論 文

Technical Papers

# 自己増殖型ニューラルネットワークを用いたヒューマノイドロボットの発達的言語獲得

# Developmental Word Acquisition through Self-Organized Incremental Neural Network with A Humanoid Robot

岡田 将吾
Shogo Okada
東京工業大学大学院総合理工学研究科知能システム科学専攻
Department of Computatinal Intelligence and Systems Science, Tokyo Institute of Thechnology
okada.s.ab@m.titech.ac.jp

賀 小淵
Xiaoyuan He
（同　上）

小島 量
Ryo Kojima
（同　上）

長谷川 修
Osamu Hasegawa
東京工業大学大学院理工学研究科像情報工学研究施設
Imaging Science and Engineering Lab.,Tokyo Insititute of Thecnology
hasegawa.o.aa@m.titech.ac.jp

**keywords:** robot, word grounding, language, mental models, interactive learning, incremental learning

**Summary**

This paper presents an unsupervised approach of integrating speech and visual information without using any prepared data(training data). The approach enables a humanoid robot, Incremental Knowledge Robot 1 (IKR1), to learn words' meanings. The approach is different from most existing approaches in that the robot learns online from audio-visual input, rather than from stationary data provided in advance. In addition, the robot is capable of incremental learning, which is considered to be indispensable to lifelong learning. A noise-robust self-organized incremental neural network(SOINN) is developed to represent the topological structure of unsupervised online data. We are also developing an active learning mechanism, called "desire for knowledge", to let the robot select the object for which it possesses the least information for subsequent learning. Experimental results show that the approach raises the efficiency of the learning process. Based on audio and visual data, we construct a mental model for the robot, which forms a basis for constructing IKR1's inner world and builds a bridge connecting the learned concepts with current and past scenes.

## 1. はじめに

人間は視覚情報を手がかりに，物事に関する概念を獲得することが可能である．この概念は先天的に与えられるものではなく，人間自らが経験的に獲得していくものであり，もとの視覚情報と密接な対応関係を持つ．

一方，コンピュータが扱う記号には，実世界の物事との繋がりは存在しない．そのため，コンピュータ内に記述されている記号と実世界に存在する物事との繋がりをどのように表現すれば良いかという問題 (シンボルグラウンディング問題) が指摘されている [Harnad 90]．そこで近年，物体に関連する発話 (音声情報) とその物体から抽出した画像情報とを統合し，その物体の意味 (概念) を獲得する手法の研究が盛んである．

本研究では，音声情報および画像情報を統合的に用いた，シンボルグラウンディング問題への 1 つのアプローチを提案する．さらに提案メカニズムをヒューマノイドロボットに搭載して稼動させる．本研究の目的をより明確にするために，1･1 節で本研究に関連する研究を紹介し，1･2 節において 1･1 節で述べた関連研究に対する本研究の新規性・有用性について議論する．

### 1･1 関 連 研 究

[Siskind 96, Siskind 00, Wachsmuth 00, Gorin 99, Regier 96, Oates 00, Roy 02, Iwahashi 03, Iwahashi 04, Yu 04] では，音声情報と視覚情報から言語の意味を獲得するモデルが提案されている．まず [Siskind 96] では，幼児が対面する単語の意味獲得問題を参考に，学習アルゴリズムが提案されている．しかしヒューマノイドロボットのような実システム上で，このアルゴリズムが

稼動可能かどうかの議論はなされていない．また，ここで提案されたアルゴリズムは文法に関する予備知識を必要とするため，予備知識がない場合に文法の獲得は不可能である．次に [Wachsmuth 00] では異なったレベルの表現を使用することで視覚情報と音声情報を統合する手法が提案されている．ここでは，異なったレベル間の関連性および，視覚情報と音声情報の関連性をモデル化する手法としてベイジアンネットワークが用いられている．[Regier 96] では独自のアルゴリズムにより空間的な概念 (例えば, in, on, through) が獲得されている．この手法では学習時に教師付き訓練データを必要とする．

[Roy 02] では視覚情報と音声情報の相互情報量を定義し，これを最大化するアルゴリズムが提案されている．この研究では，幼児と擁護者との自然なインタラクションの音声データから言語獲得が行われている．ここで視覚データには対象物体の静止画が用いられている．[Iwahashi 03] では，移動物体から得られる 2 次元動画像と，その動画像を表現する言語の組み合わせを入力として，動きの概念と文法の獲得を行う手法が提案されている．[Iwahashi 04] では能動学習と教師なし学習を用い，視覚概念と音声概念のペアを追加学習可能な方法が提案されている．

以上をまとめて [Wachsmuth 00, Roy 02, Iwahashi 03, Iwahashi 04] では，音声データ (言語データ) と画像データのどちらか，または両方が静的データである．このうち [Wachsmuth 00, Iwahashi 03] では，あらかじめ音声データと，それに対応する視覚データがセットとして与えられ，学習が行われる．[Roy 02] では，自然なインタラクションから得られた音声データが用いられているものの，それらを用いたオンライン学習および追加学習は行われていない．[Iwahashi 04] では，追加的単語獲得（追加学習）を可能としている．この手法では，単語音声概念が，音素を表す Hidden markov model(HMM) の結合によって，また視覚概念が Gaussian 関数によってモデル化されている．この音素を表す HMM のセットは，単語獲得を行う前に音声データを用いた教師なしバッチ学習によって行われており，完全なオンライン学習は実現されていない．

次に [Yu 04] では，発話された物体の名称 (音声データ) と物体に関する視覚情報をグラウンディング可能なマルチモーダル学習システムが提案されている．ここで各物体 (視覚特徴空間ではクラスタとして表現されている) は，発話より得られた言語的ラベルに基づき分類される． しかしこのシステムでは，視覚特徴空間に未知の物体が入力された場合，新しいクラスタ (未知の物体) と既存のクラスタを混同してしまうため，新しいクラスタを追加的に獲得すること (追加学習) が困難である．

[Steels 97, Steels 03, Steels 02, Vogt 05] では人間とロボットの相互作用に基づいて概念を獲得するのでなく，ロボット同士の相互作用を通じた概念の獲得が試みられている．

### 1･2 本研究の特性

本研究では，視聴覚情報から能動的・追加的に概念を獲得 (学習) 可能なシステムを提案し，これをヒューマノイドロボットに搭載し稼動させる．本研究で提案する手法は 1･1 節で述べた従来手法に比べて，以下の 3 点で異なっている．

(1) 提案手法では，Self Organizing Incremental Neural Network(SOINN)[Shen 06] という自己増殖型ニューラルネットワークを用いることにより追加学習 *incremental learning* ([Elman 93, Thrun 95]) が実現されている．追加学習においては，どれだけ既学習の情報を温存して新しい情報を記憶するか (安定性/柔軟性ジレンマ) という基本的問題が存在する．SOINN はこの問題に対し，既学習のデータを忘却せず新規のデータを記憶することが可能である．また SOINN は教師ラベル付き訓練データを必要としない．総じて SOINN は，教師ラベル付き訓練データを必要とする手法 ([Wachsmuth 00, Regier 96]) や追加学習が困難であるである手法 ([Yu 04]) と異なり，教師なし追加学習を可能とする．加えて SOINN はノイズを含む入力データからノイズを除去することも可能である．

(2) 提案システムでは，訓練データを事前に与えてバッチ学習を行う必要がなく，事前知識のない状態からオンラインで追加的に視覚・音声データを学習 (教師なしオンライン学習) 可能である他，視覚情報と音声情報の対応付けも完全にオンラインで行われる．

(3) 提案システムを搭載したロボットは能動的に学習を行うことが出来る．

提案システムでは，学習対象の物体に関して十分な情報が得られたかどうかを判断し，足りない情報をパートナー (教示者) に要求する．具体的には，提案システムである物体に関する情報が足りないと判断された場合，ロボットはその物体を指し示す．ロボットの指示に従い，パートナーはその物体に関する情報をロボットに与える．結果的に，ロボットは物体に関する情報を能動的に獲得することが可能である．これに対し，既存のシステムやロボットは受身に情報を学習するのみで，能動的に情報を得ることは出来ない．

提案アルゴリズムを搭載したヒューマノイドロボットが以上の 3 点を実現することを，複数の概念 (例えば，赤い，丸い) を持つ，72 種類の物体を用いた実験により検証する．

## 2. 提案システム

### 2･1 ロボット環境

本研究で用いたヒューマノイドロボットは，IKR1 と呼ばれ上半身のみを持つ．IKR1 は自由度 2 の可動式ヘッド，およびステレオカメラで構成された目を持つ (図 1).

