼度が算出される．物体特定のテストデータには，その 514 カテゴリの中から，家庭内に存在しうる 100 カテゴリを用いた．音響モデルには，Julius ディクテーションキット [42] に含まれている，不特定話者の PTM トライフォン Hidden Markov Model (HMM) を用いた．特徴ベクトルとして，メル周波数ケプストラム係数 (MFCC)12 次元とそのデルタ，対数パワーの合計 25 次元のベクトルを用いた．話者数は 2 名であり，ノイズがほとんどない環境下で 100 カテゴリ名を各 1 回発話し，それらを録音した．2 名のうち，1 名の発話音声をテストデータ，もう 1 名の発話音声をロジスティック回帰の学習データとした．

OverFeat [55] を用いて一般物体認識を行った．画像モデルは CNN で，ILSVRC2013 の CLS-LOC データセット中の学習データセットに含まれる 1,281,167 枚の画像を用いて学習されている．画像認識の際も音声認識と同様に，ある画像の認識につき，514 カテゴリに対する信頼度が算出される．画像は，ILSVRC2013 の CLS-LOC データセットのバリデーションデータセットに含まれる画像を用いた．各カテゴリにつき，5 枚をテストデータ，25 枚をロジスティック回帰の学習データとして使用した．テストデータは全て，画像モデルの学習データには含まれていない．

物体特定データセットの作成手順:　まず，音声で指示する物体カテゴリを 1 つ決め，その音声発話をテストデータとして用いる．次に，画像の組み合わせを決定する．例えば，画像が 3 枚入力され，その中から物体特定を行う際は，3 枚のうち 1 枚を音声と対応する物体カテゴリ，残りの 2 枚を音声と対応しない物体カテゴリとする．音声と対応しない物体カテゴリは，514 カテゴリから音声と対応する物体カテゴリを 1 つ取り除き，513 カテゴリの中から特定する．2 枚に対する物体カテゴリの組み合わせは，${}_{513}P_2$ 組の中からランダムに 1 組を選ぶ．物体カテゴリの組み合わせが決定すれば，それぞれの物体カテゴリに対して，画像を 1 枚使用する．こうして得られた 1 音声と 3 画像を，物体特定実験の 1 テストデータとして用いる．各物体カテゴリに対して，テストデータとして使用可能な 5

枚の画像があるので，1 音声発話に対して，物体特定実験のテストデータが 5 個作成できる．以上の手順を，指示音声を変えて 100 回繰り返し，物体特定データセットを作成した．すなわち，合計で $5 \times 100 = 500$ 個の画像の組が作成できる．ただし，特定された 3 画像の中に，1 枚でもグレースケール画像が含まれる場合は，OverFeat による画像認識が不可能であるため，テストデータから除いた．そのような場合は 40 回あったため，合計で $500 - 40 = 460$ 回の物体特定実験を行った．

ロジスティック回帰の学習:　ロジスティック回帰の学習のために必要となる，Case 1∼Case 5 の 5 クラスに対する学習データの作成法について述べる．ある 1 音声と 1 画像の組が与えられたとき，両者が対応関係にあるのかそうでないのかの二種類に分けられる．音声と画像の組について，音声認識・画像認識を行うと，それぞれ 514 個の信頼度が得られる．この 514 個の音声信頼度と画像信頼度は，各モデルに対して組となって紐づけられている．i) 対応関係にある組が与えられたとき，音声信頼度と画像信頼度の組 514 組のうち，1 組のみが Case 1，それ以外の 513 組が Case 2 となる．ただし，数が膨大となるため，513 組の中からランダムに 1 組を選んだ．ii) 対応関係にない組が与えられたとき，音声信頼度と画像信頼度の組 514 個のうち，1 組が Case 3，1 組が Case 4，それ以外の 512 組が Case 5 となる．ただし，その 512 組の中からランダムに 1 組を選んだ．

このような作業を，以下に示す回数だけ，繰り返し行うことで，学習データを作成した．i) 対応関係にある音声の物体カテゴリと画像の物体カテゴリの組は 100 通りである．これはテストデータが 100 カテゴリあることに起因する．音声の物体カテゴリと画像の物体カテゴリの組が決まれば，その物体カテゴリに対して音声は 1 音声，画像は 25 画像を使用できるので，$1 \times 25 = 25$ 通りの音声と画像の組が作成される．これは，1 物体カテゴリ当たりの学習用物体画像枚数が 25 枚であることに起因する．よって，作業の繰り返し回数は，$100 \times 25 = 2500$ 回となる．ii) 対応関係にない音声の物体カテゴリと画像の物体カテゴリの組は 200 通りである．その中からランダムに 100 組を選んだ．音声の物体カテゴリと画像の物体カテゴリの組が決まれば，i) と同様にして，音声は 1 音声，画像は 25 画像を使用できるので，$1 \times 25 = 25$ 通りの音声と画像の組が作成され，作業の繰り返し回数は，$100 \times 25 = 2500$ 回となる．

3.2.4.2　物体特定実験

提案手法である一般物体認識に基づく物体特定法を評価した．音声認識精度は 93%，画像認識精度は 63.2% であった．物体特定の結果を図 3.10 に示す．入力画像枚数が 1 枚の場合は，物体が正しく認識された場合にタスクが達成されたとしている．入力画像枚数が 2，3 枚の場合は，音声によって指示された物体を特定でき，かつ，特定された物体が正しく認識された場合にタスクが達成されたとしている．“no integration”は，音声認識を行い，音声による候補の絞り込みは行うが，音声認識信頼度と統合することはせず，入力画

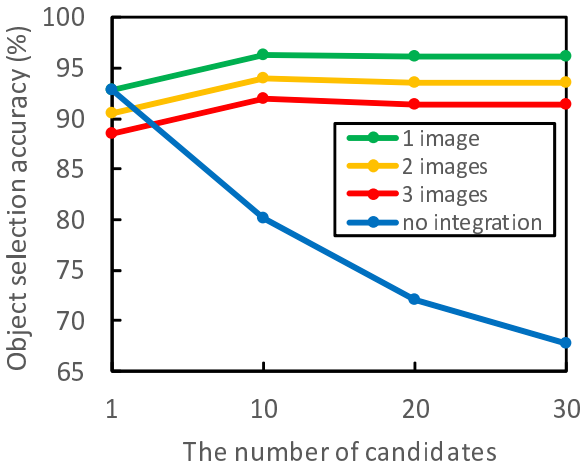


図 3.10　Accuracy of object identification.

像 1 枚に対して，画像認識信頼度のみを用いて物体認識を行った際の結果である．すなわち，音声認識により候補を絞り込み，この候補を用いて入力画像を画像モデルで認識し，画像認識信頼度の最も高いものを認識結果とした結果である．

物体特定精度は，入力画像枚数が 3 枚のときでも約 92% (候補数 10) と高い精度となり，提案手法の有効性が確かめられた．また，候補数が増えると “no integration” は下がるが，音声信頼度と統合することによって，“1 image” は約 96% と高くなる．これは，音声認識精度が 93% と高いからだと考えられる．さらに，例えば入力画像枚数が 1 枚のとき，候補数が 1 の場合と比べて，$10, 20, 30$ の場合の物体特定精度は約 3 ポイント高い．これは，候補数を減らしすぎると，正解がその候補内に含まれていない可能性が高くなるためだと考えれる．入力画像枚数が $2, 3$ 枚のときも同様の傾向が見られる．

3.2.5　まとめと今後の課題

本節では，一般物体認識に基づく物体特定手法を提案した．一般物体認識の手法として CNN を用いた．また，物体特定を行う際，音声認識結果候補の絞り込みを行った．候補を絞り込む際，絞り込む候補数は，実験的に決定したが，絞り込む候補数の自動設定が今後必要となる．また，絞り込みの際，音声認識精度が物体認識精度よりも高かったため，音声認識結果を用いたが，逆である場合，物体認識結果を用いた絞り込みが有効であると考えられる．つまり，音声認識精度，画像認識精度を考慮し，自動的に絞り込み手法を選択し，さらに絞り込みの候補数も自動的に設定するような手法の開発が望ましい．また，本節で提案した一般物体認識に基づく物体特定手法も，特定物体認識に基づく物体特定手法と同様，複数属性に拡張することが望ましい．

第 4 章

色属性による物体特定のための顕著性

人の認識とロボットの認識は異なる．この理由のひとつとして，人とロボットの視覚属性が異なることが考えられる．本論文では，人の視覚属性を考慮した物体特定タスクのための顕著性検出手法を提案する．

4.1 あらまし

物体名称による物体特定を扱う研究は多数提案されている [18, 19, 4, 6] が，色や大きさ，位置などの属性にも対応した研究はまだ少ない．同じ物体名称の物体があるとき，色や材質，大きさなどの属性を用いて指示をするのが自然である．複数の属性を扱う枠組みの提案はされている [16, 8, 9, 10] が，コンピュータビジョンにおける各属性の扱いについて十分に議論を行っている研究は少ない．例えば，色属性は物体画像領域の RGB [60] 値や Lab [51] 値の平均値を特徴として用いている研究が多い [16, 8]．明らかに単色で構成された物体ばかりであれば，問題ないが，実際は，複数の色で構成された物体の数の方が多いことは明らかである（図 4.1）．複数の色で構成された物体であったとしても，人が色で物体の指示を行う場合，人は 1 色で指示を行うのが自然である．「ピンクで黄色のものとって」という言い方よりも，「ピンクのものとって」という言い方の方が多いだろう．ロボットがコンピュータビジョンを用いて，複数の色で構成された物体の色を 1 色で認識することは困難である．

ロボットの視覚属性と人の視覚属性は異なる．ロボットが物体領域全域を見て色を見るのに対し，人は物体で最も顕著な部分の色を見て，その物体が何色であるか判断する．人



図 4.1　Examples of various-colored objects.

の視覚属性を考慮したコンピュータビジョンによる色名称モデルが必要とされる．

画像研究において，人の視覚属性を考慮したモデルのひとつとして [61] が提案した顕著性が挙げられる．顕著性とは人が画像中でどの部分を注視しやすいかを表したものである．画像から各ピクセルの顕著性を算出したものを顕著性マップと呼ぶ．物体画像に有効な顕著性検出手法のひとつとして，[20] のマニフォールドランキングを用いたグラフベースの顕著性検出手法がある．マニフォールドランキングとは，順位付けしたいデータで多様体を構築し，多様体の構造を考慮することで，データのランキングを行う手法である [62]．物体画像の前景と背景の色相関を用いて，マニフォールドランキングを行い，物体領域の顕著性を検出することに成功している．

本研究の目的は，人の視覚属性を考慮することで，複数の色で構成された物体であったとしても，人がその物体を見て思う 1 色で，ロボットに物体を特定させることである．本論文では，色による物体特定のための顕著性検出手法を提案する．顕著性検出には人に指示された色との相関に基づいたマニフォールドランキングを用いた．マニフォールドランキングを用いて [20] と同様に，グラフベースで顕著性マップを構築した．実験では提案手法で検出した顕著性を示すと共に，提案手法で構築した顕著性マップを用いた物体特定を行うことで，提案手法の有効性を示す．

4.2　関連研究

[8, 16] は，色による物体特定に関する手法を提案している．[8] は，物体領域の Lab の平均値を特徴とし，正規分布でモデル化することで色による物体認識を行い，物体特定に用いている．この手法では，物体画像において支配的な色を特徴量として抽出していることになる．[8] は，ほぼ単色で構成された物体のみを扱っているためこの方法で十分であるが，複数色で構成された物体を扱う場合には適さないことは明らかである．[16] は，RGB 画像とデプス画像のパッチを学習データとして，階層型スパースコーディングによって特徴を学習し，得られた特徴をヒストグラム化することで色モデルを構築してい

る．[16] は，物体画像ごとに色モデルを構築している．この手法は特定物体認識ベースで色を扱っていると言える．特定物体認識とは，インスタンスレベルでの物体認識のことである．色は本来，一般物体認識ベースで扱われるべきである．一般物体認識とはカテゴリレベルでの認識のことである．色は物体のインスタンスレベルで学習されるものではなく，色のカテゴリごとに学習されるべきである．勿論，物体固有の色はあるかもしれないが，人が物体を指示する際は，物体固有の色のみを考慮するのではなく，人の色概念に基づいて指示すると考えられるからである．本研究では，色を色のカテゴリごとに学習し，一般物体認識ベースで物体特定を行う．

人が物体を色で指示する際，人は視覚属性に基づいて物体の色を決定していると考えられる．物体において顕著な色をその物体の色として，色の指示に用いる．本研究では，人の視覚属性を考慮する．[8, 16] は，物体の画像属性は考慮して色を扱っていたが，人の視覚属性は十分に考慮していない．

[61] は人の視覚属性を考慮したモデルである視覚顕著性を提案している．人の視覚には，人の注意を惹く場所に視覚が反応する仕組みがある．この仕組みを画像によりモデル化したものが顕著性となる．近年，[61] の提案した顕著性を基に，様々な顕著性に関する研究が提案されている．顕著性の研究は大きく 2 つに分けられる．1 つ目は，画像に基づいた顕著性を検出する研究で，2 つ目は，タスクに基づいた顕著性を検出する研究である．画像に基づいた顕著性を検出する研究の例として，[20, 63] が挙げられる．[63] はグラフベースの顕著性検出手法を提案している．スーパーピクセルにより分割された画像からグラフを構築し，得られたグラフからアフィン行列を求め，伝播させることでグラフノードの重みを最適化している．[20] もまたグラフベースの顕著性検出手法を提案している．[20] もまた画像をスーパーピクセルに分割し，画像の周囲に張り付くスーパーピクセル（背景であるとする）と，それ以外のスーパーピクセル（前景であるとする）との色相関を用いて，グラフベースのマニフォールドランキングを行うことで，前景にあるはずの物体領域の顕著性を検出している．タスクに基づいた顕著性検出手法として [64] が例として挙げられる．[64] は，一度特定した人物を再度特定する際に服の色が有効であると考え，服の色を表現する顕著性検出手法を提案している．特徴量には RGB 値を用いて，人ごとに色特徴記述子を算出し，色特徴記述子を用いて顕著性の検出を行っている．本研究で扱う顕著性は，タスクに基づいた顕著性である．ただし，[64] とは扱うタスクが異なる．本研究では色による物体特定タスクを扱う．

4.3 問題設定

本研究では，指示対象物体の色を用いなければ特定不可能な場合に，色を用いて物体を特定するタスクを扱う．例えば，特定対象と，物体名称が同じである物体が他にも存在す



(a) “Bring me a yellow one.”



(b) “Bring me a purple one.”

図 4.2　Examples of object identification task using color name.

る場合，物体名称での指示は不可能である．

本研究では，人に色で指示された物体を，ロボットが複数物体の中から特定するタスクを扱う．このとき，物体名称では物体を特定不可能な場合のみを扱う．具体的には，指示対象物体と，物体名称（ペットボトル，お菓子など）が同じ物体が他にもある状況で，指示対象物体を色のみで特定するというタスクを扱う．色での指示は，1 色のみで行う．たとえ，指示対象物体が複数の色で構成された物体であったとしても，1 色のみで物体を指示するとする．以降このタスクのことを色による物体特定タスクと呼ぶ．具体的なタスク例を図 4.2 に示す．例えば，図 4.2(a) は，お菓子が 1 つとペットボトルが 2 本ある場合である．あるペットボトルを指示したいが，「ペットボトル」という情報だけでは，「ペットボトル」が複数存在するため特定不可能である．そこで，色を使ってロボットに指示をする．

本研究では，人が実物体を見て，色のアノテーションを行い，色名で物体を指示する．色のアノテーションには人の視覚属性が影響すると考え，物体特定には，人の視覚属性を考慮して物体画像から検出した顕著性を用いる．[16] と同様に，人はテキストを用いてロボットに「赤いのをとって」と指示を行うとする．音声認識を行い，物体特定タスクを達成することが望ましい [18, 19] が，これは今後の予定とする．本研究では，音声認識が 100% の精度でなされ，発話がテキスト化されロボットに入力されたと仮定する．よって，人が指示した色のテキストのみがロボットに入力されるとする．

特定対象候補となる物体のバウンディングボックスは得られているという前提を置く．バウンディングボックスとは，物体を囲む矩形のことである．物体のバウンディングボックスの中に必ず指示対象となる物体が含まれているとする．ロボットは，特定対象候補となる物体のバウンディングボックス内の画像を用いる．また，物体とカメラの間には遮蔽はないとする．物体の遮蔽については今後の予定とする．

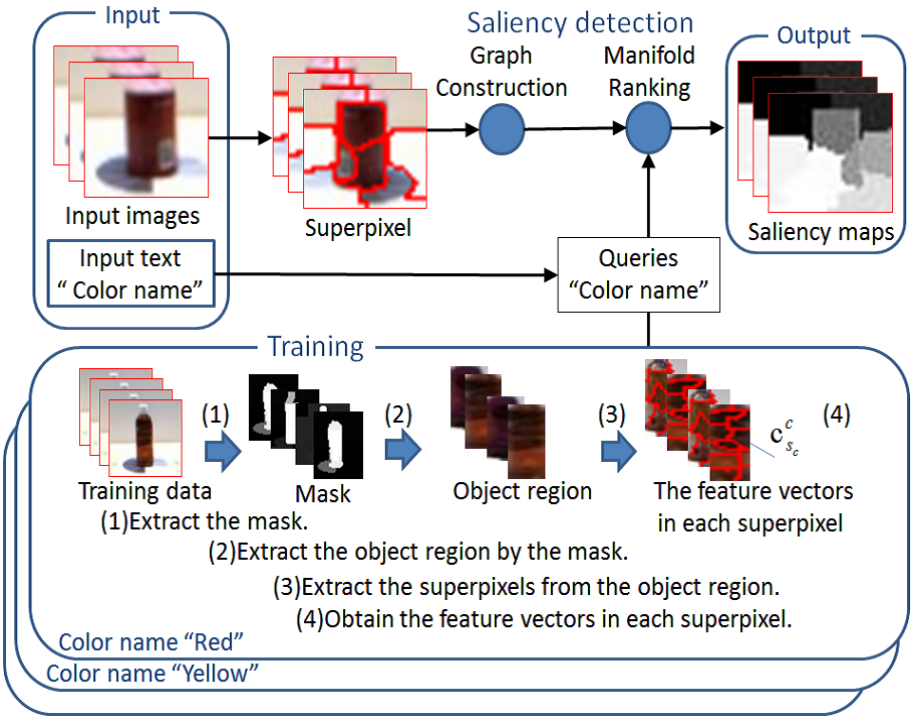


図 4.3　Overview of the saliency detection of color names for object identification.