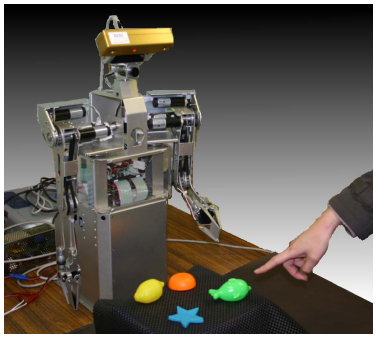


**図 1** 本研究で用いたロボット (IKR1)

**図 2** IKR1 のワークスペースと学習対象例

また IKR1 はワークスペース全体を探索することが可能である．ここでワークスペースとは IKR1 の前に設置されたテーブルの一部であり，IKR1 が物をつかむことが可能な領域とする．IKR1 は 2 本のグリッパ (腕) を持ち (各 5 自由度)，このグリッパを用いて物体を操作する．IKR1 のワークスペースは様々な物体が無作為に置かれた状態となっている．ここで物体はおもちゃ，文房具，紙などであり，これらの物体は色，形状，サイズが異なる (図 2)． パートナー (教示者) は，テーブルを挟んで IKR1 の対面に座る． 最初に，IKR1 はワークスペースにある物体の概念に関する事前知識は保持しない．ここで概念とは物体の名前，色，および形状の 3 つとする．各々の概念は (1) 発話された単語 (音声情報)，(2) 視覚情報，の 2 種類から構成される．

まず学習を始める前に，IKR1 はメンタルモデルを初期化し，ワークスペースに注意を向ける．ここでメンタルモデルとは，IKR1 が周りの環境を理解するために構築された IKR1 の内部モデルである (メンタルモデルの詳細は 3 章で述べる)．次にパートナーはワークスペース内の物体を指差し，その物体についての情報を発話する．発話された音声情報は，音声処理モジュールに受け渡される．この後に，IKR1 は単語テンプレートと入力音声データを DTW(Dynamic Time Warping) により比較し，入力音声データが既存の音声クラスに属するか未知の音声クラスかを識別する．未知であると識別した場合，新しい音声クラステンプレートが作成される．

視覚情報はステレオカメラより得られる．得られた視覚情報から画像処理によりノイズが除去され，物体の形状特徴および色特徴が抽出される．

IKR1 はこれらの音声と視覚の特徴量を統合的に用いて，言葉の意味 (概念) を学習・獲得する．

### 2·2 提案システムの処理過程

提案システムは 5 つの機能を持つモジュール群により構成される．各モジュール間の関係性を図 3 に示す．図 3 において太い円で囲まれた部分が各モジュールを示す．各モジュールは，画像処理モジュール (Image Processing)，音声処理モジュール (Speech Processing)，SOINN モジュール (SOINN)，ロボット制御モジュール (Robot Control)，そしてメンタルモデル (Mental Model) の 5 種類である．

ここでシステムの処理の流れについて説明する．まずパートナーはロボットの前に数種類の物体を置く．ロボットはステレオカメラを用いて，視覚情報を画像処理モジュールに送り，物体の切り出しおよび特徴抽出を行う．次に，制御モジュールからはロボットの頭の水平角度と垂直角度 (Robot's Status) が ，また画像処理モジュールからは物体に関する情報 (色，形状，大きさ，位置座標)(Current Scene Information) がメンタルモデルに送信される．メンタルモデルでは，カメラから得られた画像上での物体の位置座標がロボットの絶対位置座標に変換される．ここで物体の絶対位置座標 $(x', y', z')$ は画像から得られる物体の位置座標 $(x, y, z)$ を用いて，オイラー角の公式より以下の式で計算される．

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\varphi & 0 & -\sin\varphi \\ 0 & 1 & 0 \\ \sin\varphi & 0 & \cos\varphi \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (1)$$

式 (1) で $\theta$,$\varphi$ はそれぞれ x 座標のオイラー角，y 座標のオイラー角を示す．メンタルモデルは，ロボットの視界に映った物体の情報 (Current Scene Information)，およびロボットの姿勢に関する情報 (Robot's Status) を常に保持する．ロボットはこのメンタルモデルを持つことで，環境の変化に適応出来る．

一方，画像処理モジュールで処理が行われた後に，物体に関して 3 次元の色ベクトルと 8 次元の形状ベクトル，及び色ベクトルと形状ベクトルを併せた 11 次元のベクトルの計 3 種類のベクトルが抽出される (画像処理の詳細は 2·3 節で述べる)．これらのベクトルは，SOINN を用いてクラスタに分類される．SOINN ではニューロが自己増殖しながら入力ベクトルの分布が近似され，入力ベクトル群が分類される．また SOINN では入力データのオンライン追加学習が可能である (SOINN の詳細は 2·5 節で述べる)．この SOINN の機能により，入力画像データは追加的に獲得される．

次にパートナーはロボットと向き合って特定の物体に

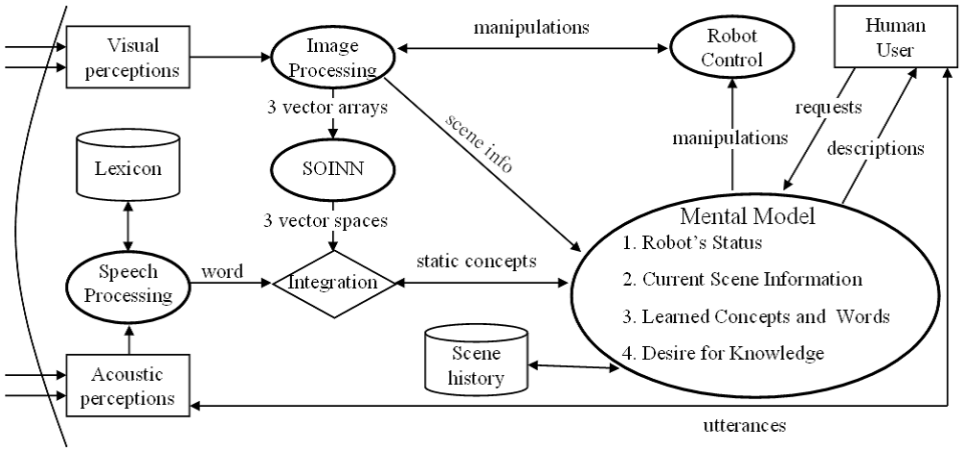



**図 3** 提案システムの概要 (各モジュール間の関係性を示す．)

対して音声を 1 つずつ発話していく．音声は DTW 法により分類される (音声処理の詳細は 2･4 節で述べる)．DTW 法において閾値設定を行うことで，テンプレートと入力音声の距離が閾値より大きい場合は，テンプレートとは異なる単語であると認識し，その入力音声を新規のテンプレートとして構成する．この手法を用いることで，簡易的に音声の追加学習を可能とした．最終的に，分類された画像データと分類された音声データが統合され，概念が獲得される (音声と画像の統合および概念の形成についての詳細は 2･6 節で述べる)．総じて提案手法では，別個に入力される音声データと視覚データを学習し，これらを関連付けることで追加的に概念を獲得する．

加えて，ロボットはパートナーが発話する単語を一方的に聞いて学習 (受動的な学習) するだけではない．ロボットは，ある物体に対して音声が入力された際に，ワークスペース内にある物体全ての既知度を算出することで，既知度の低い物体を探索する．探索後にロボットはグリッパで指し示すことにより，人間にその物体に関する情報を要求すること (能動的な学習) が可能である．本研究では，以上のようなロボットが能動的に情報を要求する機能を“ 知識欲 ”と定義する．このような“ 知識欲 ”を持ったロボットがパートナーと相互的に学習を進めることにより，円滑に効率よく概念の獲得が行われる (知識欲に関する詳細は 2･7 節で述べる)．

### 2･3 画 像 処 理

提案システムではステレオカメラ (2 つの CCD カメラ) を用いて画像処理を行う．ここで図 4 に実際にカメラから得られた画像例を示す．最初に，画像処理を行って距離画像を算出する (図 5(a))．距離について閾値を設けた後，この閾値を用いて物体を抽出する．物体を抽出後，物体ごとにラベルが付けられ，それらの物体のピクセルの数を算出する．ここでピクセル数が 100 以下の物体はノイズであると見なす．この閾値は実験的に決定した．これらの処理を行った後に得られた画像を図 5(b) に示す．物体として抽出された画像の領域に関して，その領域を構成する全画素の RGB 値の平均値を算出する．この RGB の平均値を色特徴ベクトル (**Color:** 3 次元) とする．物体の形状特徴ベクトル (**Shape:** 8 次元) は次のように算出する．

(1) 物体上の点で最も重心から離れている点を求め，重心からその点までの距離を $r$ とする．

(2) 図 6 に示すように，物体の重心を中心として半径 $\frac{i \times r}{12} (1 \leq i \leq 12)$ の同心円を描く．

(3) 中心から順々に，隣り合う 2 つの同心円の間の領域 $P_j$ の面積 $S_j$(内側から $j$ 番目の円と $j+1$ 番目の円に挟まれた部分) を算出する．図 6 において，$S_j$ は赤い部分を示す．

(4) 領域 $P_j$(面積 $S_j$) の中で物体が占める領域の面積 $T_j$ を求める．図 6 において，面積 $T_j$ は物体 (水色部分) と $S_j$(赤色部分) が重なった領域 (青色部分) の面積である．

(5) 形状特徴は次式から算出される．

$$\mathbf{Shape}[j] = \frac{T_{j+3}}{S_{j+3}} (1 \leq j \leq 8) \qquad (2)$$

ここで，中心に近い同心円 ($i = 1, 2, 3$) では $\mathbf{Shape}[i] \cong 1.0$ となることから， $i = 1, 2, 3$ についての形状特徴は用いなかった．この手法では，位置の変化と面内の回転

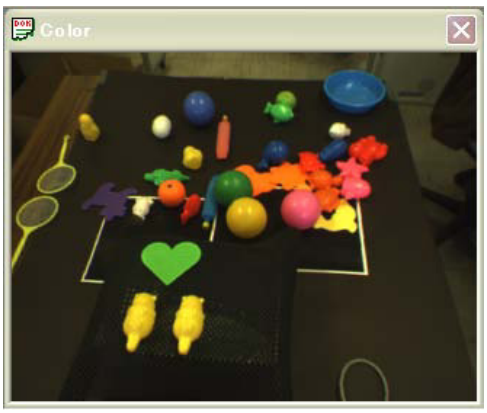



**図 4** ロボットに搭載したカメラから得られたカラー画像例


(a) (b)

**図 5** (a) ステレオカメラから得られる距離画像，(b) 閾値処理を行った後に物体が抽出された様子

に対して不変的な特徴を抽出しており，チェインコード法よりも良い性能が得られた．結果的に対象物から抽出された色，形状特徴は以下のように表される (**Object:** 11 次元).

$$\mathbf{Object} = \begin{bmatrix} \frac{1}{256}\mathbf{Color} \\ \mathbf{Shape} \end{bmatrix} \tag{3}$$

パートナーは教示したい物体をロボットに理解させるために，物体を指し示す．ここで人間の指と物体を区別するため，予め人間の指に類似した細長い肌色の物体を学習させる．またパートナーが指し示すものが，ワークスペース上の物体かどうかを認識するために，指の位置と物体の中心の間の距離が用いられる．

### 2·4 音声処理 (認識・生成)

提案システムでは，単語の音声処理に Speech Signal Processing Toolkit (SPTK) [Imai 02] を用いた．また音声データの識別には，時系列パターン同士の距離を算出するための手法である Dynamic Time Warping (DTW)

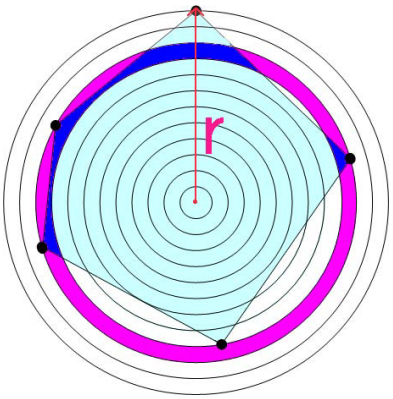



**図 6** 形状特徴 (ベクトル要素 **Shape**[6]) の算出例 (**Shape**[6] = $S1/(S1+S2)$ とした場合，$S1$ は青い部分の面積，$S2$ は赤い部分の面積を示す．)

[Myers 81] を用いた．本研究の目的より，ロボットは事前知識なしの状態から，オンラインで音声情報を受け取り，音声処理を行う必要がある．したがって提案システムでは，大量の訓練データを用いてバッチ学習を行う HMM を用いず，DTW を用いた．

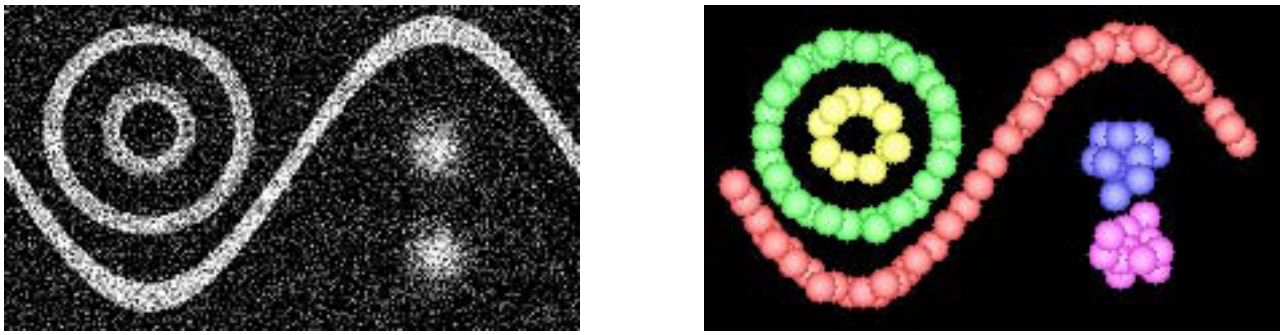
次に音声の特徴抽出手法について説明する．まずパートナーがマイクで単語を発話する．発話された単語から，1 フレームにつき 16 次元のメルケプストラム係数特徴を抽出する．このメルケプストラム係数特徴を用いて単語間の距離を算出する場合，特徴量の次元数が大きいため単語同士の比較が困難となる．これより，本研究ではベクトル量子化法 (vector quantization (VQ))[Gersho 92] を用いて次元の削減を行う．次に DTW で用いる累積距離 $D(i,j)$ と各フレーム間の距離 $d(i,j)$ を定義する．ここで $i$,$j$ はそれぞれ比較対象の 2 つの音声入力パターンの $i$ 番目および $j$ 番目のフレームである．$D(i,j)$ の算出は以下の漸化式によって求める．

$$D(i,j) = min\left\{ \begin{matrix} D(i-1,j-1) \\ D(i-1,j) \\ D(i,j-1) \end{matrix} \right\} + d(i,j) \tag{4}$$

$$D(i,j) = min\left\{ \begin{matrix} D(i-1,j-1) + 2d(i,j) \\ D(i-1,j) + d(i,j) \\ D(i,j-1) + d(i,j) \end{matrix} \right\} \tag{5}$$

漸化式 (4) では，対角，垂直，水平方向の全ての重みは等しい．これに対し漸化式 (5) では，対角方向の重みを，垂直・水平方向の重みの 2 倍としている．この 2 つの漸化式による識別性能を予備実験により検証したところ，漸化式 (5) を用いた方が識別精度が高いことを確認した．これより提案システムでは音声データの識別に漸化式 (5) を用いた．