4.4　物体特定のための色顕著性

顕著性の検出は [20] の手法を参考にした．[20] は，前景と背景の色の相関度を基に，マニフォールドランキングを行うグラフベースの顕著性検出手法を提案している．本研究では前景と背景の相関度の代わりに，テストデータ（画像）と学習データ（画像）の色相関度を用いてグラフベースで顕著性を検出する．

色の相関度は物体画像のスーパーピクセルごとに算出する．まず学習データの各画像ごとに物体画像のスーパーピクセルを算出する．各スーパーピクセル内の RGB [60] の平均値を算出する．これらと，テストデータの各スーパーピクセルの RGB の平均値との相関度を基にマニフォールドランキングを行い，顕著性の検出に用いる．相関度は，マハラノビス距離 [65] と部分空間における投影距離 [66] によって求める．顕著性検出の流れを図 4.3 に示す．

4.4.1　色名称との相関度

4.4.1.1　色名称の学習

人に物体の色名称のアノテーションをしてもらい，色名称ごとに物体画像の学習データを構築する．人のアノテーションは，照明変化が起こったとしても，色恒常性により変化しないが，物体画像は照明変化の影響を受けやすい．そこで，本研究では，物体を様々な角度から撮影した画像を用いる．

まず，物体のデプス画像を用いて，物体領域のバウンディングボックスを抽出する．バウンディングボックス内の物体画像のみを用いる．次に，SLIC [67] を用いて，物体画像からスーパーピクセルを抽出する．各スーパーピクセルの RGB の平均値を，各色名称ごとに算出し，学習データ $\mathbf{C}^c = \{\mathbf{c}^c_1, \ldots, \mathbf{c}^c_{s_c}\}$ とする．ここで，c は色名称のインデックスを，s_c は c 番目の色名称のスーパーピクセルの数を，$\mathbf{c}^c_{s_c}$ を s_c 番目のスーパーピクセルの RGB の平均値とする．c 番目の色名称との相関は，$\mathbf{C}^c$ を用いて算出する．

4.4.1.2　マハラノビス距離を用いた相関度

テストデータも学習データと同様に，スーパーピクセルで領域分割を行い，スーパーピクセルごとに RGB の平均値を求める．i 番目のテストデータの j 番目のスーパーピクセルの RGB の平均値を $\mathbf{t}^i_j$ とする．$\mathbf{t}^i_j$ と $\mathbf{C}^c$ 間のマハラノビス距離を，$\mathbf{t}^i_j$ の c 番目の色名称との相関度とする．このマハラノビス距離を $d^i_j(c)$ とすると，

$$d^i_j(c) = (\mathbf{t}^i_j - \mu_c)^{\mathbf{T}} \sum\nolimits_c^{-1} (\mathbf{t}^i_j - \mu_c). \tag{4.1}$$

ここで，μ_c は $\mathbf{C}^c$ の平均ベクトルを，$\sum_c$ は，$\mathbf{C}^c$ の共分散行列とする．$d^i_j(c)$ が，長いほど相関は低くなり，短いほど高くなる．

4.4.1.3　投影距離を用いた相関度

マハラノビス距離は，正規分布をパターン分布とするベイズ識別の近似解である [68]．部分空間における投影距離もまた，正規分布をパターン分布とするベイズ識別の近似解である．マハラノビス距離から識別関数を導出する過程で，投影距離が得られる．投影距離もまた，マハラノビス距離と同様，相関度として用いることができる．

色名称ごとに部分空間を構築し，部分空間とテストデータの投影距離を色名称との相関度とする．相関度は各テストデータの各スーパーピクセルごとに求める．つまり，c 番目の色名称との相関は，$\mathbf{C}^c$ で構築した部分空間とテストデータの投影距離とする．

i 番目のテストデータの j 番目のスーパーピクセルと，色名称 c の部分空間における投

影距離は，以下のようにして求められる．

$$d_j^i(c) = \|\mathbf{t}_j^i - \mu_c\|^2 - \|\mathbf{U_c}^{\mathrm{T}}(\mathbf{t}_j^i - \mu_c)\|^2, \tag{4.2}$$

ここで，$\mathbf{U_c}$ は色名称 c の部分空間の投影行列である．投影距離が長くなると，相関は低くなり，短くなると，相関は高くなる．投影距離 [66] は，CLAFIC 法でも算出可能である．CLAFIC 法を用いると，相関度は，以下のようにして求められる．

$$d_j^i(c) = \|\mathbf{t}_j^i\|^2 - \|\mathbf{U_c}^{\mathrm{T}}\mathbf{t}_j^i\|^2. \tag{4.3}$$

投影距離法と CLAFIC 法の両者の違いは，部分空間の原点のとり方である．CLAFIC 法では，各色名称における部分空間において原点は共通であるが，投影距離法は各色名称ごとに異なる．投影距離法における各部分空間の原点は，各部分空間の平均によって決定される [69].

部分空間を構築する際，特徴ベクトルの次元は高い方が望ましい．各スーパーピクセルにおける RGB の平均値を特徴ベクトルとして用いていたが，各スーパーピクセルにおける色ヒストグラムを算出して用いることが有効である．具体的な算出方法は以下のようになる．各スーパーピクセルから K 個の RGB 値をランダムに選択し，その値を用いて色ヒストグラムを求める．R，G，B ごとにヒストグラムを算出し，それらを $L2$ ノルムで正則化した後に結合し，特徴ベクトルを求める．つまり，ビン数 8 (等分割) でヒストグラムを算出する際，特徴ベクトルは 24 次元となる．

4.4.2 色顕著性を用いた物体特定

提案手法では，色顕著性を用いて物体特定を行う．人が色名称により物体を指示すると，ロボットは目の前にある複数の物体の色顕著性を検出し，色顕著性が最も高い物体を特定対象であると判断する．色顕著性を検出する際は，指示された色名称における顕著性を検出する．例えば，「赤いものをとって」と指示された場合，色名称「赤」の学習データを用いて特定対象候補物体の画像（入力画像）との相関度を算出し，顕著性を検出する．検出した顕著性を 0 から 1 の値に正規化し，顕著性マップを算出し，顕著性マップを用いて物体の特定を行う．顕著性マップ算出の流れと，算出された顕著性マップの例を図 4.3 示す．顕著性マップとは，顕著性が高い部分が白くなり，低い部分が黒くなるような画像のことである [20].

具体的な物体特定手順は以下のようになる．

Step.1 各テストデータのスーパーピクセルを抽出する．
Step.2 各テストデータの全スーパーピクセルにおいて特徴ベクトル $\mathbf{t}_j^i$ を求める．

表 4.1　Examples of objects in dataset.

Color name	Red	($c=1$)	Orange	($c=2$)	Yellow	($c=3$)	Green	($c=4$)	Blue	($c=5$)
Example										
Color name	Purple	($c=6$)	Pink	($c=7$)	White	($c=8$)	Black	($c=9$)	—	—
Example									—	—

Step.3　Step.2 で求めた特徴ベクトルを用いて，指示された色名称 c との相関度 $d_j^i(c)$ を各スーパーピクセルごとに求める．

Step.4　Step.3 で求めた相関度を用いて顕著性を検出し，[20] と同様の方法で顕著性マップを算出する．

Step.5　顕著性マップの画素値の平均値が最も高かった物体を特定結果とする．

4.5　実験

4.5.1　データセット

人は物体画像を見るのではなく，実物体を見て，物体の色を判断し，ロボットに指示する．ロボットは，人が見ていた実物体の画像を取得し，人の指示と取得した物体画像を用いて物体特定を行う．人が実物体を見て，色のアノテーションを行い，人が見た実物体の物体画像を撮影したデータセットは，存在しない．本研究では，タスク設定に合ったデータセットを構築した．

実験に用いた物体は 18 物体であり，うち 9 物体はペットボトル，残りはお菓子である．3 被験者に 18 物体の色のアノテーションをしてもらったところ，全員のアノテーションが一致した．色のアノテーション結果を基に，色インデックスを物体に付加した．色のインデックス数は 9 となり，実験に用いた 18 物体の内訳は，9 色，各 2 物体となった．実験に用いた物体の画像例と色インデックスを表 4.1 に示す．表 4.1 より，お菓子同士の形状（$c=3$ のお菓子を除く），ペットボトル同士の形状は全色で類似していることが分かった．また，物体画像を見ただけでは直感的に色インデックスが分かりにくいものもあることが分かった．

各物体を 10 方向から撮影し，10 枚の画像を得た．撮影には，Point Grey 社 [48] の近距離撮影用複眼カメラ Bumblebee2（BB2-03S2）を使用した．物体画像を撮影する際，高さ 70cm の机の中心に物体を置き，物体からカメラまでの距離を 112cm，カメラの高さを 119cm とした．複眼カメラの視差より，デプス画像を求め，物体のバウンディング

表 4.2 Saliency maps of 1 object. Colors of each object is different each other.

Method	Input image	Red(c=1)	Orange(c=2)	Yellow(c=3)	Green(c=4)	Blue(c=5)	Purple(c=6)	Pink(c=7)	White(c=8)	Black(c=9)
Mahalanobis (RGB)										
Projection (RGB)										
CLAFIC (RGB)										
Projection (Histogram)										
CLAFIC (Histogram)										

ボックスを抽出した．カメラと撮影対象の物体間に遮蔽は存在しないと仮定したため，一番手前の奥行きを基にバウンディングボックスを抽出した．

4.5.2 色顕著性の考察

各色名における顕著性がどのようなものであるかを考察するため，物体画像 1 枚につき，9 色の顕著性マップを出力した．学習データには，各物体につき 5 方向の画像を用い，テストデータには残りの 5 方向の画像を用いた．顕著性を検出する際の色相関は，マハラノビス距離を用いた手法，投影距離を用いた手法，CLAFIC 法を用いた手法，色ヒストグラムにより構築した投影距離，CLAFIC 法を用いた手法で算出した．顕著性マップは，各色に対して，算出した．

表 4.2 に，9 色各画像の顕著性マップを色ごとに示す．表にはペットボトルの顕著性マップを示した．今，ペットボトルは赤とアノテーションされているため，赤 $(c = 1)$ のときのみ，ペットボトル部分が白くなり，それ以外の色のときは黒くなることが望ましい．全ての顕著性マップにおいて，ペットボトル部分のみならず，背景部分が白く反応してしまっている．これは，学習データに背景部分が含まれてしまっていることが原因と考えられる．ペットボトル領域部分に着目すると，一番理想的であったのは，色ヒストグラムを特徴量とし，CLAFIC 法を用いた場合であった．この場合のみ，赤のときはペットボトル部分が白くなり，他の色のときはペットボトル部分が黒くなっていた．

4.5.3 色顕著性を用いた物体特定実験

3 物体画像の中からテキスト（色のインデックス）で指示された物体を特定し，物体特定率を示した．

学習データには各物体につき 5 方向の画像を用いた．つまり，5 画像 ×9 色 ×2 物体 = 90 画像を学習データに用いた．テストデータにはデータ集合構築に用いなかった残りの 90 画像を用いた．テストデータは，3 物体画像と 1 テキスト（色のインデックス）の

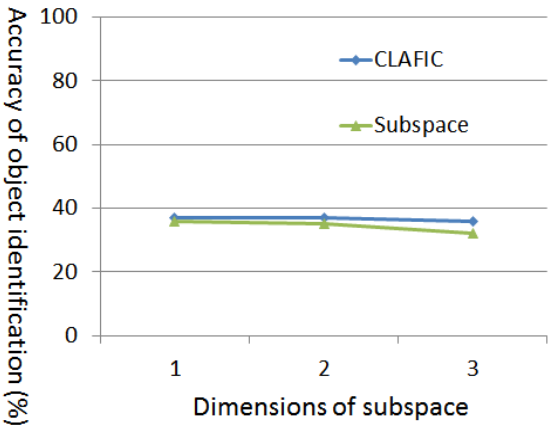


図 4.4　Relation between accuracy of the object identification and the dimension of subspace (RGB).

組となる．テストデータを選択する手順を説明する．今，特定対象が色のみで指示可能な場合を想定するため，特定対象と他の 2 物体の色のインデックスは異なる．また，物体名のみでは，特定不可能な場合を想定するため，特定対象と同じ物体名の物体が少なくとも 1 物体，特定対象以外の 2 物体に含まれているとする．以下に，テストデータの選択方法の手順を示す．特定対象となる物体画像の色のインデックスを i_0，物体の種類のインデックスを t_0，画像のインデックスを v_0 とする．残りの 2 物体の色インデックス，物体の種類のインデックス，画像のインデックスをそれぞれ，(i_1, t_1, v_1)，$(i_2; t_2, v_2)$ とする．

テストデータの選択手順

1. 特定対象とする物体画像を画像のテストデータから 1 画像選択し，(i_0, t_0, v_0) を決定する．
2. i_1，i_2 は，色インデックスの中から i_0 以外のインデックスをランダムに 1 つずつ選択し，決定する．t_1 は，t_0 と同じ物体インデックスにする．t_2 は，t_0 を除き，物体インデックスの中からランダムに 1 つを選択し，決定する．v_1，v_2 は，画像インデックスの中からランダムに 1 つずつ選択し，決定する．

以上の手順をランダムに 100 回繰り返し，100 回中，物体特定に成功した回数の割合を物体特定率とした．評価では，画像から抽出されるスーパーピクセル数の最大は 15 個とした．色ヒストグラムを抽出する際に，スーパーピクセルからサンプリングされる画素数は $K = 10$ とした．

図 4.4 は，特徴ベクトルに各スーパーピクセルにおける RGB の平均値を用いた場合の，部分空間（投影距離/CLAFIC 法）を用いた顕著性マップによる物体特定率と，部分空間の次元数の関係を示している．特徴ベクトルは 3 次元であるため，部分空間は高々 3 次元

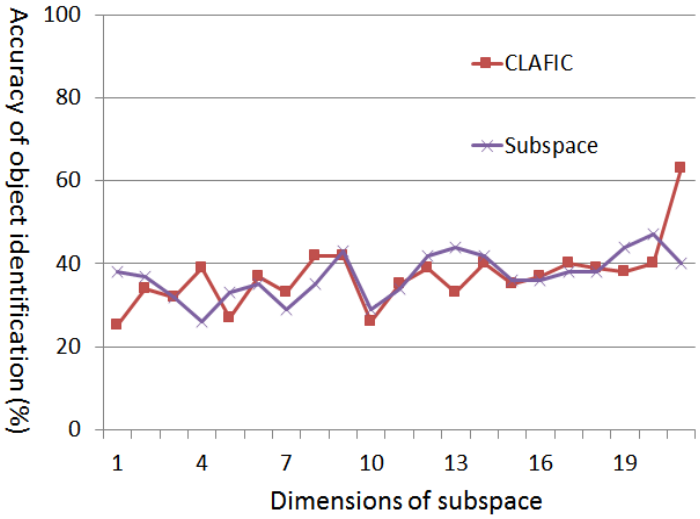


図 4.5 Relation between accuracy of the object identification and the dimension of subspace (color histogram).

表 4.3 Accuracy of the object identification (%).