音声処理部では，まずパートナーがある単語を発話すると，その発話から得られる音声データがその単語のテンプレートとなる．以後このテンプレートと入力単語を



(a) (b)

**図 7** (a)2 次元の人工データセット (データセットは 2 つのガウス分布，2 つの同心円及び *Sin* 曲線の合計 5 つのクラスによって構成されている．またデータセットにはノイズが含まれている．)

(b)SOINN に (a) の人工データを入力した結果 (入力データに含まれるノイズは削除され，入力データのクラスタ数とそのトポロジーが正しく抽出されている．)

照合することにより認識を行う．ここでテンプレートと入力音声の距離に対し，ある閾値を事前に定義した．テンプレートと入力音声の距離がこの閾値より小さい場合，その入力音声はそのテンプレートと同じ単語であると認識する．逆にテンプレートと入力音声の距離が閾値より大きい場合は，テンプレートとは異なる単語であると認識し，その入力音声を新規のテンプレートとして構成する．閾値は予備実験の結果より 6875 とした．

### 2･5 SOINN を用いた教師なし学習

#### § 1 SOINN の特徴

Self Organizing Incremental Neural Network (SOINN) [Shen 06] は Growing Neural Gas(GNG)[Fritzke 95] を拡張した自己増殖型ニューラルネットワークと呼ばれる教師なし追加学習手法である．SOINN は連続的にオンラインで入力されるデータを，ニューロを自己増殖することで追加的に学習可能である．SOINN を用いることによる主要な利点を，以下にまとめる．

(1) 過去に学習したクラスタを破綻させずに，新規に入力される未知クラスのデータを追加的に学習し，新規のクラスタを構築することが出来る
(2) 入力データに含まれるノイズを除去することが可能である
(3) 追加的に入力される教師ラベルなしデータの位相構造を表現することが可能である
(4) 入力データを近似するために用いる，ノード (ニューロ) 数や初期値を事前に設定する必要がない

以上の 4 点は，本研究で提案するロボットシステムが実環境下で稼動し，追加的に学習を行うための重要な機能である．この機能の有効性を検証するために，人工データと実画像データを用いた 2 つの予備実験を行った．

#### § 2 人工データを用いた予備実験

図 7($a$) に示す 2 次元の人工データをオンラインで追加的に入力した場合の SOINN の挙動を検証した．データセットは 2 つのガウス分布，2 つの同心円，及び *Sin* 曲線の合計 5 つのクラスによって構成されている．また，実世界の環境を想定して，5 つのクラスから生起するデータに 10%の一様ノイズが加えられている．このデータセットをオンラインで追加的に入力し，SOINN に教師なし分類を行わせた．

この入力データが SOINN によって分類された後の出力結果を図 7 ($b$) に示す．図 7 ($b$) より入力データに含まれるノイズは削除され，入力データのクラスタ数とそのトポロジーが正しく抽出されていることがわかる．SOINN のアルゴリズムの詳細については [Shen 06] に記載されている．

#### § 3 画像データを用いた予備実験

さらに SOINN の基本的機能を検証するために，画像データを用いた予備実験を行った．この実験では，追加的に入力される視覚情報 (画像データ) を SOINN への入力ベクトルとした．実験には，黒い背景の上に物体が映っている画像を用いた．2･3 節の手法を用いて，画像から特徴抽出を行った．特徴抽出の後，これらの特徴を SOINN に入力した．

ワークスペースには何もない状態から実験を開始した．[*STEP 1.*] まずワークスペースの中に 4 つの物体を順々に置く．ここで 4 つの物体はそれぞれ赤色の正方形，黄色の正方形，黄色の長方形，緑色の細長い長方形である．次にこれらの物体が SOINN により教師なし分類される．この教師なし分類過程を図 8 に示す．図 8(a) は 4 つの物体が順々に置かれる様子 (上段) と，SOINN 空間でのニューロの様子 (下段) を示す．図 8(b) は SOINN により 4 つの物体が学習された後の様子を示す．図 8(b) の下段より 4 つ全ての物体が置かれ SOINN により学習された後に，色空間 (color vector space) で 3 つの，形状空間 (shape vector space) では 3 つの，物体空間 (object vector space) では 4 つのクラスタがそれぞれ構成されている．

なお図 8〜図 10 では共通して，上段の図はワークスペース上の様子を示し，下段の図はその時の SOINN 空間で

のニューロの様子を示している．

[*STEP 2.*] 新たに緑色の三角形の物体をワークスペースに置く．新しい物体 (緑色の三角形) が入力された際の SOINN 特徴空間の変化を図 9 に示す．ここで新しい物体をワークスペースに置く際，パートナーの手が画面に映りこむ．この手に反応してニューロ群の数が増加している (図 9(a))．しかし SOINN では一時的な情報はノイズとして解釈されるため，時間が経つとこのノイズは削除される (図 9(b))．この結果より SOINN がノイズに対して頑健性を有することが示された． また，ノイズが消去された後，形状特徴空間および物体特徴空間に新しいクラスタが形成されている (図 10(a))． この新しいクラスタは，新たにワークスペースに置かれた物体に対応する．ここで注目すべきは，新規のクラスタの形成が既存のクラスタに影響を与えていない点である．
実験の結果，提案システムでは既学習のクラスタを壊すことなく，新規のクラスタが学習された．ここで新しい形状 (三角形) および新しい物体 (緑色の三角形) に対応するクラスタは追加的に獲得されたが，緑色に対応するクラスタは既知 (STEP1. で学習済み) であるため，色特徴空間に変化は見られていない (図 10(b))．以上より提案システムではオンラインで追加的に視覚情報を学習可能であることを示した．

## § 4 SOINN のパラメータ設定

SOINN の学習に必要な，パラメータ設定について説明する．SOINN において，各クラスタはニューロとニューロ間をつなぐエッジ (辺) で表現される (図 11)．このニューロとエッジは入力データに応じて生成･削除を繰り返す．ここでノイズに対してニューロを生成した場合，分類結果が悪化する．SOINN ではノイズ除去を行うために，ニューロとエッジの生成・削除に関するパラメータ $\lambda$,$age_{dead}$ を導入している．これらのパラメータをいかに設定するかが，SOINN の機能にとって重要となる．

まず $\lambda$ はノイズとおぼしきニューロを削除する周期である．$\lambda$ を小さな値に設定すると頻繁にノイズ処理が行われるが，極端に小さくすると実際はノイズではないニューロを誤って削除してしまう．逆に $\lambda$ を極端に大きな値に設定するとノイズの影響で生成されたニューロを適切に取り除くことができない．

次に $age_{dead}$ はノイズなどの影響で誤って生成されたエッジを削除するために用いられる．$age_{dead}$ を小さな値に設定するとエッジが削除されやすくなりノイズによる影響を防ぐことができるが，極端に小さくすると頻繁にエッジが削除され学習結果が不安定になる．逆に $age_{dead}$ を極端に大きな値に設定すると，ノイズの影響で生成されたエッジを適切に取り除くことができない．したがって学習対象のデータやタスクに応じて，これら 2 つのパラメータを設定する必要がある．

本研究では予備実験 (2･5･3 節) の結果から，SOINN のパラメータをそれぞれ $\lambda = 100$,$age_{dead} = 20$ に設定した．$\lambda$,$age_{dead}$ の他に SOINN にはパラメータ $c, \alpha_1, \alpha_2, \alpha_3, \beta, \gamma$ が存在するが，これらについては [Shen 06] に示された値 $c = 1, \alpha_1 = 1/6, \alpha_2 = 1/4, \alpha_3 = 1/4, \beta = 2/3, \gamma = 3/4$ を使用した．

## 2･6 概念の表現 (音声・視覚情報の統合)

提案システムでは，SOINN によって得られる物体に関する視覚情報 (2･3 節) と，その物体に関する発話から得られる音声情報 (2･4 節) が関連付けられる (統合される) ことで，物体の概念が獲得される．