Mahalanobis distance (RGB)	Projection distance		CLAFIC distance	
	RGB	Histogram	RGB	Histogram
54.0	36.0	47.0	37.0	63.0

である．「CLAFIC」は CLAFIC 法を用いた手法，「Subspace」は投影距離を用いた手法を示している．物体特定率は，次元数を変化させても，大きな変化が見られなかった．図 4.5 は，色ヒストグラムから得た特徴ベクトルを用いた場合の，部分空間を用いた顕著性マップによる物体特定率と，部分空間の次元数の関係を示している．色ヒストグラムのビン数は 8 とし，特徴ベクトルの次元数は 24 とした．部分空間の次元数は高々 24 であるが，本実験で用いたデータセットでは，最大で 21 となった．図 4.4 の結果と，図 4.5 の結果を比べると，色ヒストグラムを用いた手法の方が有効であることが分かる．

表 4.3 に各手法を用いた場合の物体特定率を示す．部分空間を用いた手法では，最も精度のよかった次元数での精度を表に記載した．表中の「RGB」は各スーパーピクセルにおける RGB 値の平均値を用いた手法を示している．最も精度がよかったのは，色ヒストグラムによる特徴ベクトルを用いて，CLAFIC 法により相関度を求めた手法であった．

4.6　まとめと今後の課題

本論文では，色による物体特定タスクのための，人の視覚属性を考慮した顕著性検出手法を提案した．提案手法で算出した顕著性を定性的に評価することで，提案手法は人の視覚属性に近い顕著性を検出できる可能性があることを確かめた．また，提案手法を用いて算出した顕著性マップを用いて物体特定タスクを行い，提案手法の定量的な評価を行った．提案手法により検出した顕著性は，物体の識別能力が高いとは言いがたいが，人の視覚属性をよく反映することができた．今後，提案手法を物体の識別能力向上も考慮した手法に拡張していきたい．また，本研究では物体特定に用いる属性を色のみに限定したが，この制約を外し，物体の材質やテクスチャなど様々な属性にも対応できるように拡張していきたい．人とロボットの視覚属性の違いを明らかにし，人がロボットとの違いを感じることなく，ロボットとの自然なインタラクションを楽しみながら共生していく未来を実現する手助けがしたいと考えている．

第 5 章

マルチモーダル情報を用いた未知物体識別

物体特定タスクを通じて，ロボットが未知物体を学習していくためには，まず，指示された物体が未知であるのか既知であるのかを判定する必要がある．本章では，音声情報と画像情報を用いることで，指示されている物体が未知であるのか，既知であるのかの判定を行う手法を提案する．音声認識結果と画像認識結果をロジスティック回帰で統合することにより，未知物体と既知物体の識別を行う．

5.1 あらまし

家庭用サービスロボットが家庭環境下で作業を行う際，ロボットは人の言葉を実世界にグラウンディングし，理解することが必要となる．言語を実世界にグラウンディングするためには，実世界にある物体を様々なモダリティによって表現する必要がある．Roy [4] は，複数のモダリティを用いた言語獲得のための計算モデルを提案している．モダリティには，視覚情報や聴覚情報を用いている．物体を介したコミュニケーションには，特に複数のモダリティが必要となる．例えば，言語情報や，人の声，物理的に観測された物体情報などが必要となる．言語や意味を学習するために，音声や物体画像などの複数のモダリティを用いた手法が数多く提案されている [4, 5, 6, 7, 8, 9, 10].

人は，本から知識を得るだけでなく，他者との会話やインタラクションを通じて知識を得る．ロボットも，人との相互的なインタラクションを通じて知識を得ていくことが望ましい．人とのインタラクションを通じて，ロボットに知識を学習させていく研究がいくつかある [11, 12, 13, 14, 15]．これらの研究には 2 つのアプローチがある．一つ目は，幼児



図 5.1　Autonomous discrimination of unknown objects and their names by a robot.

の学習を模したものである [11]．幼児が，親とのインタラクションを通じて，言語や概念を獲得していく様子を模したアプローチである．二つ目のアプローチは，タスクを通じて学習させていくアプローチである [12, 13, 14, 15]．タスクには，グラウンディング問題に関するタスクを設定している．例えば，人が声で指示した物体を特定して，運んでくるタスクを考える．このタスクを達成するためには，人が物体名を発話した音声が，物体画像にグラウンディングされている必要がある．本研究は，タスクを通じて学習していくアプローチをとる．本研究の目的は，ロボットが人との自然なインタラクションを通じて，知識を学習していくことである．インタラクションとして，物体特定タスクを扱う．タスクでは，ロボットの目の前にある複数の物体 (未知物体も含まれる) があり，人がロボットに「⟨ 物体名 ⟩ とって」と発話する状況を設定する (図 5.1)．本研究と似たような物体特定タスクを扱っている研究はいくつかあるが，それらの研究はロボットの目の前にある物体は全て既知であるという前提を置いていた．物体が既知であるとは限らず，未知物体もありうるという前提をおいた研究は存在しない．

人は，しばしば相手の知らない物体を指示するだろう．このとき，指示された人は，自分がその物体を知っているのか，知らないのかの判断を行う．物体を知らない場合は，そのときに物体を学習するだろう．本研究では，物体名称も物体画像モデルも聞き手が知らない物体を未知物体と定義する．未知物体は聞き手にとってアウトオブボキャブラリの物体となる．

本研究では，未知物体を学習するための第一歩として，音声認識結果と画像認識結果の信頼度を統合したマルチモーダル情報を用いて未知物体を識別する手法を提案する．マルチモーダル情報を用いた物体認識を拡張することによって，本手法は未知物体識別を可能とする．物体特定タスクの際，まず指示された物体が未知であるのか，既知であるのかの判断を行う．物体が既知である場合は，物体を特定する．物体が未知である場合は，物体を学習する．学習は今後の予定とするが，例えば，物体の音声を収録し，物体の画像を撮影するという単純な学習方法もある．

タスクを達成するためには，音声認識と物体認識が必要となる．入力される音声と物体



(a) Bring the object and say “Here you are.”.



(b) Say “I don’t know.”.

図 5.2　Variation of robot behaviors.

画像の組み合わせは，4 種類ある．既知音声と既知画像，既知音声と未知画像，未知音声と既知画像，未知音声と未知画像である．ロボットはこの 4 種類を識別する必要がある．ロボットがこの 4 種類を識別することができれば，ロボットがタスクを達成することができる．また，物体が未知であったとしても，物体特定可能であるケースもある．

5.2　タスクの設定

本研究で扱うタスクは，ロボットが，人に声で指示された物体を，複数の未知もしくは既知の物体の中から選択するというタスクである．このタスクは，多くのロボット研究者の扱っている物体把持タスクとは異なる．本研究で設定したタスクは家庭環境内で作業するロボットにとって重要であるにもかかわらず，私の知る限りこのタスクを扱っている研究は存在しない．

本研究で設定するタスクの詳細は以下のようになる．

- 複数の物体が机の上にある．これらの物体は全て既知であるとは限らない．未知の物体が含まれることもある．
- 人はロボットに「机の上にある〈物体名〉をとって」と発話指示し，ロボットは返答を行う．

2 種類の返答をロボットに用意する．返答を図 5.2 に示す．ロボットの返答は以下の 2 種類である．

1. ロボットが物体を特定できた場合，物体を把持し，「はい，どうぞ」と返答する．
2. ロボットが物体を特定できなかった場合，「分かりません」と返答する．

物体が既知であれば，ロボットは物体を特定することができる．物体が未知であれば，物体を特定することができない．しかし，あるケースにおいては，物体は未知であったと

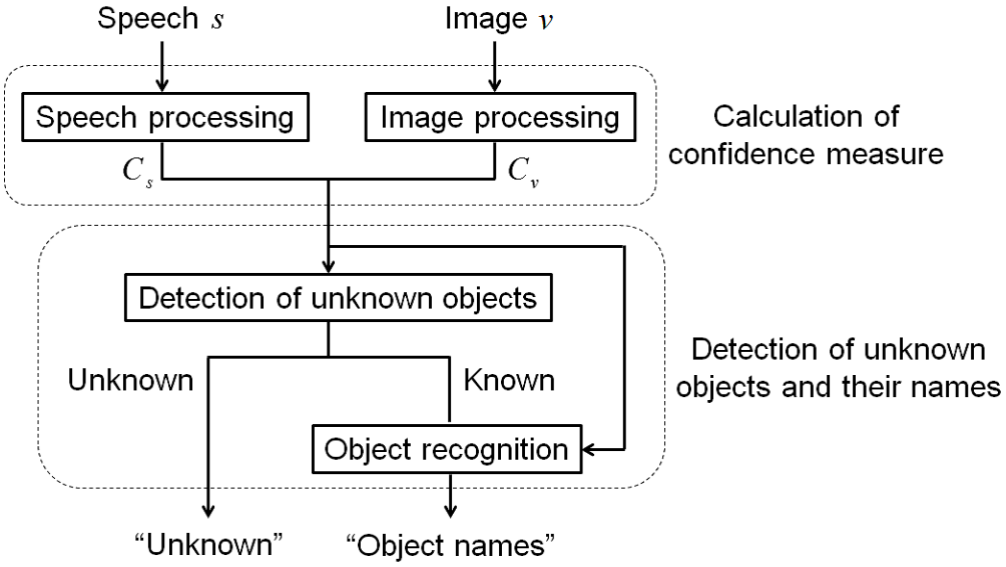


図 5.3　Proposed system configuration diagram.

しても，物体を特定することができる．このケースは，机の上にある 1 物体のみが未知であり，他の物体が全て既知であるケースで，指示された発話がロボットにとって未知である場合である．特定対象以外の物体が既知である場合は，既知物体を候補から外し，物体を特定することができる．

提案手法では，音声と画像の情報を統合したマルチモーダル情報を用いる．音声のみが既知，画像のみが既知であったとしても，物体を特定することができない．物体特定タスクを通じて，ロボットが未知物体を識別し，未知物体を学習していくことを目指す．

タスクにおいて，起こりうるケースは 3 種類ある．

- 入力音声が机の上にある既知物体の音声であるケース．
- 机の上にある複数の物体のうち，1 物体が未知であり，それ以外が既知であり．入力された音声がその未知物体に関する音声であるケース．
- 入力された音声が未知であり，机の上に複数の未知物体があるケース．

一つ目と二つ目のケースでは，ロボットは図 5.2(a) の返答を行う．三つ目のケースでは，図 5.2(b) の返答を行う．

5.3　マルチモーダル情報を用いた未知物体識別手法

一般的な物体把持タスクでは，ただ物体を把持するのみで，物体を認識することも，物体が未知であるか既知であるかの判定を行うこともない．本研究では，未知物体の検出を

行い，物体が既知であった場合は物体の認識まで行う．

提案手法は 2 つのパートで構成される．一つ目のパートは，認識結果の信頼度の推定であり，二つ目のパートは未知物体の検出である．提案手法の流れを図 5.3 に示す．未知物体検出アルゴリズムは下記のようになる．

未知物体検出アルゴリズム

入力: C_s, C_v
出力: “既知/未知”, 物体名

if $\max_i(F(C_s(s;\Lambda_i), C_v(v;o_i))) < \delta$ **then**
　出力:　“未知”, 物体名 of i
else
　出力:　“既知”, 物体名 of i

未知物体検出には，音声情報と画像情報を統合したマルチモーダル情報を用いる．音声認識結果の信頼度 $C_s(s;\Lambda_i)$ と，画像認識結果の信頼度 $C_v(v;o_i)$ を推定し，未知物体検出に用いる．ここで s は入力された音声を示し，Λ_i は，i 番目の物体の音声モデルを示す．v は入力された画像を示し，v_i は，i 番目の物体の画像モデルを示す．信頼度は，ロジスティック回帰 $F(C_s, C_v)$ により統合され，統合値を閾値 δ を用いて判定することによって，未知物体が検出される．

5.3.1 認識信頼度

提案手法では，音声認識結果と画像認識結果の信頼度を統合し，未知物体とその名前の検出に用いる．音声認識信頼度は，3 章で定義したものと同じである．入力音声を s，i 番目の物体の名前の単語 HMM モデルを Λ_i とし，$C_s(s;\Lambda_i)$ を音声認識信頼度として算出する．3 章では，色名称の画像特徴量に Lab（3 次元），物体名称の画像特徴量に輪郭のフーリエ係数（8 次元）を用いて，事後確率化により画像認識信頼度を算出していた．本章では，物体名称による指示のみを扱うが，特徴量には，輪郭のフーリエ級数（8 次元）のみならず，物体領域の Lab 特徴量の平均値（3 次元）と，物体の面積（1 次元）も用いる．物体画像の信頼度 $C_v(v;o_i)$ の算出方法は，3 章と同様である．ここで，v は入力画像，o_i は i 番目の物体のモデル (正規分布) である．

5.3.2　ロジスティック回帰を用いたモダリティ統合

音声認識と画像認識の信頼度は，必ずしも信頼できるとは限らない．例えば，音声認識信頼度はノイズのような音響条件によって影響を受けたり，画像認識信頼度は照明条件によって影響を受けることがある．よって，音声認識信頼度と画像認識信頼度を統合して，よりよく推定された信頼度を得ることが有効であると考えられる．本章では，3 章と同様に，統合のためにロジスティック回帰を用い，未知物体の識別のために，統合された信頼度を用いる．

5.3.2.1　ロジスティック回帰

音声認識信頼度と物体認識信頼度は，3 章と同様，以下のロジスティック回帰の関数によって統合される [58].

$$F(\mathbf{C}) = \frac{1}{1+\exp\{-\alpha^{\mathrm{T}}\mathbf{C}\}}, \tag{5.1}$$

ここで，$\mathbf{C}^{\mathrm{T}} = (1, C_s, C_v)$ であり，$\alpha^{\mathrm{T}} = (\alpha_0, \alpha_1, \alpha_2)$ は，ロジスティック回帰の係数である．ロジスティック回帰の学習は，3 章と同様の方法で行う．3 章では，マルチモーダル物体認識と物体特定にのみ，ロジスティック回帰による統合値を用いていたが，本章では，未知物体識別にも，ロジスティック回帰による統合値を用いる．ロジスティック回帰により，入力がモデルにマッチしているか，していないかを判定し，ロジスティック回帰の出力値を用いて，入力が未知物体か，そうでないかを判定する．各物体モデルに対応するロジスティック回帰を用いて，入力が，全ての既知モデルのうち一つにマッチしているかしていないかを判定し，入力がどのモデルにもマッチしていなければ，未知物体であるとみなす．

5.3.2.2　正則化ロジスティック回帰

ロジスティック回帰の過学習は深刻な問題である．過学習を防ぐために，正則化ロジスティック回帰を用いる．正則化ロジスティック回帰の対数尤度関数は下記のようになる [46].

$$\begin{aligned} l_R(\alpha) = &\sum_{j=1}^{M}\sum_{i=1}^{N}\{d_{i,j}\alpha^{\mathrm{T}}\mathbf{C}_j^i - \log(1+\exp(\alpha^{\mathrm{T}}\mathbf{C}_j^i))\} \\ &+ \frac{\lambda}{2}\|\alpha\|^2 \end{aligned} \tag{5.2}$$

ここで，λ は正則化係数である．パラメータ α は，3.1.4.4 と同様にして求める．

5.3.2.3　カーネルロジスティック回帰

回帰には，線形回帰と非線形回帰がある．ロジスティック回帰は非線形回帰であるが，その識別境界は線形である．識別境界が非線形であるロジスティック回帰のひとつに，カーネルロジスティック回帰がある [46]．基底関数を用いることで，ロジスティック回帰のカーネル化が可能となる．本研究では，基底関数に下記のガウシアン基底を用いる．

$$\phi_m(\mathbf{C}) = \exp(-\frac{\|\mathbf{C}-\mu_m\|^2}{2s_m^2}), \tag{5.3}$$

ここで，μ_m は基底関数の中心ベクトルであり，s_m は基底関数の勾配である．カーネルロジスティック回帰は下記のようになる．

$$F_K(\mathbf{C}) = \frac{1}{1+\exp\{-\alpha^{\mathrm{T}}\phi(\mathbf{C})\}}, \tag{5.4}$$

ここで，$\phi(\mathbf{C})$ は，(i,j) 番目の学習データに対するガウシアン基底の要素ベクトルである．つまり，$\mu_m = \mu_{i,j} = \mathbf{C}_j^i = (1, C_{s_j}^i, C_{o_j}^i,)^{\mathrm{T}}$ である．カーネルロジスティック回帰の対数尤度関数は下記のようになる．

$$l_K(\alpha) = \sum_{j=1}^{M}\sum_{i=1}^{N}\{d_{i,j}\alpha^{\mathrm{T}}\phi(\mathbf{C}_j^i) - \log(1+\exp(\alpha^{\mathrm{T}}\phi(\mathbf{C}_j^i)))\}. \tag{5.5}$$

パラメータ α はロジスティック回帰と同様にして求める．

5.3.2.4　多クラスロジスティック回帰

前述のロジスティック回帰はマルチモーダル情報を 2 クラスに識別する，2 クラスロジスティック回帰であった．2 クラスロジスティック回帰を拡張した多クラスロジスティック回帰がある [46]．

ここで，K クラスロジスティック回帰を考える．k 番目のクラスのロジスティック回帰は下記のようになる．

$$F_{M,k}(\mathbf{C}_j^i) = \frac{\exp(\alpha_k^{\mathrm{T}}\mathbf{C}_j^i)}{\sum_{p=1}^{K}\exp(\alpha_p^{\mathrm{T}}\mathbf{C}_j^i)}. \tag{5.6}$$

このときの対数尤度関数は下記のようになる．

$$\begin{aligned} l_M(\alpha) = &\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=1}^{K} d_{i,j,k}\log F_{M,k}(\mathbf{C}_j^i) \\ &+ (1-d_{i,j,k})\log(1-F_{M,k}(\mathbf{C}_j^i)), \end{aligned} \tag{5.7}$$

ここで，$d_{i,j,k}$ は教師信号であり，0 もしくは 1 となる．

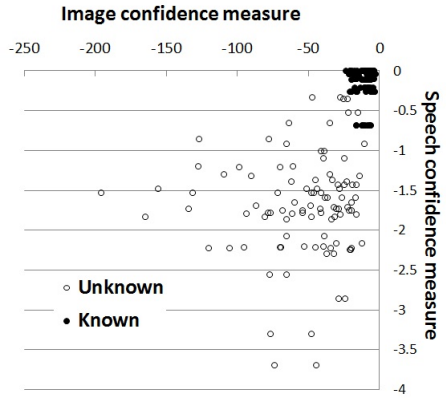


図 5.4　Joint distribution of values of speech and object confidence.

5.3.3　未知物体識別手法

入力された音声と画像のマルチモーダル情報は，認識結果の信頼度を用いて未知物体であるか，既知物体であるかの判定がなされる．未知物体と判定された場合は，入力された音声から物体名を得る．既知物体と判定された場合は，物体が特定され，かつその物体名が出力される．

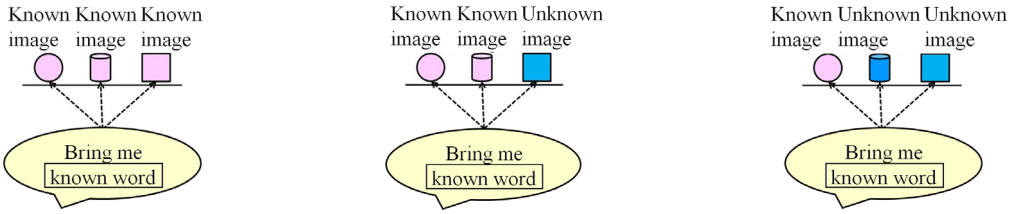
5.3.3.1　未知物体識別

図 5.4 に音声認識結果の信頼度と画像認識結果の信頼度の分布を示す．信頼度の推定には，5.4 節に示す実験データを用いた．図にプロットされたデータは 10 物体のデータである．各物体には 11 方向から撮影された画像と，各音声を示す 1 音声の 110 組（10 物体 ×11 方向 ×1 音声）が用意されている．10 物体に対する認識結果の信頼度が算出され，合計 110 データが示されている．図には，物体を既知とした場合と，未知とした場合の信頼度が示されている．この図より，音声信頼度と画像信頼度の共起性に着目することによって，未知物体と既知物体を識別可能であることが分かる．音声信頼度と画像信頼度の共起性を閾値 δ で，閾値処理することにより，未知物体と既知物体を識別する．

未知物体の識別には，ロジスティック回帰 $F(\mathbf{C}_j^i)$ を用いる．以下の条件を満たした場合，未知物体であると判定し，それ以外の場合は既知物体であると判定する．

$$\max_i F(\mathbf{C}_j^i) < \delta, \tag{5.8}$$

ロジスティック回帰の閾値 δ には 2 種類がある [70]．一つ目は，決定境界閾値であり，二つ目は，信頼境界閾値である．前者は，0.5 に設定され，後者は 0.9 に設定される．一般的に，決定境界閾値がよく使用されるが，実データの判定には信頼境界閾値の方が適して



(a)　Case 1　　(b)　Case 2　　(c)　Case 3

図 5.5　Cases where the input word is known.

いることが知られている [70]．実験では，信頼境界閾値を判定に用いた．

5.3.3.2　物体認識

既知物体であると判定された場合，物体の ID を認識する．認識は下記の方法で行われる．

$$\hat{i} = \arg\max_{i} F(\mathbf{C}_j^i). \tag{5.9}$$

物体認識結果 $\hat{i}$ が出力される．

5.3.4　複数の物体がある場合の未知物体識別

5.3.4.1　複数の物体がある場合

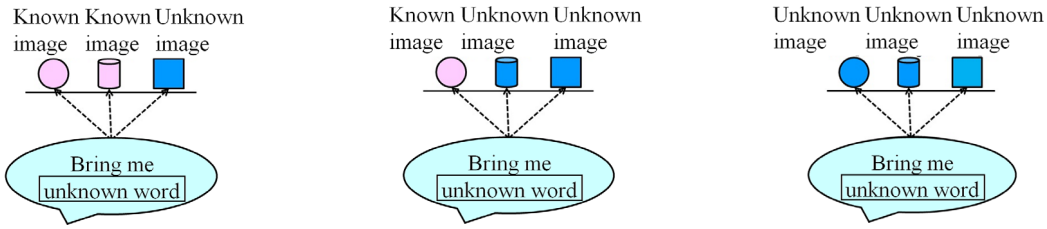
5.3.3 項で提案した未知物体識別手法を拡張することで，複数の物体がある場合でも未知物体を識別するための手法を提案する．

5.3.3 項では，入力された音声と画像は対応関係にあるという仮定を置いていた．複数の物体画像が入力された場合，音声と対応関係にある物体画像は 1 つのみであるため，この仮定が崩れる．入力された各画像が，音声と対応関係にあるか判定する必要がある．たとえ，入力された音声が既知であったとしても，入力された画像が対応関係にある画像とは限らない場合がある．

例として，3 物体が机の上にあり，その物体の中のある物体を音声で指示する場合を図 5.5，5.6 で示す．

ケース 1： 3 つの既知物体が机の上にあり，既知音声が入力される．1 つの物体が音声と対応関係にあり，それ以外の物体は対応関係にない．ロボットは，音声と対応関係にある既知物体と，音声と対応関係にない既知物体を識別することができれば，対象物体を特定することができる．

ケース 2 と 3： 入力された音声は既知であり，音声と対応関係にある既知物体があり，



(d) Case 4　　(e) Case 5　　(f) Case 6

図 5.6　Cases where the input word is unknown.

それ以外の物体が音声と対応関係にない既知もしくは未知物体である場合．このとき，音声と対応関係にある既知物体と，音声と対応関係にない既知物体，音声と対応関係にない未知物体を識別することが出来れば物体を特定することができる．

ケース 4： 入力された音声が未知物体と対応関係にあり，それ以外の物体が全て既知物体で，かつ，音声と対応関係にない場合．このとき，音声と対応関係にある未知物体と，音声と対応関係にない未知物体を識別することができれば，物体を特定することができる．

ケース 5： 入力された音声が未知物体と対応関係にあり，それ以外の物体が既知物体もしくは未知物体で，かつ，それらと音声が対応関係にない場合．このとき，音声と対応関係にない既知物体と，それ以外の未知物体を識別することができれば，特定対象候補を絞り込むことができる．

ケース 6： 机の上にある全ての物体が未知物体であり，入力された音声も未知である場合．未知物体を検出することができれば，「分かりません」と答えることができる．

本項では，これらのケースに対応可能な未知物体検出手法を提案する．提案手法は 3 つのパートから構成される．

1 つ目のパートでは，入力された音声と画像がモデルにマッチするか否かの判定を行う．まず，図 5.7 のように，音声と画像の組を認識することで得られた信頼度を C_1，C_2，C_3 に分類する．C_1 は，音声と画像の信頼度が共にモデルにマッチしない場合である．C_2 は，音声もしくは画像のいずれかの信頼度はモデルにマッチするが，もう片方はマッチしない場合である．C_3 は，音声と画像の信頼度が共にモデルにマッチする場合である．分類には 2 種類の方法がある．1 つ目の方法は，2 クラスロジスティック回帰を 2 つ用いる方法である．C_3 とそれ以外，C_1 とそれ以外を識別する 2 つのロジスティック回帰を用いる．もう 1 つの方法は，3 クラスロジスティック回帰を用いる方法である．3 クラスロジスティック回帰を用いて，C_1，C_2，C_3 に分類する．

2 つ目のパートでは，入力が未知物体であるか否かを判定する．判定では，1 つ目のパートで得られた分類結果を用いる．

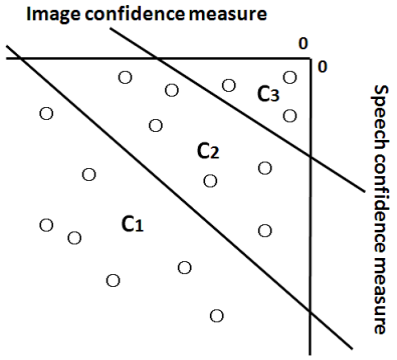


図 5.7 Matching input pairs of a speech and an image with object models.

(U) ある音声と画像の，全ての既知物体モデルに対する信頼度が C_1 に分類されている物体は，未知物体である．

(K) ある音声と画像の，全ての既知物体モデルに対する信頼度のうち，少なくとも 1 つの信頼度の組が C_3 に分類されている物体は，既知物体である．

(O) それ以外の場合は，既知音声と未知画像の組，もしくは未知音声と既知画像の組である．

3 つ目のパートでは，2 つ目のパートの結果を用いて，音声で指示された物体を特定する．例えば，図 5.5 のケース 1 の場合，1 つの物体は K と判定され，それ以外の物体は O と判定される．特定対象は K と判定された物体となる．ケース 2 の場合，3 つの物体が，それぞれ K，O，U と判定される．ケース 3 の場合，1 つの物体は K と判定され，それ以外の物体は U と判定される．ケース 4 の場合，1 つの物体が U と判定され，それ以外の物体は O と判定される．ケース 5 の場合，1 つの物体が O と判定され，それ以外の物体が U と判定される．ケース 6 の場合は，全ての物体が U と判定される．

5.4 実験

未知物体検出手法についての評価と，マルチモーダル情報を用いた物体認識手法の評価を行った．パラメータ α_0，α_1，α_2 は，実験毎に最適化した．

50 物体と，それらの物体と対応関係にある音声を各 1 音声用意した．50 物体を 11 方向から撮影し，各物体につき 11 枚の画像を用意した．物体画像の例を図 5.8 に示す．

データセットは 2 種類用意した．データセット 1，データセット 2 とする．データセット 1 は，11 方向から撮影した物体画像で構成した．データセット 2 は，5 方向から撮影した物体画像で構成した．図 5.9 に，ぬいぐるみを 11 方向から撮影した 11 画像を示す．物体画像のサイズは 640x480 ピクセルである．RGB 画像とデプス画像を Kinect [71] によ



図 5.8　Examples of objects used in the experiment.

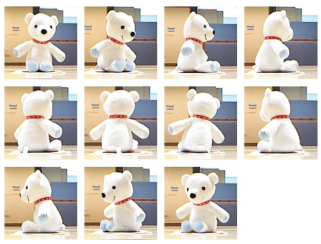


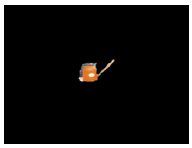
図 5.9　11 images of a bear taken from 11 different angles.



RGB image



Depth map



Extracted object region

図 5.10　Example of information obtained by Kinect.

表 5.1　Accuracy of multiple unknown object discrimination (%).

Two types of two-class logistic regression				Three class logistic regression				SVM
L	RL	KL	RKL	L	RL	KL	RKL	
82.3	85.8	92.4	97.6	88.2	90.1	91.5	98.0	95.6

り撮影し，物体領域の抽出を行った．画像例を図 5.10 に示す．実験には，抽出された物体領域画像を用いた．全ての発話は，1 人の話者によって発声されたものである．

5.4.1　未知物体検出手法の評価

本項では，leave-one-out-crossvaridation を用いて評価を行う．本実験にはデータセット 1 を用いる．未知物体検出の評価のために，音声と画像の 4 つ組み合わせの全ての組み

表 5.2 Accuracy of object recognition for respective feature (%).

Feature	Image					Speech	Logistic
	Lab	Area	Fourier	L*a*b+Fourier	All		
Dataset 1	73.6	14.6	38.0	89.6	93.0		100.0
Dataset 2	69.6	11.4	29.2	84.8	88.2	96.0	100.0

合わせを用いた．音声と画像の 4 つ組み合わせは，既知物体の名前の音声と既知物体の画像, 未知物体の名前の音声と既知物体の画像, 既知物体の名前の音声と未知物体の画像, 未知物体の名前の音声と未知物体の画像である．テストデータは 1 ペアの音声と画像とし，それ以外を学習データとする．

テストデータが既知物体の名前の音声と既知物体の画像である場合，50 物体各 9 画像 (450 画像) を画像モデルの学習データとする．テストデータが未知物体の名前の音声と画像である場合，この未知物体の画像は学習データから除外する．よって 49 物体各 9 画像 (441 画像) を画像モデルの学習データとする．テストデータの音声と画像を除いたデータを学習データとして用いて，ロジスティック回帰により物体モデルの学習を行う．49 物体に対し各 1 音声，1 画像のペアがロジスティック回帰の学習データとして使用される．本項では 2 クラスのロジスティック回帰と 3 クラスのロジスティック回帰の比較を行う．評価にはロジスティック回帰，正則化ロジスティック回帰，カーネルロジスティック回帰，正則化カーネルロジスティック回帰の 4 つのロジスティック回帰を用いる．学習の計算コストを考慮し，カーネルパラメーターおよび正規化パラメータは least-squares probabilistic classifier (LSPC) [72] を用いて決定した．パラメータは各実験で 1 つずつ最適化される．

実験結果を表 5.1 に示す．表 5.1 の L, RL, KL, と RKL はそれぞれロジスティック回帰，正則化ロジスティック回帰，カーネルロジスティック回帰，正則化カーネルロジスティック回帰を示す．polynomial kernel SVM (Support Vector Machine) を用いた手法の精度も比較した．

比較の結果，正則化ロジスティック回帰はロジスティック回帰に比べて有効であった．また，カーネルロジスティック回帰は正則化ロジスティック回帰に比べて有効であった．この結果はデータセットのペアの信頼度が大きく異なり，カーネルロジスティック回帰が有効であることを示している．また，全ての手法の中で正則化カーネルロジスティック回帰が最も有効であった．SVM はロジスティック回帰，正則化ロジスティック回帰，カーネルロジスティック回帰に比べて有効であったが，正則化カーネルロジスティック回帰に比べると有効ではなかった．本実験では 3 クラスの正則化カーネルロジスティック回帰が最も有効であった．

5.4.2　物体認識の評価

本項では，leave-one-out-crossvaridation を用いて評価を行う．既知物体が入力された場合，50 物体からテストデータとして 1 画像が選択され，残りの画像を学習データとする．データセット 1 を使用する場合，学習データは 549 画像 (50 物体 ×11 枚-1 枚) とする．データセット 2 を使用する場合，学習データは 249 画像 (50 物体 ×5 枚-1 枚) とする．実験は全ての画像に対して行われる．パラメータ α, λ は各実験ごとに最適化される．

画像認識に使用する特徴量は，前項で物体特定に用いた特徴量と同様である．物体認識の各特徴量の精度を表 5.2 に示す．表 5.2 では画像の信頼度，音声の信頼度とロジスティック回帰で統合した信頼度による物体認識の精度が示されている．各特徴の画像信頼度の精度の中で，Lab 特徴が両方のデータセットにおいて最も有効であった．統合した信頼度の精度が最も有効であった．

5.5　まとめと今後の課題

実環境下のロボットのためのインタラクティブ学習により新たな知識を獲得することは生活環境下においてロボットに必要な能力である．新たな知識を得るために，未知物体およびそれらの名前を識別し，学習することは必要である．未知物体学習のための第一歩として，未知物体を検出する手法を提案した．未知物体の検出には，音声情報を画像情報を統合したマルチモーダル情報を用いた．統合にはロジスティック回帰を用いた．ロジスティック回帰を統合に用いることは一般的であるが，未知物体の検出と既知物体の認識にどのように適用するかは決して容易な問題ではない．実験の結果，正則化カーネルロジスティック回帰が未知物体の検出に対して最も有効であることを示した．今後も生活環境下における未知物体の学習を目指す．

本手法では，入力がシステムの持つモデルのうちの 1 つと一致するかどうかを判断する際，ロジスティック回帰という簡易な方法で統合したモダリティ情報を用いた．この簡易性のために，本手法は顔画像や話者認証などの他のタスクへの適用が期待できる．

第 6 章

未知物体学習のための未知モダリティ検出

家庭用ロボットにとって，人に発話指示された物体を特定し，把持することは必要なタスクのひとつである．ロボットは，このタスクを通じて，未知の知識を獲得していくことが望ましい．知識が増えれば増えるほど，特定できる物体の数が増えるが，認識候補が増えるため，物体特定の精度は低くなる．やみくもに知識を増やすのではなく，効果的に知識を学習していく仕組みが必要である．本論文では，学習すべきモデルの数を極力抑えながら，物体特定を通じて，音声や画像を学習していくことを可能とする手法を提案する．

6.1　あらまし

物体を特定するためには，音声が画像にグラウンディングしていなければならない．本研究では，音声が画像にグラウンディングしていることを，音声と画像が同じ物体を指すことと定義する (図 6.1)．物体特定のためには，同じ物体を表している音声と画像を，ペアにして学習しておく必要がある．しかし，すべての音声と画像のペアを学習しておくことは困難である．実環境を考慮すると，音声や画像が未知である場合は勿論，音声と画像が既知であったとしても，それらのグラウンディング関係が未知である場合もある．

5 章では，未知物体を学習するための第一歩として，物体特定タスクにおいて，指示された物体が既知物体であるのか，未知物体であるのかを判定する手法を提案している．未知物体判定と物体特定には，音声認識結果と画像認識結果を統合したマルチモーダル情報を用いている．音声と画像が共に既知であり，同じモデルにマッチする場合は物体特定が可能であるが，それ以外では物体特定ができず，どのモダリティが未知であるのか判断す

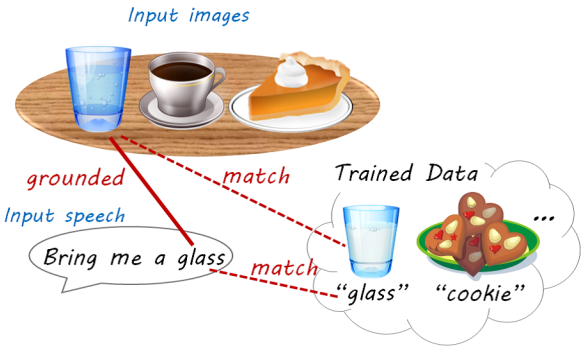


図 6.1　Object identification.