視覚情報からの入力データの分類結果が安定した (SOINN での学習が収束した) 後，パートナーはワークスペース上の任意の物体を指差しながら，マイクでその物体に関する単語を発話する．次に，発話から得られる音声データを VQ コードに変換する．VQ コードを，音声処理モジュールで得られた各クラスの参照テンプレートと比較して認識する．どのテンプレートとも入力 VQ コードが異なる場合，この入力 VQ コードを新規の音声クラスの参照テンプレートとし，音声クラスのインデックス数を増やす．このインデックス数は視覚情報と統合するために用いられる．次に SOINN の 3 種類 (色，形状，物体) の特徴空間上で，指示された物体に関する発話がどのクラスタ (例えば，黄色，正方形) と対応するかが認識される．例えばパートナーがワークスペースのレモンを指差すと同時に「黄色」と発話したとすると，3 種類の特徴空間のそれぞれで，対応するクラスタに属するニューロの関連度 (関連度については 2･6･2 節で説明する) が変化する．これは，学習の初期段階では「黄色」という発話が何の属性 (色の名称なのか，形状の名称なのか，物体の名称なのか) を示すのかがロボットにはわからないためである．

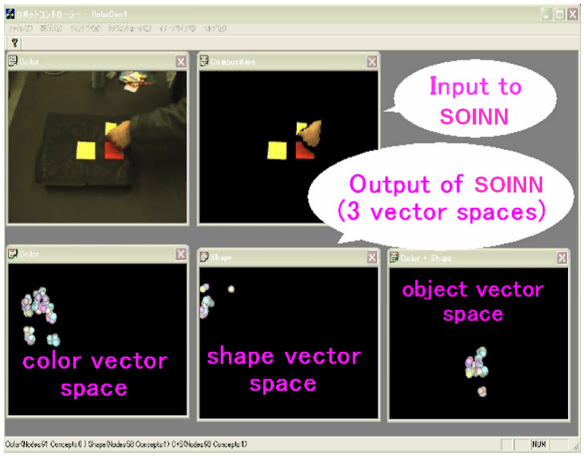
以上のように，物体をロボットに見せて，その物体に関する単語を発話する作業を繰り返し行う．この操作を通して，ロボットは「黄色」という言葉の意味が (物体の名称および形状の名称ではなく) 色の名称であるということを理解する．この理解が，本研究で定義する概念の獲得である．以下では，音声・視覚情報の統合により概念を獲得するアルゴリズムを説明する．このアルゴリズムは 5 つの主要な部分で構成される．

### § 1 概念の定義

SOINN により形成されたクラスタに 4 つ以上のニューロが含まれている場合，本研究ではそのクラスタを概念に対応付ける．ここでニューロの数 4 は実験的に決定した．クラスタに含まれるニューロの数が 4 つ未満の場合，そのクラスタは除去される．

### § 2 クラスタと概念の関連度の定義

クラスタに含まれるニューロと発話された単語との関連性の度合い (関連度)$S$ を $S = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_N\}$ と定義する．ここで $N$ は SOINN 空間上のニューロの総数を




(a)


(b)

**図 8** (a) ワークスペースに物体を置いた時の SOINN 空間の様子 (上段は，4 つの物体が順々にワークスペースに置かれる様子を表している．下段は物体が順々に置かれた場合の，SOINN 空間でのニューロの様子を表している．下段は左からそれぞれ，色特徴空間 (color vector space)，形状特徴空間 (shape vector space)，物体特徴 (色と形状を併せた特徴) 空間 (object vector space) を示す．)

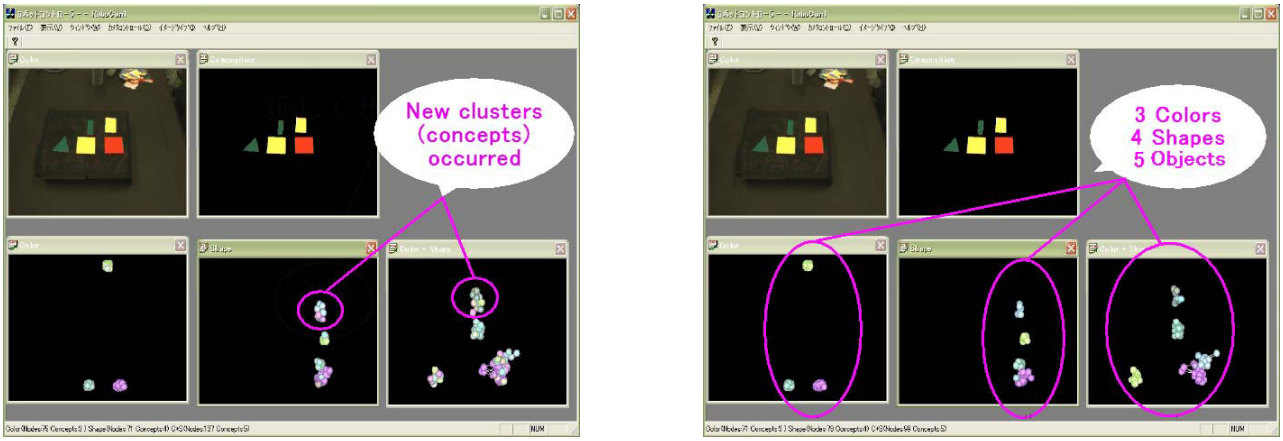
(b) SOINN により 4 つの物体が学習された後の様子 (色特徴空間で 3 つの，形状特徴空間では 3 つの，物体特徴空間では 4 つのクラスタがそれぞれ構成されている．)


(a)


(b)

**図 9** (a) 新しい物体 (緑色の三角形) が入力された際の SOINN 空間の様子 (新しい物体をワークスペースに置く際，パートナーの手が画面に映りこむ (上段)．この手に反応して SOINN 空間のニューロ群の数が増加している (下段)．)

(b)(a) の状態から一定時間が経過した後の様子 (SOINN では一時的な情報はノイズとして解釈し，一定時間後に削除する)




(a) (b)

**図 10** (a) ノイズが消去された後の SOINN 空間の様子 (ノイズが消去された後，形状特徴空間および物体特徴空間に，新しいクラスタ (緑色の三角形) が形成されている．)

(b) ニューロの増減が安定した後 (学習が終了した後) の SOINN 空間の様子 (新しい形状 (三角形) および新しい物体 (緑色の三角形) に対応するクラスタは追加的に獲得されたが，色特徴空間に変化は見られない．)

示し，$\mathbf{c}_i$ は $i$ 番目のニューロと各単語との関連度ベクトルである．$j$ を単語のクラスタ番号とすると，$\mathbf{c}_i$ は $\mathbf{c}_i = \{d_{i1}, d_{i2}, \cdots, d_{ij}\}$ と表せる．ここで $d_{ij}$ は初期値を 0 とした正の値であり，$i$ 番目のニューロと $j$ 番目の単語の関連度である．また 1 つの概念は全ての単語との関連度を保持する．$p$ 番目のクラスタ $C_p$($p$ 番目の概念) の関連度 $\mathbf{cc}_p$ は以下のように表せる．

$$\begin{aligned}\mathbf{cc}_p &= \frac{1}{a_p}\sum_{n_i \in C_p} \mathbf{c}_i \\ &= \frac{1}{a_p}\sum_{n_i \in C_p} \{d_{i1}, d_{i2}, \cdots, d_{ij}\} \\ &= \{\frac{1}{a_p}\sum_{n_i \in C_p} d_{i1}, \cdots, \frac{1}{a_p}\sum_{n_i \in C_p} d_{ij}\} \\ &= \{\tilde{d}_{p1}, \tilde{d}_{p2}, \cdots, \tilde{d}_{pj}\}. \end{aligned} \quad (6)$$

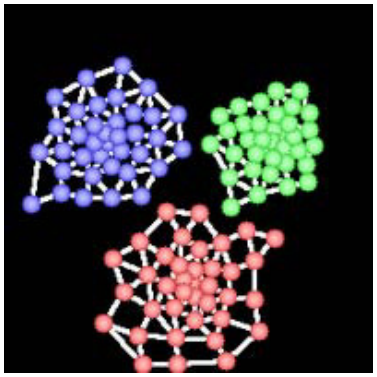


**図 11** SOINN による分類結果 (ここでデータ群は 3 クラス (青，緑，赤) に分類されている．各クラスタはニューロとエッジ (白色) で表現されている．)

式 (6) で $a_p$ は $p$ 番目のクラスタに含まれるニューロ $n_i$ の総数を示す．また関連度 $\mathbf{cc}_p$ は，あるクラスタに含まれる全ニューロの関連度の平均値とする．