ることができなかった．そのため，未知物体と判定された場合は，音声情報と画像情報の両モダリティ情報を学習するしかなかった．

物体特定の精度は，音声認識，画像認識の精度によって変わる．両認識の精度が高ければ，物体特定の精度は向上するが，逆の場合は，精度が下がってしまう．音声と画像をひもづけた知識が少なければ少ないほど，知っている物体に対しては物体特定の精度が上がるが，知らないものが多くなり，物体特定をできない場合が多くなる．知識が多ければ多いほど，知らないものが少なくなるため，物体特定をできる場合が増えるが，その場合の物体特定の精度は低くなる．知識は増やしたいが，物体認識，物体特定の精度を下げないために，構築するモデルの数は極力おさえたい．

5 章では，ひとつの音声にはひとつの物体しかひもづけられていないという前提を置いていた．そのため，音声モデルの数だけ，物体モデルの数は増加し，物体モデルの数だけ，音声モデルが増加していく．しかし，ひとつの物体に複数の音声がひもづけられる場合も，逆に，ひとつの音声に複数の物体がひもづけられる場合もある．これに着目すると，やみくもに音声モデルの数だけ，物体モデルを増やすのではなく，モデルを増やす必要のない場合はモデルを増やさなくてもいいことが分かる．このような着眼点をもった研究は私の知る限り存在しない．

提案手法は，物体特定が不可能であった場合に，学習すべきモデルの数を極力抑えながら，音声や画像を学習していくことを目標とする．未知物体判定の際に，どのモダリティが未知であるのかを判定することができれば，未知のモダリティのみを学習すればよい．本章では，物体特定が不可能であった場合に，未知のモダリティを検出する方法を提案する．どれが未知であるのかの組み合わせにより，学習のさせ方が異なる．入力された音声と画像がどの種類であるのかの判別を行い，学習を行う．

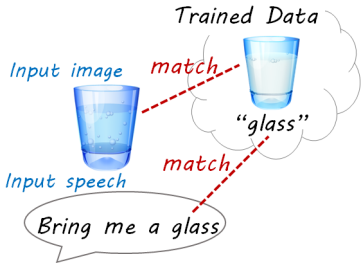


図 6.2　Case when speech and image are grounded.

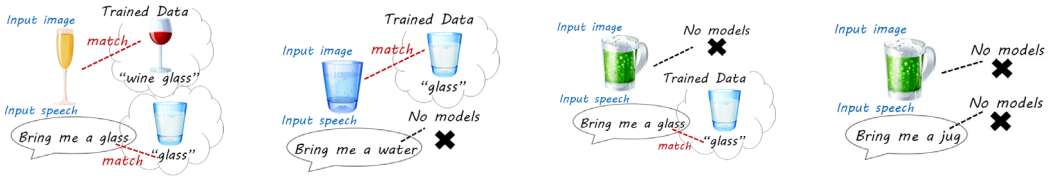


図 6.3　Case when speech and image are not grounded.

6.2　問題設定

図 6.2 に，指示された物体の音声と画像がグラウンディングされている状況を示す．この場合は，物体特定可能である．図 6.3 に，指示された物体の音声と画像がグラウンディングされていない状況を示す．指示された物体の音声と画像がグラウンディングされていなければ，どのような状況であれ，物体を特定することは不可能である．よって，物体特定のための学習データは，すべての音声と画像がペアで学習されている必要がある．本研究では，音声と画像の学習データはすべてペアで学習されていると仮定する．

物体特定が不可能である場合，どの物体がその音声が指す物体なのかを人に教えてもらうしかない．このような場合の音声と画像の組は以下のような 4 ケースがある．

(a) 音声と画像は既知であるが，マッチしていない場合．つまり，音声は既知であるが本来対応すべき画像とペアにして学習されておらず，画像は本来対応すべき音声とペアにして学習されていない場合である．この場合，音声と画像のモデルは新たに学習する必要はない．ただ，この音声と画像のモデルをペアにすればよいだけである．

(b) 音声は既知であるが，画像は未知である場合．つまり，音声と対応すべき物体画像がまだ学習されていない場合である．この場合は，音声モデルを学習する必要はなく，画像モデルのみを構築すればよい．[19] の手法により，音声認識結果から，画

表 6.1　5 cases.

	Speech	Image	Grounding
Case 1	Known	Known	○
Case 2	Known	Known	×
Case 3	Known	Unknown	×
Case 4	Unknown	Known	×
Case 5	Unknown	Unknown	×

像検索を行い，画像を収集すればよい．収集した画像で構築した画像モデルと，音声モデルをペアにすればよい．

(c) 画像は既知であるが，音声は未知である場合．つまり，物体画像と対応すべき音声がまだ学習されていない場合である．この場合は，画像モデルの学習は不要である．音声のみを学習すればよい．音声から音響モデルを構築し，音声モデルを構築すればよい．画像モデルとこの音声モデルをペアにすればよい．

(d) 音声，画像ともに未知である場合．つまり，音声も画像も学習されていない場合．この場合のみ，音声，画像ともに学習が必要となる．

これらのケース (a)，(b)，(c)，(d) は，ぞれぞれ，図 6.3 の (a)，(b)，(c)，(d) のケースに相当する．モダリティが既知であるのか，未知であるのかの判断ができれば，この 4 ケースに音声と画像の組を分類することができる．どのケースであるのかが判断できれば，そのケースにおいて効果的な学習を行えばよい．

音声が（画像が）未知であるということは，音声 (画像) モデルが学習されていないこと，つまり，モデルがないことである．音声が既知の場合は，既知の音声モデルとマッチする．音声が未知の場合は，既知の音声モデルとマッチしない．この事実に着目して，本論文では，未知のモダリティの検出を行い，音声と画像のペアのケースを判断する．

6.3　未知モダリティ検出

ある音声と画像が与えられたとき，それぞれを音声モデルと画像モデルのペア集合を用いて音声認識，画像認識し，音声信頼度群と画像信頼度群を算出する．得られた音声信頼度群と画像信頼度群を用いて，未知モダリティを検出する．また，音声・画像がともに既知の場合でも，音声と画像のグラウンディングが既知かどうかの識別も行う．

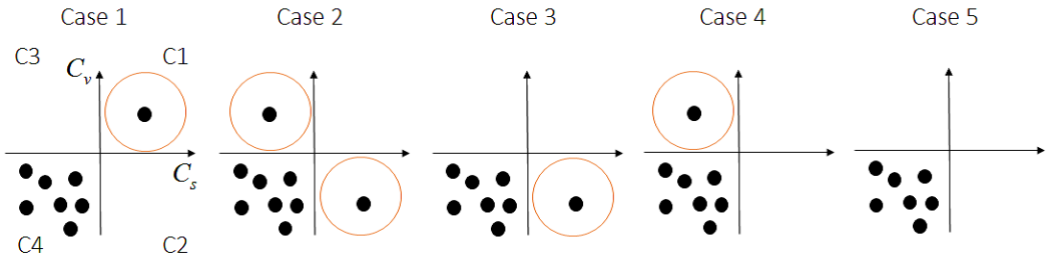


図 6.4 Conceptual diagram of confidence distribution in 5 cases.

6.3.1 未知モダリティ

ある音声と画像のペアが与えられたとき，音声，画像，グラウンディングに着目すると，図 6.2 の 1 ケースと図 6.3 の 4 ケースの合わせて 5 ケースに分けられる．表にまとめると，表 6.1 のように表せる．これらの 5 ケースに分類することで，未知モダリティの検出を達成できる．つまり，音声 s と画像 v が与えられたとき，ケース 1 からケース 5 の 5 クラスに分類することが目標となる．

まず，音声認識と画像認識によって，音声信頼度と画像信頼度を算出する．音声信頼度と画像信頼度は 3.2 節と同様の方法で算出する．入力音声 s の，m 番目の物体の音声モデル Λ_m に対する音声信頼度を $C_s(s;\Lambda_m)$ とする．m 番目の物体の画像モデル o_m に対する，入力画像 v の信頼度を $C_v(v;o_m)$ とする．

ある入力音声を音声認識した際，入力音声とマッチする音声モデルとマッチしない音声モデルの二つが存在する．認識の際に得られる信頼度は，前者は高い値，後者は低い値となる．画像についても同様のことが言えるため，ある入力音声と入力画像のペアに対して，音声信頼度と画像信頼度のペアは，4 クラスに分けることができる．これらを C1 (音声信頼度高，画像信頼度高)，C2 (高，低)，C3 (低，高)，C4 (低，低) と呼ぶ．

そこで，5 クラスの音声信頼度と画像信頼度をプロットすると，図 6.4 のようになる．C1, C2, C3 領域にそれぞれ点が存在するかしないかで，5 クラスを分類するというアプローチをとる．それらの 3 点を本論文ではキーポイントと呼び，C1, C2, C3 に対応するキーポイントを，それぞれキーポイント 1, キーポイント 2, キーポイント 3 と呼ぶ．3 つのキーポイントがそれぞれ存在するかしないかを判定することが重要な要素となる．

6.3.2 未知モダリティ検出のための識別器

5 クラス分類は，以下の 2 つの識別器を組み合わせて実現する．識別器は，キーポイントが存在するかしないかを識別することができ，3 つのキーポイントを順番に識別する．

本項では，2 つの識別器の説明をする．

識別器 1:　ある音声 s が与えられたとき，それが既知であるかどうかは，学習済みの音声モデル集合の中にマッチする音声モデルが存在するかどうかで識別する．つまり，以下の式のように，全音声モデルに対して音声 s の信頼度を算出し，その中に閾値よりも高い値があるかどうかで，既知か未知かを識別する．

$$\max_m C_s(s;\Lambda_m) > \delta_s, \tag{6.1}$$

画像 v が与えられたときも，同様に以下の条件式を用いて，既知か未知かを識別する．

$$\max_m C_v(v;o_m) > \delta_v, \tag{6.2}$$

δ_s, δ_v は，それぞれ，音声信頼度，画像信頼度の閾値を示す．

識別器 2:　音声認識と画像認識の信頼度は，必ずしも信頼できるとは限らない．例えば，音声信頼度はノイズのような音響条件によって影響を受けたり，画像信頼度は照明条件によって影響を受けることがある．よって，音声信頼度と画像信頼度を統合して，よりよく推定された信頼度を得ることが有効であると考えられる．本章では，5 章と同様，以下のロジスティック回帰を用いて信頼度を統合し，未知物体の識別に用いる．

$$F(\mathbf{C}) = \frac{1}{1+\exp\{-\alpha^{\mathrm{T}}\mathbf{C}\}}, \tag{6.3}$$

ここで，$\mathbf{C}^{\mathrm{T}} = (1, C_s, C_v)$ であり，$\alpha^{\mathrm{T}} = (\alpha_0, \alpha_1, \alpha_2)$ は，ロジスティック回帰の係数である．ロジスティック回帰の学習は 5 章と同様にして行う．

本章では，3 種類のロジスティック回帰 $F_n(n = 1, 2, 3)$ を用いる．F_1 は C1 と C2, C3, C4，F_2 は C2 と C3, C4，F_3 は C3 と C4 のように学習させておく．F_1, F_2, F_3 を用いて，それぞれ，C1 かどうか，C2 かどうか，C3 かどうかを識別する．

$$\max_m F_n(C_s(s;\Lambda_m), C_v(v;o_m)) > \delta_n, \tag{6.4}$$

δ_n は，n 番目のロジスティックの識別境界を表す閾値である．

6.3.3　クラス分類アルゴリズム

前項の 2 つの識別器を組み合わせて，5 クラス分類を実現する．図 6.5 にアルゴリズムを示す．3 つのキーポイントを順番に識別するため，3 段階で構成される．2 つの識別器を用いる順番と組み合わせは，複数存在するが，本論文では以下の 2 つの手法について述べる．以下に詳細を述べる．

アルゴリズム 1:　識別器 2 によってキーポイント 1 を識別することで，ケース 1 かどうかを分類する．識別器 1 によってキーポイント 2 を識別し，識別器 1 によってキーポイント 3 を識別することで，残りの 4 クラスに分類する．

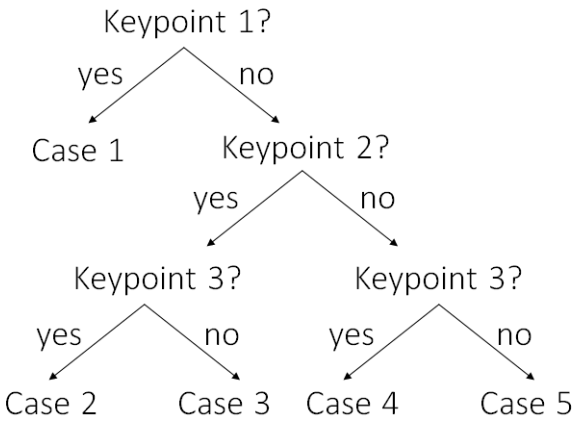


図 6.5 Algorithm for 5-class classification.

表 6.2 Composition of dataset.

	Test data	Training data	Varidation data
Known speech (Object 1-100)	2 speeches / object	2 speeches / object	2 speeches / object
Unknown speeche (Object 101-200)	2 speeches / object		2 speeches / object
Known image (Object 1-100)	25 images / object	20 images / object	5 images / object
Unknown image (Object 101-200)	25 images / object		25 images / object

アルゴリズム 2: 識別器 2 によってキーポイント 1 を識別することで，ケース 1 かどうかを分類する．識別器 2 によってキーポイント 2 を識別し，識別器 2 によってキーポイント 3 を識別することで，残りの 4 クラスに分類する．

6.3.4 実験

データセットの作成法:

ケース 1 から 5 のサンプルを作成した．データセットは，ILSVRC2013 の CLS-LOC データセットに含まれる 1000 カテゴリの中から，artifact ノードより下層にある 514 個の物体カテゴリ名のみで構成した．この 514 物体の中から画像認識精度が高い 100 物体を既知物体として用いた．514 物体から既知 100 物体を除き，その中から未知 100 物体をランダムに選択した．

既知 100 物体は学習済みとした．その 100 物体を入力として音声・画像認識を行うこ

とで，既知物体を想定し，未学習の 100 物体を入力として音声・画像認識を行うことで，未知物体を想定した．既知物体には，各物体につき 50 枚の画像と 6 発話の音声が付与されており，どちらも，テストデータ，ロジスティック回帰の学習データ，ロジスティック回帰のハイパーパラメータのバリデーションデータの 3 つに分けられる．未知物体には，各物体につき 25 枚の画像と，2 発話の音声が付与されており，どちらも，テストデータとして用いた．データセットの詳細な内訳は表 6.2 に示す通りである．数値は 1 物体あたりのものである．

音声はノイズがほとんどない環境で，2 名の話者が物体名を発話したものを録音した．それぞれの話者が，既知 100 物体に対してそれぞれ 3 回ずつと，未知 100 物体に対してそれぞれ 1 回ずつ発話した．画像は，ILSVRC2013 の CLS-LOC データセットのバリデーションデータセットに含まれる画像を用いた．各物体につき 50 枚の画像が含まれているので，既知 100 物体についてはその中から 50 枚全て，未知 100 物体については 50 枚のうち 25 枚を使用した．

音声認識と画像認識: Julius で HMM による不特定話者孤立単語認識を行った．音響モデルには，Julius ディクテーションキット [49] に含まれている，不特定話者の PTM トライフォン Hidden Markov Model (HMM) を用いた．特徴ベクトルとして，メル周波数ケプストラム係数（MFCC）12 次元とそのデルタ，対数パワーの合計 25 次元のベクトルを用いた．OverFeat で CNN による一般物体認識を行う．画像モデルは，ILSVRC2013 の CLS-LOC データセット中の学習データセットに含まれる 1,281,167 枚の画像を用いて学習されている．1000 物体があらかじめ学習されているため，1000 物体全てについてのスコアが得られるが，その中から既知 100 物体のみのスコアを用いた．

ロジスティック回帰の学習: ロジスティック回帰の学習のために必要となる，C1, C2, C3, C4 の 4 クラスの学習データの作成法について述べる．学習には，データセットの中の学習データを用いた．未知音声・未知画像は用いず，既知音声・既知画像のみで，ロジスティック回帰は学習できることに注意する．ある 1 音声と 1 画像のペアが与えられたとき，両者が対応関係にあるのかそうでないのかの二種類に分けられる．音声と画像のペアについて，音声認識・画像認識を行うと，それぞれ 100 個の信頼度が得られる．音声信頼度群と画像信頼度群は，100 個それぞれモデルごとにペアとなって紐づけられている．i) 対応関係にあるペアが与えられたとき，音声信頼度と画像信頼度のペア 100 個のうち，1 ペアのみが C1，それ以外の 99 個が C4 となる．ただし，その 99 個の中からランダムに 1 個を学習データとして選んだ．ii) 対応関係にないペアが与えられたとき，音声信頼度と画像信頼度のペア 100 個のうち，1 個が C2，1 個が C3，それ以外の 98 個が C4 となる．ただし，その 98 個の中からランダムに 1 個を学習データとして選んだ．このような音声信頼度と画像信頼度のペアをデータセット中の学習データを利用して作成した．

このような作業を，以下に示す回数だけ繰り返し行うことで，学習データを作成した．

表 6.3 Accuracy of unknown modarity detection (%).

Algorithm 1	Algorithm 2
68.1	67.9

表 6.4 Accuracy of unknown/known modarity detection by thresholding confidence measure of speech and image (%).

Speech confidence	Image confidence
90.5	73.1

表 6.5 Accuracy of discrimination by logistic regression used in detection of unknown/known modarity (%).

Logistic regression F_1	98.3
Logistic regression F_2	96.8
Logistic regression F_3	94.3

1. 対応関係にあるペアの数は，音声が示す物体と画像が示す物体のペア 100 通りである．音声の物体と画像の物体のペアが決まれば，音声はその物体の中から 2 音声，画像は 20 画像を使用できるので，$2 \times 20 = 40$ 通りの音声と画像のペアを作成した．よって，作業の繰り返し回数は，$100 \times 40 = 4000$ 回となる．
2. 対応関係にないペアの数は，音声が示す物体と画像が示す物体のペアは ${}_{100}P_2$ 通りである．その中からランダムに 100 ペアを選んだ．音声が示す物体と画像が示す物体のペアが決まれば，1. と同様にして，音声は 2 発話，画像は 20 画像を学習データとして使用できるので，$2 \times 20 = 40$ 通りの音声と画像のペアを作成した．よって，作業の繰り返し回数は，$100 \times 40 = 4000$ 回となる．

6.3.4.1 実験結果

アルゴリズム 1 と 2 を用いた未知モダリティ検出結果を表 6.3 に示す．全ての処理でロジスティック回帰を用いて未知モダリティを検出したアルゴリズム 2 よりも，音声信頼度，もしくは，画像信頼度を閾値判定することで未知モダリティを検出したアルゴリズム 1 の方が精度が高かった．

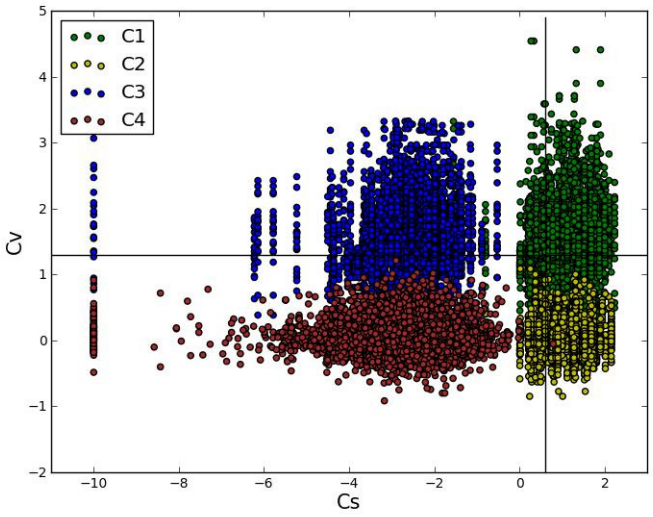


図 6.6　Thresholding border and confidence measure of known/unknown speech and image. Green dots denote C1, yellow dots denote C2, blue dots denote C3, and red dots denote C4.

6.3.4.2　考察

アルゴリズム 1 の閾値判定による未知モダリティ検出精度を表 6.4 に示す．また，アルゴリズム 2 の未知モダリティ検出に用いられたロジスティック回帰の識別精度を表 6.5 に示す．ロジスティック回帰 1 は，音声信頼度と画像信頼度がモデルにマッチしており，かつ，対応関係にあるか否かを判定するロジスティック回帰である．ロジスティック回帰 2 は，音声信頼度がモデルにマッチしており，画像信頼度がモデルにマッチしていない場合であるか，否かを判定するロジスティック回帰である．ロジスティック回帰 3 は，音声信頼度がモデルにマッチしておらず，画像信頼度がモデルにマッチしている場合であるか，否かを判定するロジスティック回帰である．図 6.6 に，既知/未知の音声と画像の信頼度と閾値境界を示す．図 6.7 に，既知/未知の音声と画像の信頼度と，ロジスティック回帰による識別境界を示す．アルゴリズム 1 とアルゴリズム 2 での，未知モダリティ検出精度を比べると，精度はほぼ同じとなった．

6.3.5　まとめと今後の課題

本章では，物体特定タスクにおいて，特定対象のグラウンディングが未知であった場合に，未知のモダリティを検出するための手法を提案した．音声信頼度や，画像信頼度を閾値処理することで，未知モダリティを検出する手法と，音声信頼度と画像信頼度をロジス

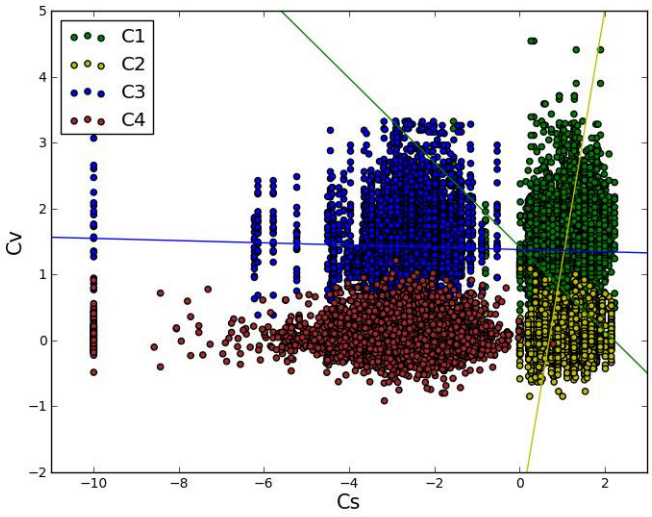


図 6.7 Discriminant border by logistic regression and confidence measure of known/unknown speech and image. Green dots denote C1, yellow dots denote C2, blue dots denote C3, and red dots denote C4.

ティック回帰で統合し，未知モダリティを検出する手法を提案した．実験により，閾値判定を行う手法とロジスティック回帰を用いる手法は，ほぼ同じ精度であることが分かった．しかし，識別器の精度を比較すると，ロジスティック回帰を用いる手法の方が精度がよかった．ロジスティック回帰を用いた未知モダリティ検出アルゴリズムの開発を今後の予定とする．

今後は，検出された未知モダリティに合わせた，効果的な学習手法と物体特定手法が必要となる．次節では，グラウンディング関係と画像のみが未知である場合 (ケース 3) の物体の学習方法と物体特定手法を提案する．他のケースにおける物体学習手法と物体特定手法の開発を今後の課題とする．

6.4 Web 画像を用いた物体特定手法

音声が既知であるが，グラウンディング関係と画像が未知である場合に，物体を学習する手法と物体を特定する手法について本節で提案する．

音声が既知である場合，音声認識結果を用いて Web 画像検索を行い，収集した画像を用いて一般物体認識を行い，物体特定を行えばよい．Web 画像を用いた物体特定手法の流れは以下のようになる．

1. 発話 s を音声認識する．

2. 音声認識の結果を用いて，候補の絞り込みを行う．
3. 音声認識結果 (複数) をクエリとして Web 画像検索を行い，物体カテゴリごとに画像を収集する．
4. 得られた Web 画像によって，物体ごとに画像モデルを構築する．
5. 目の前にある複数の物体 $v_n(n = 1, 2, \cdots, N)$ を画像認識する．
6. それらの認識結果を統合して，物体特定を行い，認識結果とともに出力する．

6.4.1　Web 画像による画像モデルの構築

Web 画像を用いて物体カテゴリごとに画像モデルを構築し，それらのモデルを用いて画像認識を行う．画像から，スケール変化等に頑健な局所特徴量である SIFT [56] を抽出する．以下のように，画像 v_n から，D 次元の局所特徴量が L 個得られたとする．

$$\boldsymbol{X}_n = [\boldsymbol{x_{n1}}, \boldsymbol{x_{n2}}, \cdots, \boldsymbol{x_{nL}}] \in \boldsymbol{R}^{D \times L}. \tag{6.5}$$

得られた局所特徴量の集合 $\boldsymbol{X}_n$ を，Locality-constrained Linear Coding (LLC) [53] のコード化関数 Φ を用いて，特徴ベクトル $\boldsymbol{x_n}$ に変換する．

$$\boldsymbol{x_n} = \Phi(\boldsymbol{X}_n). \tag{6.6}$$

[73] では，学習データに Web 画像を用いる場合，画像枚数が十分であれば，k-Nearest Neighbor (kNN) が有効であることが示されている．学習データは大量に収集可能であるため，本研究でも kNN に基づいた画像モデルを構築する．m 番目の物体カテゴリの画像モデル $o_m(m \in A)$ に対して，入力画像 v_n の信頼度 $C_v(v_n; o_m)$ は，以下の式によって得られる．

$$C_v(v_n; o_m) = -\sum_{k=1}^{K} d(\boldsymbol{x_n}, neigh(k, \boldsymbol{x_n}, o_m)), \tag{6.7}$$

ただし，$d(\boldsymbol{a}, \boldsymbol{b})$ は，$\boldsymbol{a}, \boldsymbol{b}$ 間のユークリッド距離を示し，$neigh(k, \boldsymbol{x_n}, o_m)$ は，カテゴリ m に属する特徴ベクトルの中で，$\boldsymbol{x_n}$ との距離が k 番目に小さい特徴ベクトルを示す．K は kNN のパラメータの値を示す．音声 s と画像 $v_n(n = 1, 2, \cdots, N)$ の組 $\{s, v_n\}$ に対するマルチモーダル物体認識結果 $\hat{m_n}$ は式 (3.13)，物体特定結果は式 (3.14) と同様にして得られる．

6.4.2　実験

Julius [42] を用いて不特定話者孤立単語音声認識を行った．単語辞書は，家庭内にあると想定される物体カテゴリ 1000 個の名前で構成した．物体特定のテストデータには，そ

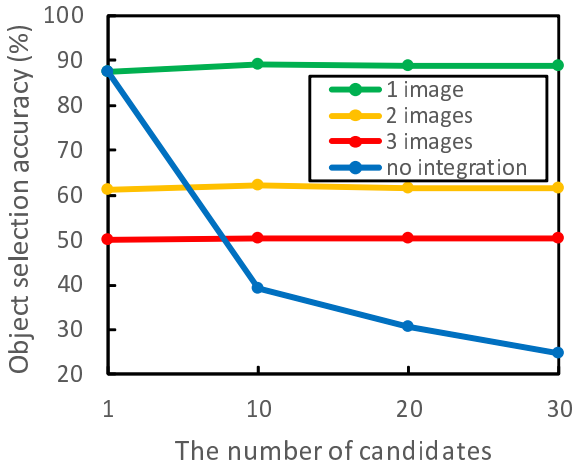


図 6.8　Accuracy of object identification（%）.

の 1000 カテゴリの中から，ランダムに選んだ 100 カテゴリを用いた．音響モデルには，Julius ディクテーションキット [49] に含まれている，不特定話者の PTM トライフォン HMM を用いた．特徴ベクトルとして，MFCC12 次元とそのデルタ，対数パワーの合計 25 次元のベクトルを用いた．話者数は 1 名であり，ノイズがほとんどない環境下で 100 カテゴリ名を各 1 回発話し，それらを録音した．100 発話のうち，80 発話をテストデータ，残りの 20 発話をロジスティック回帰の学習データとした．

音声認識結果にしたがって，Web 画像によって構築された画像モデルを用いて画像認識を行った．Web 画像の収集には，ImageSpider [74] と ImageGetter [75] を用い，kNN の実装には Scikit-learn [76] を用いた．各物体カテゴリにつき，210 画像を収集した．210 画像のうち，10 画像はテストデータ，残りの 200 画像は画像モデルとロジスティック回帰の学習データとした．画像は，アスペクト比を固定したまま，短辺が 150 ピクセル以下となるようにリサイズした．SIFT は画像に対して，10 ピクセルごとのグリッド点から抽出した．LLC で用いるコードブックは，k-means によって得られた 100 個のセントロイドで構成した．LLC についての他のパラメータは，Wang ら [53] の実装で設定されている値と同じである．画像モデルのパラメータ K は，予備実験によって最も物体特定精度が高くなる 10 と設定した．

物体特定データセットの作成は，物体カテゴリ 100 個の物体カテゴリを使用し，3.2.4.1 と同様にして行った．ロジスティック回帰の学習データの作成には，100 個の物体カテゴリを使用した．各物体カテゴリに対して，1 音声と 200 画像が使用可能であるので，3.2.4.1 と同様にして学習データを作成した．

音声認識精度は 88% であった．物体特定の結果を図 6.8 に示す．入力画像枚数が 1 枚

の場合は，物体が正しく認識された場合にタスクが達成されたとしている．入力画像枚数が 2，3 枚の場合は，音声によって指示された物体を特定でき，かつ，特定された物体が正しく認識された場合にタスクが達成されたとしている．“no integration”は，音声認識を行い，音声による候補の絞り込みは行うが，音声認識信頼度と統合することはせず，入力画像 1 枚に対して，画像認識信頼度のみを用いて物体認識を行った際の結果である．すなわち，音声認識により候補を絞り込み，この候補を用いて入力画像を画像モデルで認識し，画像認識信頼度の最も高いものを認識結果とした結果である．物体特定精度は，入力画像枚数が 3 枚のとき，約 50% (候補数 10) となった．さらに精度を上げるためには，物体特定の基となる “no integration”，つまり画像認識精度を向上させる必要がある．

6.4.3　まとめと今後の課題

本節では，音声が既知である場合に，物体が未知であったとしても物体を特定可能な手法を提案した．提案手法では，音声認識結果を用いて，Web で画像を収集することによって物体モデルを構築し，物体特定を行った．本手法により，音声が既知である場合に物体を学習すると同時に物体を特定することが可能となる．しかし，物体によっては，Web 画像を用いた物体認識の精度が十分に高くなかったため，物体特定率が下がってしまった．Web 画像を収集する際の検索としての精度向上と，Web 画像を用いた物体認識率の向上が望まれる．前者では，収集した画像内に合い異なる概念を持つ画像（同音異義語など）が入り込む問題や，対象物体以外の物体や，人などが同一画像内に入り込む問題などを解決する必要がある．後者の問題は，一般物体認識の抱える問題と同一の問題である．より頑健な特徴量，識別器，検出器などの開発が望まれる．

第 7 章

おわりに

家庭用サービスロボットが，人と共生するためには，ロボットが人の指示を理解する必要がある．指示を理解するためには，指示された事象や概念を実世界へのグラウンディングする必要がある．

本研究では，ロボットが音声や画像などの複数のモダリティを用いて，物体を実世界にグラウンディングするための方法について論じた．具体的には，ロボットが人に指示された物体を複数の物体の中から特定する方法についての提案を行った．物体を特定するためには，マルチモーダル処理が必要となる．本研究ではパターン認識のアプローチを用いて，マルチモーダル処理を行った．マルチモーダル認識を拡張することにより，物体の特定を可能とした．この処理はグラウンディングのひとつであると言える．本稿では，物体特定タスクにまつわる問題を提起し，各問題についての解決策を提案した．

第 3 章では，複数の属性に基づく表現を用いた物体特定手法を提案した．本研究では，色名称と物体名称の 2 属性を用いて物体の特定を行った．音声認識と画像認識の結果をロジスティック回帰で統合し，統合した値を用いて物体特定を行った．また，ある物体をどのように表現するかは人によって異なることに着目し，人によって表現が異なる場合でも，物体特定ができるような枠組みを提案した．実験により，1 属性を含む表現で指示した場合よりも，複数属性を含む表現を用いた方が物体特定精度が高くなることが確かめられた．本研究ではロボットに学習させていない表現による指示には対応できていない．今後は，未学習の指示においても対応可能な手法を研究する予定である．また，本研究では，設定したタスクに対する提案手法の有効性を示すためにデータセットを構築した．構築したデータセットでは，人が実物体を見て，発話指示を行っている．人が物体画像を見て発話指示する場合との比較は今後の予定である．色名称や物体名称に基づいて物体認識を行う際，最も適した画像特徴量や，認識手法について研究することも今後必要な課題の

ひとつである．本研究では，属性として色名称と物体名称のみしか扱っていないが，提案手法の枠組みは，他の属性にも拡張可能である．大きさや，テクスチャ，位置関係などの属性を今後与えていく予定である．発話指示のフォーマットや，物体に遮蔽がないという制約も緩和し，さらに自然な発話の中でロボットが物体を特定できるよう，本研究を発展させていく予定である．

第 4 章では，色による物体特定タスクのための，人の視覚属性を考慮した顕著性検出手法を提案した．提案手法で算出した顕著性を定性的に評価することで，提案手法は人の視覚属性に近い顕著性を検出できる可能性があることを確かめた．また，提案手法を用いて算出した顕著性マップを用いて物体特定タスクを行い，提案手法の定量的な評価を行った．提案手法により検出した顕著性は，物体の識別能力が高いとは言いがたいが，人の視覚属性をよく反映することができた．今後，提案手法を物体の識別能力向上も考慮した手法に拡張していきたい．また，本研究では物体特定に用いる属性を色のみに限定したが，この制約を外し，物体の材質やテクスチャなど様々な属性にも対応できるように拡張していきたい．人とロボットの視覚属性の違いを明らかにし，人がロボットとの違いを感じることなく，ロボットとの自然なインタラクションを楽しみながら共生していく未来を実現する手助けがしたいと考えている．

第 5 章では，物体特定タスクにおいて，未知物体を指示された際に，未知物体を検出するための手法を提案した．未知物体の検出は，未知物体の学習に繋がる．未知物体の検出には，音声情報と画像情報を統合したマルチモーダル情報を用いた．統合にはロジスティック回帰を用いた．ロジスティック回帰を統合に用いることは一般的であるが，未知物体の検出と既知物体の認識にどのように適用するかは決して容易な問題ではない．実験の結果，正則化カーネルロジスティック回帰が，未知物体の検出に対して最も有効であることが示された．今後も生活環境下における未知物体の学習を目指す．

第 6 章では，物体特定タスクにおいて，特定対象のグラウンディングが未知であった場合に，未知のモダリティを検出するための手法を提案した．音声認識の信頼度や，画像認識の信頼度を閾値処理することで，未知モダリティを検出する手法と，音声認識の信頼度と画像認識の信頼度をロジスティック回帰で統合し，未知モダリティを検出する手法を提案した．実験により，閾値判定を行う手法の方が有効であることが分かった．しかし，識別器の精度を比較すると，ロジスティック回帰を用いる手法の方が精度がよかった．ロジスティック回帰を用いた未知モダリティ検出アルゴリズムの開発を今後の予定とする．

本稿で扱った問題は，本研究の抱える問題の一部に過ぎない．パターン認識，コンピュータビジョン，自然言語処理，認知科学など，問題は多岐に渡る．マルチモーダル処理を可能とするロボットには，様々な分野の技術と，それらの統合技術が必要である．また，人とロボットが自然な形で共生するためには，工学技術以外の要素，心理学などの社会科学的要素も必要とされる．本研究がひとつの分野として確立され，様々な分野の研究

者が協力し合って，研究が進んでいくことを強く願う．

付録

A 有効視野の算出方法

3.1.6.1 節で用いた有効視野範囲の算出方法について述べる．本研究における計測条件を図 7.1 に示す．人の有効視野は，視力が 1.0 の人であれば，左右約 15 度，上約 8 度，下

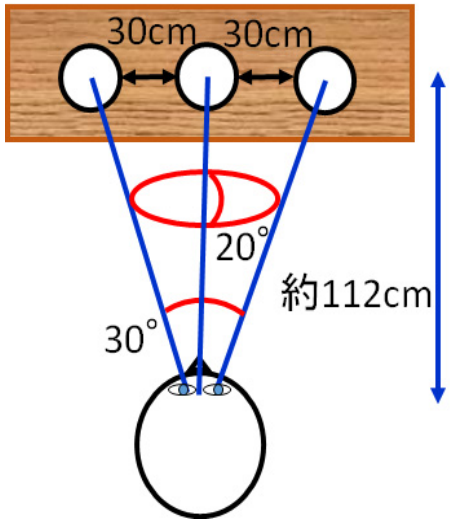


図 7.1　Effective visual field.