## § 3 物体および概念に関する既知度の定義

各概念 $C_p$ に関してロボットが知っている度合い (既知度)$k_p$ を以下の式で定義する．

$$k_p = \frac{\max(\mathbf{cc}_p)^2}{\sum_i (\tilde{d}_{pi})^2} \quad (7)$$

ここで $\max(\mathbf{cc}_p)$ は $\{\tilde{d}_{p1}, \tilde{d}_{p2}, \cdots, \tilde{d}_{pj}\}$ の中での最大値である．物体 $ok$ に関する既知度は，3 種類の概念に関する既知度の線形和として定義され，以下の手順で計算される．

まず，物体から画像処理により得られる色ベクトル (3 次元)，形状ベクトル (8 次元)，及び物体そのものを表すベクトル (11 次元) のそれぞれが，どのクラスタに所属するものなのかを，ベクトルごとに最近傍決定則により決定する．特徴空間 (例えば色特徴空間) で，対応するクラスタが一つも存在しない場合，その属性 (色) に関する既知度は 0 とする．一方 3 つの特徴ベクトルが，それぞれの空間内の $p$ 番目と $q$ 番目と $r$ 番目の概念に最も近かった場合，これらの概念の関連度を $\mathbf{ccc}_p, \mathbf{ccs}_q, \mathbf{ccn}_r$ で表す．以下の式 (8),(9),(10) は一般的な概念の関連度を表す式 (6) を色や形状などの概念の関連度の式に直したものである．

$$\mathbf{ccc}_p = \{\tilde{dc}_{p1}, \tilde{dc}_{p2}, ..., \tilde{dc}_{pj}\} \quad (8)$$

$$\mathbf{ccs}_q = \{\tilde{ds}_{q1}, \tilde{ds}_{q2}, ..., \tilde{ds}_{qj}\} \quad (9)$$

$$\mathbf{ccn}_r = \{\tilde{dn}_{r1}, \tilde{dn}_{r2}, ..., \tilde{dn}_{rj}\} \quad (10)$$

提案システムでは，「色，形状，及び物体そのものの名称が一致することはない」と仮定している．そのため，

$\arg\max(\mathbf{ccc}_p)$ と $\arg\max(\mathbf{ccs}_q)$ と $\arg\max(\mathbf{ccn}_r)$ のいずれかが一致する場合，物体の既知度を低く設定する必要がある．そこで式 (7) に対して次式のような拡張を行う．

$$kc_p(a,b) = \frac{\max(\mathbf{c\acute{c}c}_p)^2}{\sum_i (\tilde{dc}_{pi})^2} \quad (11)$$

$$(where\ \mathbf{c\acute{c}c}_p := \mathbf{ccc}_p \setminus \{\tilde{dc}_{pa}, \tilde{dc}_{pb}\})$$

$$ks_q(a,b) = \frac{\max(\mathbf{c\acute{c}s}_q)^2}{\sum_i (\tilde{ds}_{qi})^2} \quad (12)$$

$$(where\ \mathbf{c\acute{c}s}_q := \mathbf{ccs}_q \setminus \{\tilde{ds}_{qa}, \tilde{ds}_{qb}\})$$

$$kn_r(a,b) = \frac{\max(\mathbf{c\acute{c}n}_r)^2}{\sum_i (\tilde{dn}_{ri})^2} \quad (13)$$

$$(where\ \mathbf{c\acute{c}n}_r := \mathbf{ccn}_r \setminus \{\tilde{dn}_{ra}, \tilde{dn}_{rb}\}),$$

式 (11) で，$a$,$b$ は単語のインデックスを示す．$a$,$b$ が与えられない場合には，式 (7) を用いて直接 $kc_p$ が計算される．

次に色，形状，及び物体そのものの概念が持つ既知度 $(kc_p, ks_q, kn_r)$ を以下のように定義する．

(1) When $\max(\mathbf{ccc}_p) \geq \max(\mathbf{ccs}_q)$,

$$\acute{kc}_p = kc_p(\emptyset,\emptyset) \quad (14)$$

$$\acute{ks}_q = ks_q(\arg\max(\mathbf{ccc}_p),\emptyset) \quad (15)$$

$$\acute{kn}_r = kn_r(\arg\max(\mathbf{ccc}_p), \arg\max(\mathbf{ccs}_q \setminus \{\tilde{ds}_{q(\arg\max(\mathbf{ccc}_p))}\})) \quad (16)$$

(2) When $\max(\mathbf{ccc}_p) < \max(\mathbf{ccs}_q)$,

$$\acute{ks}_q = ks_q(\emptyset,\emptyset) \quad (17)$$

$$\acute{kc}_p = kc_p(\arg\max(\mathbf{ccs}_q),\emptyset) \quad (18)$$

$$\acute{kn}_r = kn_r(\arg\max(\mathbf{ccs}_q), \arg\max(\mathbf{ccc}_p \setminus \{\tilde{dc}_{p(\arg\max(\mathbf{ccs}_q))}\})) \quad (19)$$

式 (14) において $\arg\max(\mathbf{ccc}_p)$ は $\mathbf{ccc}_p$ の中で最大の要素のインデックスを表す．これは，$\mathbf{ccc}_p$ の中で最も関連度が大きい単語が $\arg\max(\mathbf{ccc}_p)$ であることを意味する．また $\arg\max(\mathbf{ccs}_q)$,$\arg\max(\mathbf{ccn}_r)$ に関しても同様に定義する．

最後に以下の式 (20) で，物体に関する総合的な既知度 $ok$ を定義する．

$$ok = \frac{kc_p + ks_q + kn_r}{3} \quad (20)$$

この既知度は，ロボットがその物体をどの程度知っているかの指標である．ある物体に対する既知度が小さければ，ロボットはその物体に関して情報が少ないと理解する．ロボットは既知度の小さい物体に関する情報を，能動的にパートナーに要求する．この機能を「知識欲」と定義し，後の 2･7 節で述べる．

## § 4 音声が入力された場合の関連度の変化

ある物体に対して $k(1 \leq k \leq j)$ 番目のクラスに分類された音声を入力したときの動作は以下のようになる ($j$ は音声クラスの数)．最初に，物体から色，形状，及び物体そのものを意味する 3 つの特徴ベクトルを抽出し，これらのベクトルから最も近いクラスタをそれぞれ最近傍決定則を用いて求める．このとき，各空間内に対応するクラスタが一つも存在しなかったらその属性に関して音声は関連付けられない．一方，そのような概念 $C$ が存在すれば，その概念に所属するすべてのニューロ $n$ の関連度の $k$ 番目の要素を次式のように 1 だけ増やす．

$$d_{ik} := d_{ik} + 1\ (\forall i\{n_i \in C\}) \quad (21)$$

ここで $k$ は $k$ 番目の単語示す．

## § 5 関連度の伝播

2･5･4 節より，SOINN では一定時間ごとにノイズ処理のため，ニューロを削除している．この機能によって，ノイズ以外のニューロが誤って削除される場合がある．ここである音声と関連度を持ったニューロが誤って削除された場合，音声情報が損失してしまう．この問題をニューロ間で関連度を伝播することで回避する．$i$ 番目のニューロを $n_i$，$i$ 番目のニューロに隣接するニューロの集合を $N_i$，$i$ 番目のニューロに隣接するニューロの数を $a_i$ とする．また時刻 $t$ における $i$ 番目のニューロの関連度 $\mathbf{c}_{i(t)}$ を次式で表す．

$$\mathbf{c}_{i(t+1)} = (1-\alpha)\mathbf{c}_{i(t)} + \frac{\alpha}{a_i} \sum_{n_j \in N_i} \mathbf{c}_{j(t)} \quad (22)$$

ここで $\alpha = 0.01$ とした．短い周期 $t$ ごとにこの操作を行うことで，ニューロが消滅する前に周囲のニューロに音声情報との関連度を伝播することが可能となる．