約 12 度以内の範囲となる．本研究の協力者は全員矯正視力 1.0 の視力であったため（内に 1 名裸眼視力 1.0 を含む），上記の範囲を有効視野と設定した．3 物体は 30cm の間で等間隔に並べた．中心に置く物体位置を視野中心とした．3 物体が有効視野いっぱいに入るよう，協力者と物体までの距離を算出した．このとき，協力者の首の角度は 15 度を仮定した．椅子に座った状態での協力者の目の高さは平均約 119cm，机の高さは 70cm であり，有効視野を考慮して，人と物体までの距離を算出すると，平均で約 112cm であった．

B　発話結果

3.1.6.3 節の実験にて得られた全シーンにおける全協力者の発話を表 7.1 に示す．協力者は 7 人である．各シーンにおいて，左から順に，物体 1，物体 2，物体 3 であるとする．各シーンにおいて，各物体をロボットに指示する想定で，色名称，物体名称，もしくは両名称を用いて指示してもらった．

表 7.1　Speeches of 7 participants in 15 scenes.

Scene index	Scene	Participant	Object 1	Object 2	Object 3
1		1	Bring me a christmas tree.	Bring me a vase.	Bring me a tea.
		2	Bring me a tree.	Bring me a green vase.	Bring me a tea.
		3	Bring me a model.	Bring me a green bottle.	Bring me a green plastic bottle.
		4	Bring me a christmas tree.	Bring me a vase.	Bring me a plastic bottle.
		5	Bring me a tree.	Bring me a vase.	Bring me a plastic bottle.
		6	Bring me a green tree.	Bring me a green bottle.	Bring me a green tea.
		7	Bring me a tree.	Bring me a bottle.	Bring me a tea.
2		1	Bring me a mugcup.	Bring me a cage.	Bring me a yellow-green case.
		2	Bring me a cup.	Bring me a cage.	Bring me a trash box.
		3	Bring me a mugcup.	Bring me a cage.	Bring me a green trash box.
		4	Bring me a mugcup.	Bring me a cage.	Bring me a green one.
		5	Bring me a mugcup.	Bring me a cage.	Bring me a trash box.
		6	Bring me a cup.	Bring me a cage.	Bring me a trash box.
		7	Bring me a cup.	Bring me a cage.	Bring me a frog.
3		1	Bring me a black one.	Bring me a pail.	Bring me a ball.
		2	Bring me a cap.	Bring me a pink one.	Bring me a pink ball.
		3	Bring me a cap.	Bring me a pink pail.	Bring me a pink ball.
		4	Bring me a silk hat.	Bring me a pail.	Bring me a ball.
		5	Bring me a silk hat.	Bring me a pail.	Bring me a ball.
		6	Bring me a cap.	Bring me a pink pail.	Bring me a pink ball.
		7	Bring me a cap.	Bring me a pail.	Bring me a ball.
4		1	Bring me a brown one.	Bring me a pan.	Bring me a sieve.
		2	Bring me a brown one.	Bring me a pan.	Bring me a basket.
		3	Bring me a flowerpot.	Bring me a pan.	Bring me a sieve.
		4	Bring me a flowerpot.	Bring me a pan.	Bring me a sieve.
		5	Bring me a brown one.	Bring me a pan.	Bring me a sieve.
		6	Bring me a brown one.	Bring me a white pan.	Bring me a sieve.
		7	Bring me a brown one.	Bring me a pan.	Bring me a sieve.
5		1	Bring me a yellow one.	Bring me a bucket.	Bring me a red book.
		2	Bring me a yellow one.	Bring me a bucket.	Bring me a book.
		3	Bring me a shampoo.	Bring me a red bucket.	Bring me a book.
		4	Bring me a pump bottle.	Bring me a red bucket.	Bring me a book.
		5	Bring me a yellow one.	Bring me a bucket.	Bring me a red book.
		6	Bring me a yellow dispenser.	Bring me a bucket.	Bring me a red book.
		7	Bring me a shampoo.	Bring me a bucket.	Bring me a book.
6		1	Bring me a tissue.	Bring me a pink cushion.	Bring me a orange one.
		2	Bring me a tissue.	Bring me a pink one.	Bring me a orange one.
		3	Bring me a tissue.	Bring me a pink cushion.	Bring me a color cone.
		4	Bring me a tissue.	Bring me a pink one.	Bring me a traffic cone.
		5	Bring me a tissue.	Bring me a pink one.	Bring me a orange one.
		6	Bring me a tissue.	Bring me a pink cushion.	Bring me a traffic cone.
		7	Bring me a tissue.	Bring me a pink one.	Bring me a pole.
7		1	Bring me a yellow one.	Bring me a light blue elephant.	Bring me a white bear.
		2	Bring me a giraffe.	Bring me a elephant.	Bring me a polar bear.
		3	Bring me a yellow .	Bring me a blue stuffed toy.	Bring me a white stuffed toy.
		4	Bring me a giraffe.	Bring me a elephant.	Bring me a polar bear.
		5	Bring me a giraffe.	Bring me a elephant.	Bring me a polar bear.
		6	Bring me a yellow stuffed toy.	Bring me a elephant.	Bring me a white bear.
		7	Bring me a giraffe.	Bring me a elephant.	Bring me a polar bear.
8		1	Bring me a tupper.	Bring me a black rubber cup.	Bring me a vacuum cleaner.
		2	Bring me a green vessel.	Bring me a black one.	Bring me a vacuum cleaner.
		3	Bring me a tupper.	Bring me a black one.	Bring me a vacuum cleaner.
		4	Bring me a tupper.	Bring me a black one.	Bring me a vacuum cleaner.
		5	Bring me a green tupper.	Bring me a rubber cup.	Bring me a vacuum cleaner.
		6	Bring me a green vessel.	Bring me a rubber cup.	Bring me a pink vacuum cleaner.
		7	Bring me a green one.	Bring me a black one.	Bring me a vacuum cleaner.

Scene index	Scene	Participant	Object 1	Object 2	Object 3
9		1	Bring me a carton of coffee.	Bring me a wet tissue.	Bring me a soccer ball.
		2	Bring me a carton of coffee.	Bring me a green vessel.	Bring me a soccer ball.
		3	Bring me a carton of coffee.	Bring me a wet tissue.	Bring me a soccer ball.
		4	Bring me a carton.	Bring me a tissue.	Bring me a soccer ball.
		5	Bring me a carton.	Bring me a wet tissue.	Bring me a soccer ball.
		6	Bring me a drink.	Bring me a wet tissue.	Bring me a soccer ball.
		7	Bring me a carton of coffee.	Bring me a wet tissue.	Bring me a soccer ball.
10		1	Bring me a ship.	Bring me a blue cap.	Bring me a spray.
		2	Bring me a ship.	Bring me a cap.	Bring me a spray.
		3	Bring me a ship.	Bring me a cap.	Bring me a blue vessel.
		4	Bring me a ship.	Bring me a cap.	Bring me a white spray.
		5	Bring me a ship.	Bring me a blue cap.	Bring me a spray.
		6	Bring me a ship.	Bring me a blue cap.	Bring me a deodorizing spray.
		7	Bring me a ship.	Bring me a cap.	Bring me a spray.
11		1	Bring me a black camera.	Bring me a black flashlight.	Bring me an electric fan.
		2	Bring me a camera.	Bring me a flashlight.	Bring me an electric fan.
		3	Bring me a video camera.	Bring me a flashlight.	Bring me an electric fan.
		4	Bring me a digital camera.	Bring me a flashlight.	Bring me an electric fan.
		5	Bring me a gray camera.	Bring me a black one.	Bring me an electric fan.
		6	Bring me a video camera.	Bring me a light.	Bring me an electric fan.
		7	Bring me a camera.	Bring me a light.	Bring me an electric fan.
12		1	Bring me a Kyorochan.	Bring me an iron.	Bring me a dolphin.
		2	Bring me a Kyorochan.	Bring me a pink one.	Bring me a dolphin.
		3	Bring me a brown one.	Bring me an iron.	Bring me a blue stuffed toy.
		4	Bring me a Kyorochan.	Bring me an iron.	Bring me a dolphin.
		5	Bring me a Kyorochan.	Bring me a pink iron.	Bring me a dolphin.
		6	Bring me a Kyorochan.	Bring me an iron.	Bring me a stuffed toy.
		7	Bring me a brown one.	Bring me an iron.	Bring me a dolphin.
13		1	Bring me a brown one.	Bring me a picture frame.	Bring me a can of cola.
		2	Bring me a ketchup.	Bring me a picture frame.	Bring me a can of cola.
		3	Bring me a ketchup.	Bring me a picture frame.	Bring me a red can.
		4	Bring me a ketchup.	Bring me a picture frame.	Bring me a red can.
		5	Bring me a ketchup.	Bring me a picture frame.	Bring me a can.
		6	Bring me a ketchup.	Bring me a picture frame.	Bring me a can of cola.
		7	Bring me a ketchup.	Bring me a mirror.	Bring me a can of cola.
14		1	Bring me a pan.	Bring me a mirror.	Bring me a pink one.
		2	Bring me a black one.	Bring me a mirror.	Bring me a pink one.
		3	Bring me a pan.	Bring me a purple one.	Bring me a microwave oven.
		4	Bring me a pan.	Bring me a mirror.	Bring me a microwave oven.
		5	Bring me a black one.	Bring me a mirror.	Bring me a pink one.
		6	Bring me a black pan.	Bring me a purple mirror.	Bring me a pink one.
		7	Bring me a pan.	Bring me a mirror.	Bring me a pink stuffed toy.
15		1	Bring me a dustpan.	Bring me a pot.	Bring me a green turtle.
		2	Bring me a dustpan.	Bring me a orange one.	Bring me a turtle.
		3	Bring me a dustpan.	Bring me a watering pot.	Bring me a green stuffed toy.
		4	Bring me a dustpan.	Bring me a watering pot.	Bring me a turtle.
		5	Bring me a dustpan.	Bring me a watering pot.	Bring me a turtle.
		6	Bring me a dustpan.	Bring me a orange watering pot.	Bring me a turtle.
		7	Bring me a dustpan.	Bring me a watering pot.	Bring me a turtle.

謝辞

神戸大学大学院工学研究科において，本研究を行う機会を与えてくださり，また本研究を進めるにあたり，有益なご指導，ご教示をいただき本研究の主査をして頂いた神戸大学システム情報学研究科　有木康雄　教授に深く感謝しております．博士後期課程 3 年目の秋に，転専攻した際，私を受け入れてくださったご恩は一生忘れません．解ける問題ではなく，解けない問題に立ち向かう勇気を持ち続けることの大切さ，既存研究への敬意を持って，サーベイをし続ける大切さ，繰り返し何度でも反芻して基礎力を強化していくことの大切さを，有木康雄教授の背中から学びました．

お忙しい中，貴重な時間を割いて本論文の副査をして頂いた神戸大学大学院システム情報学研究科 大川剛直 教授，玉置久 教授に感謝致します．

また，本研究を遂行するに当たり，多くのご支援と，ご指導をいただきました，(株) ホンダリサーチインスティチュートジャパン　中野幹生　様に深く感謝をしております．本研究の共同研究を通じて，論文の意味，論文の書き方，共著としての在り方を御教示いただいたことが，私の研究に対する姿勢の転機となりました．中野幹生様に，学位論文のひとつである論文のうちの 1 本を 2 年かけてご指導していただいたことが私の財産です．

転専攻直後，本研究テーマに関して全く知識がなかった私に，この研究テーマの面白さ，難しさを教えていただいた，岡山県立大学情報工学部　岩橋直人　教授に深く感謝しております．転専攻直後のミーティングで，私の転専攻理由をお話したときに，「その転専攻は優秀な証だ．逆に安心した．しっかり研究を進めていきましょう．」と言っていただいたこと，涙が出るほど嬉しかったです．あの一言で，走り出せました．

教員でありながら，女性研究者として同志であり，親友でもある神戸大学自然科学系先端融合研究環　榎並直子　助教に深く感謝しております．私の精神状態を支えてくださったおかげで，よい精神状態を保って学位論文に向かうことができました．

他研究室でありながら，度々研究のオフサイトミーティングをしてくださった神戸大学大学院システム情報学研究科 江口浩二 准教授に深く感謝しております．視点を変えて私の研究を見つめなおす機会をいただいたことで，私自身が見えていなかった問題を沢山見つけることができました．

また，日頃より研究生活を共にした神戸大学工学部情報知能工学科有木研究室の滝口哲也　准教授をはじめとする皆様に感謝致します．特に，人が成長するには，何よりも素直さが大切であるということを体現してくれた百合本瑞規さん，私の日頃の研究生活を支え続けてくれた西村仁志さんに深く感謝したします．

最後に，物心ついたときから，何をしても紆余曲折する私を支え，辛抱強く支援してくださった私の家族に深い感謝の意を表して謝辞と致します．

参考文献

[1] R. Descartes, Principles of philosophy, vol.24, Springer, 1984.

[2] J.W. Carey, Communication as culture, revised edition: Essays on media and society, Routledge, 2008.

[3] S. Harnad, “The symbol grounding problem,” Physica D: Nonlinear Phenomena, vol.42, no.1, pp.335–346, 1990.

[4] D.K. Roy and A.P. Pentland, “Learning words from sights and sounds: A computational model,” Cognitive science, vol.26, no.1, pp.113–146, 2002.

[5] L. Steels, Experiments in cultural language evolution, vol.3, John Benjamins Publishing, 2012.

[6] L. Steels and F. Kaplan, “Aibo’s first words: The social learning of language and meaning,” Evolution of communication, vol.4, no.1, pp.3–32, 2002.

[7] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, and N. Iwahashi, “Autonomous acquisition of multimodal information for online object concept formation by robots,” Proceedings of 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)IEEE, pp.1540–1547 2011.

[8] C. Yu and D.H. Ballard, “On the integration of grounding language and learning objects,” Proceedings of 2004 AAAI Conference on Artificial Intelligence, vol.4AAAI, pp.488–493 2004.

[9] N. Iwahashi, “Interactive learning of spoken words and their meanings through an audio-visual interface,” IEICE Transactions on Information and Systems, vol.91, no.2, pp.312–321, 2008.

[10] D.O. Johnson and A. Agah, “Human robot interaction through semantic integration of multiple modalities, dialog management, and contexts,” International Journal of Social Robotics, vol.1, no.4, pp.283–305, 2009.

[11] D. Roy, “New horizons in the study of child language acquisition,” International Speech Communication Association, pp.6–10, 2009.

[12] M. Nakano, N. Iwahashi, T. Nagai, T. Sumii, X. Zuo, R. Taguchi, T. Nose, A. Mizutani, T. Nakamura, M. Attamim, et al., "Grounding new words on the physical world in multi-domain human-robot dialogues.," AAAI Fall Symposium: Dialog with Robots, pp.74–79, 2010.

[13] H. Holzapfel, D. Neubig, and A. Waibel, "A dialogue approach to learning object descriptions and semantic categories," Robotics and Autonomous Systems, vol.56, no.11, pp.1004–1013, 2008.

[14] D. Skocaj, M. Janicek, M. Kristan, G.-J.M. Kruijff, A. Leonardis, P. Lison, A. Vrecko, and M. Zillich, "A basic cognitive system for interactive continuous learning of visual concepts," ICRA 2010 Workshop ICAIR-Interactive Communication for Autonomous Intelligent Robots, pp.30–36, 2010.

[15] F. Lömker and G. Sagerer, "A multimodal system for object learning," Pattern Recognition, pp.490–497, Springer, 2002.

[16] Y. Sun, L. Bo, and D. Fox, "Attribute based object identification," Proceedings of 2013 IEEE International Conference on the Robotics and Automation (ICRA)IEEE, pp.2096–2103 2013.

[17] R. Fang, C. Liu, and J.Y. Chai, "Integrating word acquisition and referential grounding towards physical world interaction," Proceedings of 2012 ACM International Conference on Multimodal Interaction (ICMI)ACM, pp.109–116 2012.

[18] Y. Ozasa, Y. Ariki, M. Nakano, and N. Iwahashi, "Disambiguation in unknown object detection by integrating image and speech recognition confidences," Proceedings of 2012 Asian Conference on Computer Vision (ACCV), pp.85–96, Springer, 2013.