## § 6 音声の識別

提案手法では物体に対応している音声を以下のようにして求める．色に関する音声を $s_i$，形状に関する音声を $s_j$，物体そのものに関する音声を $s_k$ とする．まず，物体から 3 つの特徴ベクトルを抽出し，最近傍決定則を使ってこれらのベクトルから最も近いクラスタをそれぞれ求める．対応するクラスタが属性空間内に一つも存在しなかったら，その属性に関する音声は不明だと判断する．一方属性ベクトルが，それぞれの空間内の $p$ 番目と $q$ 番目と $r$ 番目の概念に最も近かったら，色の概念に関する確信度を $\mathbf{ccc}_p$，形状の概念に関する確信度を $\mathbf{ccs}_q$，物体そのものの概念に関する確信度を $\mathbf{ccn}_r$ のように記す．そして以下のルールに基づいて各属性に結合している最適な音声を決定する．

(1) $\max(\mathbf{ccc}_p) \geq \max(\mathbf{ccs}_q)$ のとき

$$i = \arg\max(\mathbf{ccc}_p) \quad (23)$$

$$j = \arg\max(\mathbf{ccs}_q \setminus \{\tilde{ds}_{qi}\}) \quad (24)$$

$$k = \arg\max(\mathbf{ccn}_r \setminus \{\tilde{dn}_{ri}\} \setminus \{\tilde{dn}_{rj}\}) \quad (25)$$

(2) $\max(\mathbf{ccc}_p) < \max(\mathbf{ccs}_q)$ のとき

$$j = \arg\max(\mathbf{ccs}_q) \quad (26)$$

$$i = \arg\max(\mathbf{ccc}_p \setminus \{\tilde{dc}_{pj}\}) \quad (27)$$

$$k = \arg\max(\mathbf{ccn}_r \setminus \{\tilde{dn}_{ri}\} \setminus \{\tilde{dn}_{rj}\}) \quad (28)$$

ただし，各属性に関する既知度が 0 であるとき，すなわち $kc_p = 0, ks_q = 0, kn_r = 0$ であるとき，その属性に関する音声は不明であるものとする．

### 2·7 知識欲を利用した概念獲得

人間は何も知らない物体を提示されたとき，”これは何? ”や ”これの呼び名は何? ”といった質問をする．ロボットにもこのような知識欲を持たせることが望まれる．ロボットが知識欲を持つことで，能動的に学習に参加すればより効率的に学習が進むと考えられる．具体的には，システムが各物体に関する既知度 (式 (20)) を算出した後，既知度の低い物体 $o_i$ を選択し，この物体に関する情報 (発話) をパートナーに要求する．情報の要求は，ロボットが特定の物体をグリッパで指し示すことで行われる．ここで，ロボットの指し示しが物体に関する発話を要求していることを，パートナーは理解しているものとする．

上記の既知度の低い物体 $o_i$ は式 (29) に基づいて選択される．

$$\begin{aligned} i &= \arg_i \max \sum_{j=1}^{n} \mathrm{E}(\Delta(ok_j, ok_j|(o_i, s))) \quad (29) \\ &= \arg_i \max \sum_{j=1}^{n} \frac{1}{t+1} \big(\Delta(ok_j, ok_j|(o_i, s_1)) \\ &\quad + \Delta(ok_j, ok_j|(o_i, s_2)) \\ &\quad + \cdots \\ &\quad + \Delta(ok_j, ok_j|(o_i, s_t)) \\ &\quad + \Delta(ok_j, ok_j|(o_i, s_{t+1}))\big) \end{aligned}$$

式 (29) において以前に記憶した単語を $s$，その単語の合計数を $t$，音声クラスの数を $j$，$i$ 番目の物体の既知度を $ok_i$ とする．また $s_{j+1}$ は未知の音声を意味する．未知の音声が入力された場合，$s$ の合計数が $t$ から $t+1$ になる．なお E は $t+1$ 個の単語に関する期待値，$\Delta$ は変化量を意味する関数である．$\Delta$ は $(o_i, s)$ が与えられる前後での，$ok_j$ の差分を示す．$|(o_i, s)$ は 「$(o_i, s)$ を与えられた場合に」 という意味を表す演算子である．

## 3. メンタルモデル

ロボットの内部の世界を構築するために，メンタルモデルを定義する (図 12)．メンタルモデルは [Weng 01, Roy 04] で提案されたロボットの内部世界の表現にも利用されている．またメンタルモデルは人間の内部にも構築されており，人間同士のコミュニケーションにおいて重要な役割を果たしていることが [Piaget 56, Hsiao 03] で報告されている．

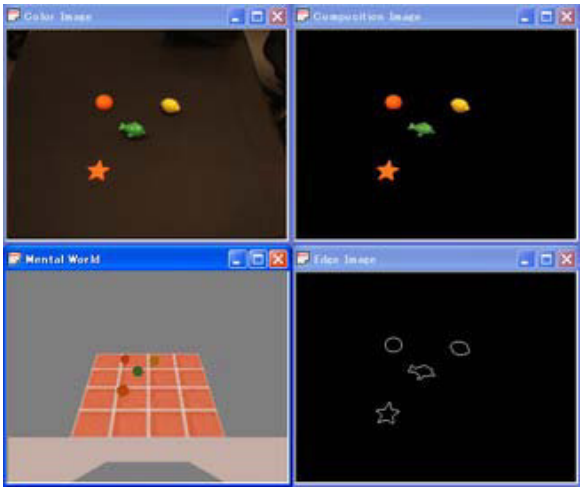



**図 12** IKR1 のメンタルモデル (メンタルモデルでは IKR1 の前に置かれている物体の数や位置，物体の特徴 (色，形状) といった情報を記録している．)

メンタルモデルには，ロボットの前に置かれている物体の数や位置，物体の特徴 (色，形状) といった情報や，各物体に関して発話された音声情報が記憶される．またメンタルモデルではロボットの視点から見た，パートナーと物体の位置関係の情報 (三次元座標) が記録される．これらの情報を保持することで，ロボットはパートナーと物体に関する知識 (例えば，ロボットと人間の間に置かれている物体の位置など) を共有することが可能となる (図 13)．さらにメンタルモデルは，時間が経つにつれて変化するワークスペースの状況 (例えば，ある時刻にロボットの前に新しい物体が置かれる状況など) を把握する役割を担う．まず，ステレオカメラ (ロボットの目) から連続的にワークスペースの情景が入力される．ここでロボットは異なるフレーム間で検出された物体が，同じ物体かどうかを認識する必要がある．このため提案手法では，前場面のフレームと現場面のフレームの差分を算出し，現場面の物体の特徴と前の場面の物体の特徴を比較する．特徴を比較して前フレームの物体の特徴と現フレームの物体の特徴が異なっている場合，現場面の物体の特徴を新たなメモリとして更新する．

ここで物体の位置と視覚特徴の情報は，新しい物体が入力されるごとに更新される．観測された物体の位置情報は，カメラの座標系からロボットの絶対座標系に変換される．以上の操作により，ロボットは現在のワークスペースの状況を把握する．

## 4. 実　　験

### 4·1 追加学習実験

本実験では，色が 9 種類，形状が 8 種類の合計 72 種類の物体を用いて，提案システムにおいてオンラインの追

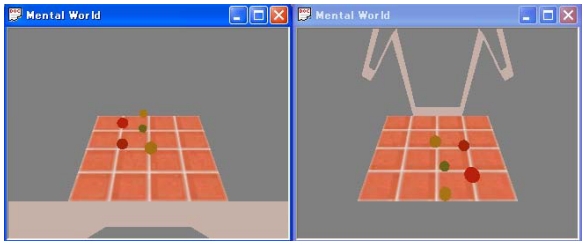



**図 13** メンタルモデルの動作 (メンタルモデルにはロボットの視点から見た，パートナーと物体の位置関係の情報 (三次元座標) が記録される．これらの情報を保持することで，ロボットはワークスペース内の物体の位置などを理解する．)

**図 14** 追加学習実験で用いた学習対象の例

加的な学習が可能であることを示す．実験で用いた物体の例を図 14 に示す．以下に実験の具体的な手順を示す．

[*STEP 1.*] 72 の物体の中から，4 種類の異なる色と 3 種類の異なる形状を持つ 12 個の物体をランダムに選択する．

[*STEP 2.*] 選択された 12 個の物体から 4 個の物体を選択し，それをロボットに見せる．それぞれの物体について，パートナーは物体を指差しながら色，形状，名前 (3 つの特性) を発話する．