[19] H. Nishimura, Y. Ozasa, Y. Ariki, and M. Nakano, "Object recognition by integrated information using web images," Proceedings of 2013 IAPR Asian Conference on Pattern Recognition (ACPR)IEEE, pp.657–661 2013.

[20] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," Proceedings of 2013 Computer Vision and Pattern Recognition (CVPR)IEEE, pp.3166–3173 2013.

[21] S. Nagaya, S. Seki, and R. Oka, "A theoretical consideration of pattern space trajectory for gesture spotting recognition," Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference onIEEE, pp.72–77 1996.

[22] G. Kress, Multimodality: A social semiotic approach to contemporary communication, Routledge, 2009.

[23] A. Jaimes and N. Sebe, “Multimodal human–computer interaction: A survey,” Computer vision and image understanding, vol.108, no.1, pp.116–134, 2007.

[24] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, “Multimodality image registration by maximization of mutual information,” Medical Imaging, IEEE Transactions on, vol.16, no.2, pp.187–198, 1997.

[25] H. Yin and S. Li, “Multimodal image fusion with joint sparsity model,” Optical Engineering, vol.50, no.6, pp.067007–067007, 2011.

[26] S. Oviatt and P. Cohen, “Perceptual user interfaces: multimodal interfaces that process what comes naturally,” Communications of the ACM, vol.43, no.3, pp.45–53, 2000.

[27] 経済産業省，“ロボット革命実現会議”. http://www.kantei.go.jp/jp/singi/robot/.

[28] 渡辺慧，認識とパタン，第 36 巻，岩波書店，1978.

[29] X.D. Huang, Y. Ariki, and M.A. Jack, Hidden Markov models for speech recognition, vol.2004, Edinburgh university press Edinburgh, 1990.

[30] P.M. Roth and M. Winter, “Survey of appearance-based methods for object recognition,” Inst. for Computer Graphics and Vision, Graz University of Technology, Austria, Technical Report ICGTR0108 (ICG-TR-01/08), pp.1–8, 2008.

[31] D. Wu and L. Shao, “Multimodal dynamic networks for gesture recognition,” Proceedings of the ACM International Conference on MultimediaACM, pp.945–948 2014.

[32] S. Ouellet, F. Grondin, F. Leconte, and F. Michaud, “Multimodal biometric identification system for mobile robots combining human metrology to face recognition and speaker identification,” Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium onIEEE, pp.323–328 2014.

[33] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.31, no.1, pp.39–58, 2009.

[34] M.H. Giard and F. Peronnet, “Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study,” Journal of cognitive neuroscience, vol.11, no.5, pp.473–490, 1999.

[35] A. Aldoma, F. Tombari, J. Prankl, A. Richtsfeld, L. Di Stefano, and M. Vincze, “Multimodal cue integration through hypotheses verification for rgb-d object recognition and 6dof pose estimation,” Robotics and Automation (ICRA), 2013

IEEE International Conference onIEEE, pp.2104–2111 2013.

[36] J.R. Searle, “Minds, brains, and programs,” Behavioral and brain sciences, vol.3, no.03, pp.417–424, 1980.

[37] L. Steels, “The symbol grounding problem has been solved. so what’ s next,” Symbols and embodiment: Debates on meaning and cognition, pp.223–244, 2008.

[38] B. Graf, M. Hans, and R.D. Schraft, “Care-o-bot II : Development of a next generation robotic home assistant,” Autonomous robots, vol.16, no.2, pp.193–205, 2004.

[39] K. Yamazaki, R. Ueda, S. Nozawa, M. Kojima, K. Okada, K. Matsumoto, M. Ishikawa, I. Shimoyama, and M. Inaba, “Home-assistant robot for an aging society,” Proceedings of the IEEE, vol.100, no.8, pp.2429–2441, 2012.

[40] R. Ros, S. Lemaignan, E.A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, “Which one? grounding the referent based on efficient human-robot interaction,” Proceedings of 2010 IEEE International Symposium on Robot and Human Interactive Communication(RO-MAN)IEEE, pp.570–575 2010.

[41] A. Lee, K. Shikano, and T. Kawahara, “Real-time word confidence scoring using local posterior probabilities on tree trellis search,” Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on, vol.1IEEE, pp.I–793 2004.

[42] Julius, “Open source large vocabulary csr engine Julius,” `http://julius.sourceforge.jp/`.

[43] H. Jiang, “Confidence measures for speech recognition: A survey,” Speech communication, vol.45, no.4, pp.455–470, 2005.

[44] L. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” Proceedings of the IEEE, vol.77, no.2, pp.257–286, 1989.

[45] T. Kurita, “Interactive weighted least squares algorithms for neural networks classifiers,” Proceedings of 1992 Workshop on Algorithmic Learning TheoryAlgorithmic Learning Theory, pp.77–86 1992.

[46] C.M. Bishop, et al., Pattern recognition and machine learning, vol.1, Springer New York, 2006.

[47] JAGAT 編集委員会，眼色光より優れた色再現を求めて，第 1 巻，JAGAT，2007.

[48] P.G.C. Ltd., “Point grey,” `http://ww2.ptgrey.com/`.

[49] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, “Recent progress of open-source LVCSR engine Julius and Japanese model repository,” Proceedings of 2004 International Conference on Spoken Language Processing (ICSLP),

pp.3069–3072, 2004.

[50] CG-ARTS 協会編集委員会，デジタル画像処理，第 1 巻，CG-ARTS 協会，2006.

[51] M.W. Schwarz, W.B. Cowan, and J.C. Beatty, “An experimental comparison of rgb, yiq, lab, hsv, and opponent color models,” ACM Transactions on Graphics (TOG), vol.6, no.2, pp.123–158, 1987.

[52] 柳井啓司，“Confidence measures for speech recognition: A survey,” コンピュータビジョン・イメージメディア, vol.48, no.SIG16(CVIM19), pp.1–24, 2007.

[53] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference onIEEE, pp.3360–3367 2010.

[54] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, pp.1097–1105, 2012.

[55] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” arXiv, vol.1312, no.6229, pp.1–16, 2013.

[56] D.G. Lowe, “Object recognition from local scale-invariant features,” Computer vision, 1999. The proceedings of the seventh IEEE international conference on, vol.2IEEE, pp.1150–1157 1999.

[57] C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol.20, no.3, pp.273–297, 1995.

[58] X. Zuo, N. Iwahashi, K. Funakoshi, M. Nakano, R. Taguchi, S. Matsuda, K. Sugiura, and N. Oka, “Detecting robot-directed speech by situated understanding in physical interaction,” Information and Media Technologies, vol.5, no.4, pp.1314–1326, 2010.

[59] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., “Imagenet large scale visual recognition challenge,” arXiv, vol.1409, no.0575, pp.1–37, 2014.

[60] M. Cowlishaw, “Fundamental requirements for picture presentation,” Proceedings of 1985 Information Display, vol.26Society of Information Display, pp.101–107 1985.

[61] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” Transactions on pattern analysis and machine intelligence, vol.20, no.11, pp.1254–1259, 1998.

[62] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, “Ranking

on data manifolds," Advances in neural information processing systems, vol.16, pp.169–176, 2004.

[63] S. Lu, V. Mahadevan, and N. Vasconcelos, "Learning optimal seeds for diffusion-based salient object detection," Proceedings of 2013 Computer Vision and Pattern Recognition (CVPR), pp.2790–2797, 2013.

[64] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S.Z. Li, "Salient color names for person re-identification," Proceedings of 2012 European Conference on Computer Vision(ECCV), pp.536–551, Springer, 2014.

[65] R. De Maesschalck, D. Jouan-Rimbaud, and D.L. Massart, "The mahalanobis distance," Chemometrics and intelligent laboratory systems, vol.50, no.1, pp.1–18, 2000.

[66] C.K.J.B. S. Watanabe, P.F. Lambert and R. Walker, "Evaluation and selection of variables in pattern recognition," Computer and Information Science, vol.2, pp.91–122, 1967.

[67] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," Transactions on pattern analysis and machine intelligence, vol.34, no.11, pp.2274–2282, 2012.

[68] M. Nishida and Y. Ariki, "Speaker verification by integrating dynamic and static features using subspace method.," INTERSPEECH, pp.1013–1016, 2000.

[69] S.-M. Park and K.-J. Kim, "Color recognition with compact color features," International Journal of Communication Systems, vol.25, no.6, pp.749–762, 2012.

[70] D. Barber, Bayesian reasoning and machine learning, Cambridge University Press, 2012.

[71] M.C. Ltd, "Kinect," `http://www.microsoft.com/en-us/kinectforwi\ \ndows/`.

[72] M. Sugiyama, H. Hachiya, M. Yamada, J. Simm, and H. Nam, "Least-squares probabilistic classifier: A computationally efficient alternative to kernel logistic regression," Proceedings of International Workshop on Statistical Machine Learning for Speech Processing (IWSML2012), Kyoto, JapanCiteseer, pp.1–10 2012.

[73] X.-J. Wang, L. Zhang, M. Liu, Y. Li, and W.-Y. Ma, "Arista-image search to annotation on billions of web photos," Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference onIEEE, pp.2987–2994 2010.

[74] ImageSpider, "Imagespider". `http://kurima.sakura.ne.jp/sb/log/eid123.html`.

[75] ImageGeter, "Imagegeter". http://uwa.potetihouse.com/soft/imagegeter.html.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," The Journal of Machine Learning Research, vol.12, pp.2825–2830, 2011.

発表論文

論文（査読あり）

1. Y. Ozasa, M. Nakano, Y. Ariki, “Discriminating Unknown Objects from Known Objects Using Image and Speech Information,” IEICE Transactions on Information and Systems, Vol.E98-D,No.3, 2015. (採録決定)

国際会議（査読あり）

1. Y. Ozasa, N. Iwahashi, T. Takiguchi, Y. Ariki, and M. Nakano, “Integrated multimodal information for detection of unknown objects and unknown names,” RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP), pp.631-634, 2012.
2. Y. Ozasa, Y. Ariki, M. Nakano, and N. Iwahashi, “Disambiguation in unknown object detection by integrating image and speech recognition confidences,” Proc. Asian Conference on Computer Vision (ACCV), pp.85-96, 2012.
3. H. Nishimura, Y. Ozasa, Y. Ariki, and M. Nakano, “Object Recognition by Integrated Information Using Web Images,” Proc. Asian Conference on Pattern Recognition (ACPR), pp.657-661, 2013.
4. H. Nishimura, Y. Ozasa, Y. Ariki, and M. Nakano, “Selection of Unknown Objects Specified by Speech Using Models Constructed from Web Images,” Proc. International Conference on Pattern Recognition (ICPR), pp.477-482, 2014.
5. H. Nishimura, Y. Ozasa, Y. Ariki, and M. Nakano, “Selection of an Object Requested by Speech Based on Generic Object Recognition.” Proc. Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction. ACM, 2014.
6. Y. Ozasa, N. Enami, Y. Ariki, “Color Saliency for Object Identification,” Proc. 21st Japan-Korea Joint Workshop on Frontiers of Computer Vision (FCV), 2015.

特許

1. M. Nakano, N. Iwahashi, Y. Ariki, Y. Ozasa, Y. Hori, R. Nakatani, “Object classification/recognition apparatus and method.” U.S. Patent Application 13/724,220.
2. M. Nakano, H. Nishimura, Y. Ozasa, Y. Ariki “Object recognition system and an object recognition method.” U.S. Patent Application 14/190,539.

口頭発表（査読なし）

1. Y. Ozasa, N. Iwahashi, T. Hori, R. Nakatani, Y. Ariki, and M. Nakano, “Detecting Unknown Objects and Unknown Names Using Multimodal Information,” 計測自動制御学会 システムインテグレーション部門講演会 (SI), pp.1629-1639, 2011.
2. 小篠　裕子, 有木康雄, 岩橋直人, 中野幹生, “マルチモーダル情報を用いた未知物体検知

手法,” 電子情報通信学会技術研究報告, パターン認識・メディア理解 (PRMU), vol.111, no.499, pp.247-252, 2012.

3. Y. Ozasa, Y. Ariki, N. Iwahashi, and M. Nakano, “Unknown Object Detection Using Multimodal Information Integrated by Kernel Logistic Regression,” 画像の認識・理解シンポジウム (MIRU), IS3-71, 2012.
4. 小篠　裕子, 中野幹生, 長谷川雄二, 中村友昭, 長井孝之, 岩橋直人, 有木康雄, “マルチモーダル情報を用いた未知物体学習のための未知物体判別手法,” 日本ロボット学会学術講演会 (RSJ), 4N1-2, 2012.
5. H. Nishimura, Y. Ozasa, Y. Ariki, and M. Nakano, “Object Recognition by Integrated Information Using Speech and Web Images,” 画像の認識・理解シンポジウム (MIRU), SS5-25, 2013.
6. 高柳陽平, 小篠　裕子, 榎並直子, 有木康雄, “人検出のための動的顕著性マップモデルの構築,” 電子情報通信学会技術研究報告, パターン認識・メディア理解 (PRMU), vol.112, no.385, pp.187-191, 2013.
7. S. Numano, N. Enami, Y. Ozasa, and Y. Ariki, “The Level of Skill Model for Piano Performance: Analyzing Gaze on Music Videos,” 画像の認識・理解シンポジウム (MIRU), SS2-34, 2014.
8. 西村仁志, 小篠　裕子, 有木康雄, 中野幹生, “Web 画像を用いた一般物体認識と指示発話の音声認識を統合した物体選択法,” 画像の認識・理解シンポジウム (MIRU), SS2-36, 2014.
9. 小篠　裕子, 榎並直子, 有木康雄, “物体特定のための顕著性,” 電子情報通信学会技術研究報告, パターン認識・メディア理解 (PRMU), vol.114, no.230, pp.19-24, 2014.
10. 小篠　裕子, 榎並直子, 有木康雄, “色属性による物体特定のための顕著性,” 電子情報通信学会技術研究報告, パターン認識・メディア理解 (PRMU), vol.114, no.356, pp.79-83, 2014.

受賞

1. 第 12 回計測自動制御学会システムインテグレーション部門講演会（SICE SI2011）優秀講演賞，2014，12．25.

奥付

神戸大学博士論文「マルチモーダル情報を用いた未知物体の認識とグラウンディングに関する研究」全 93 頁
提 出 日 2015 年 1 月 23 日
本博士論文が神戸大学機関リポジトリ Kernel にて掲載される場合、掲載登録日（公開日）はリポジトリの該当ページ上に掲載されます.

© 小篠　裕子
本論文の内容の一部あるいは全部を無断で複製・転載・翻訳することを禁じます.

本論文に基づく学位審査にあたっては、以下の参考論文とともに審査された.

OZASA Yuko, IWAHASHI Naoto, HORI Takahiro, NAKATANI Ryohei, ARIKI Yasuo, NAKANO Mikio, : Detecting Unknown Objects and Unknown Names Using Multimodal Information. 計測自動制御学会 システムインテグレーション部門講演会, 1629-1639, 2011.

OZASA Yuko, IWAHASHI Naoto, TAKIGUCHI Tetsuya, ARIKI Yasuo, NAKANO Mikio, : Integrated Multimodal Information for Detection of Unknown Objects and Unknown Names. RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing , 631-634, 2012.

小篠裕子, 有木康雄, 岩橋直人, 中野幹生：マルチモーダル情報を用いた未知物体検知手法. 電子情報通信学会技術研究報告, パターン認識・メディア理解, 111(499), 247-252, 2012.

OZASA Yuko, ARIKI Yasuo, IWAHASHI Naoto, NAKANO Mikio : Unknown Object Detection Using Multimodal Information Integrated by Kernel Logistic Regression. 画像の認識・理解シンポジウム, IS3-71, 2012.

小篠裕子, 中野幹生, 長谷川雄二, 中村友昭, 長井隆行, 岩橋直人, 有木康雄 : マルチモーダル情報を用いた未知物体学習のための未知物体判別手法. 日本ロボット学会学術講演会, 4N1-2, 2012.

西村仁志, 小篠裕子, 有木康雄, 中野幹生 : Web 画像を用いたマルチモーダル情報による物体認識. 電子情報通信学会総合大会, 2(2), 109, 2013.

NAKANO Mikio, IWAHASHI Naoto, ARIKI Yasuo, OZASA Yuko, HORI Takahiro, NAKATANI Ryohei : Object Classification/Recognition Apparatus and Method. U.S. Patent Application, 13/724, 220, 2013.

NISHIMURA Hitoshi, OZASA Yuko, ARIKI Yasuo, NAKANO Mikio : Object Recognition by Integrated Information Using Web Images. Proceedings of the Asian Conference on Pattern Recognition, 657-661, 2013.

NISHIMURA Hitoshi, OZASA Yuko, ARIKI Yasuo, NAKANO Mikio : Object Recognition by Integrated Information Using Speech and Web Images. 画像の認識・理解シンポジウム, SS5-25, 2013.

NAKANO Mikio, NISHIMURA Hitoshi, OZASA Yuko, ARIKI Yasuo : Object Recognition System and An Object Recognition Method. U.S. Patent Application, 14/190, 539, 2014.

西村仁志, 小篠裕子, 有木康雄, 中野幹生 : Web 画像を用いた一般物体認識と指示発話の音声認識を統合した物体選択法. 画像の認識・理解シンポジウム, SS2-36, 2014.

小篠裕子, 榎並直子, 有木康雄 : 物体特定のための顕著性. 電子情報通信学会技術研究報告, パターン認識・メディア理解, 114(230), 19-24, 2014.

小篠裕子, 榎並直子, 有木康雄 : 色属性による物体特定のための顕著性. 電子情報

通信学会技術研究報告, パターン認識・メディア理解, 114(356), 79-83, 2014.