[*STEP 3.*] 物体が示されている状態で，ロボットは対応する言葉 (色，形状，名前) を発話することによって反応する． ロボットが発話する内容が間違いだった場合，再度パートナーはその物体を指差しながら，正しい言葉を発話する．ここで，内容の修正を行うのは，1 物体に付き 1 回限りとする．

[*STEP 4.*] 残り 8 個の物体から 4 個の物体を選択し，*STEP 2* と *STEP 3* の操作を行う．

[*STEP 5.*] 残りの 4 個の物体を選択し， *STEP 2* と *STEP 3* の操作を行う．この時点で，12 個の物体に関連する 36(12 個× 3 個の特性) 単語が入力されたことになる．

[*STEP 6.*] 学習に用いた 12 個の物体と，それらに関連する 3 種類の単語 (色：COLOR，形状：SHAPE，物体の名称：OBJECT) の関係の正当性を評価する．12 個の物体に関連する単語をロボットが正しく発話出来た場合に，概念を獲得したと見なし正答とする．表 1 の 1 列目に正答率の結果を示す．

**表 1** 物体と発話された言語が正しく関連付けられた割合 ( % ).

| STAGE | 1st | 2nd | 3rd | 4th | 5th | 6th |
|---|---|---|---|---|---|---|
| NUMBER OF OBJECTS | 12 | 20 | 30 | 42 | 56 | 72 |
| COLOR | 100 | 100 | 100 | 100 | 100 | 100 |
| SHAPE | 100 | 100 | 100 | 100 | 100 | 100 |
| OBJECT | 100 | 95.00 | 100 | 95.24 | 98.21 | 95.83 |

[*STEP 7.*] 次に，色，形状ともに 1 種類ずつ増やし，計 5 種類の色，4 種類の形状を用意する (この操作で新しい 8 個の物体が追加され，計 20 個の物体が用意される)．ここで新たに追加された 8 物体に対して，*STEP 2* と *STEP 3*) を行う．

[*STEP 8.*] *STEP 6.* と同様に 20 個の物体と，それらに関する単語の関係の正当性を評価する．評価の結果を表 1 の 2 列目に示す．

[*STEP 9.*] さらに，色の数，形状の数をそれぞれ $(6,5),(7,6),(8,7),(9,8)$ と増加させて，*STEP 7* と *STEP 8* を行う．最終的に，72 個の物体に対して認識率を算出した．

### 4·2 追加学習実験の結果

表 1 より最初の段階 (1st) では，学習した物体に関連する全ての単語を認識 (概念を獲得) した．実験の結果を総じて，多数の物体を追加的に学習した場合でも，システムは物体を高い精度で認識しており，オンラインでの追加学習機能を有することを示した．

また本実験では，ロボットが誤り続けた場合にも，1 物体に付き 1 度の修正しか行わなかった．それにもかかわらず，本実験では物体の名称を誤答する場合があったものの，物体に関する色，形状に関する単語については一度も誤答しなかった．

### 4·3 知識欲の有効性に関する実験

次に 2·7 節で述べた「知識欲」の有効性に関する実験を行った．この実験では，学習メカニズムに「知識欲」を導入することで学習効率にどのような影響を与えるかを検証した．使用した物体は図 15 に示す，色が 4 種類，形状が 4 種類の合計 16 種類で，これらの物体から毎回ランダムに 4 種類の物体を選択しロボットの前に置いていく．そして，ロボットの前に置いた 4 つの物体から，以下の戦略のもとに音声を入力する物体を一つだけ選択する．

(戦略 $A$) ロボットが，各物体についての既知度を基準として 1 つの物体を選択する．

(戦略 $B$) 教示者が任意の物体をランダムに選択する．

選択した物体に対して，物体の色，形状，および物体

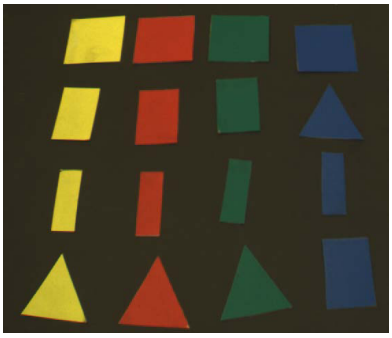


**図 15** 知識欲を検証する実験で用いた学習対象の物体

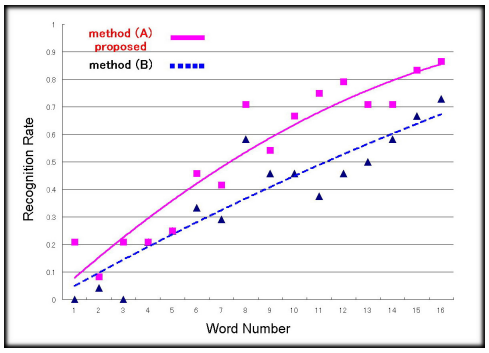



**図 16** 認識率 (横軸は追加された物体の数，縦軸はその数の物体に対する認識率を示す．)

の名称に関する教示 (発話) を行い，教示が完了したら，4 つの物体を一旦ワークスペースから外す．以上，16 種類の物体の内から 4 種類の物体をランダムに選択するところから，物体をワークスペースから外すまでの操作を 1 サイクルと考え，全部で 16 サイクルの操作を行った．ここでロボットの前に置いた 4 つの物体に対して音声を教示する前に，既知度及び音声結合正確率の測定を毎回行った．さらに，16 サイクルの操作が完了した後に，全ての物体に対して既知度と音声結合正確率を測定した．なおここで音声結合正確率とは，測定対象となっている複数の物体の色，形状，および物体の名称をロボットが発話という形状で正確に出力できた場合の割合である．

### 4·4 知識欲に関する実験結果

以上の実験を二回行った．二回の実験から得られた結果の平均を図 16 に示す．実線が (戦略 A) を取り続けた場合，点線が (戦略 B) を取り続けた場合の結果を示している．図 16 から，ロボットに教示の対象を求めさせることが学習効率の向上に寄与していることが認められる．これは，ロボットが“ 自らの知識の範囲を理解している ”ことを示している．

提案システムでは，ロボットにグリッパを用いた指し示しの機能を持たせることで，ロボットにも積極性および能動性を備えさせ,“ より自然な ”学習を行うことを試みた．この結果，学習効率の向上を確認した．ここで注意したいのが，知識欲がなかった場合 ((戦略 $B$) をとり続けた場合) でも提案システムは追加的に概念を獲得可能であるということである．総じて知識欲は，提案システムの学習効率を向上させる役割を担っている．本研究では今後のロボットの知能学習において，人間とロボットの間に提案システムのような双方向的なコミュニケーションを導入することの重要性を主張しておきたい．

## 5. 考　　察

実験 1,2 の結果より，事前知識のない状態からの追加的概念 (言語) 獲得が可能であることを示した．以下では大きく 2 つの節に分けて考察を行う．まず，提案システムの課題について議論する．次に，日常生活環境下におけるヒューマノイドロボットの言語獲得といったより高いレベルの目標を実現するために，必要な機能について議論する．

### 5·1 提案システムの課題

提案システムの機能は，音声情報処理モジュールと視覚情報処理モジュールの主要な 2 つの機能に依存する．つまりこの 2 つの機能に問題が生じた場合，システムとして破綻することになる．本研究では，比較的簡易な画像 (背景が黒) および音声 (連続音声でなく分節化された単語を使用) を用いて実験を行った．このことにより起き得る問題および必要な機能を，音声情報処理モジュールと視覚情報処理モジュールに分けて議論する．

#### § 1 音声情報処理モジュールにおける課題

本研究の目的はロボットと教示者のインタラクションを通じて，視覚情報および音声情報に関する事前知識を全く持たない状態から (訓練データを学習に用いずに)，オンラインで追加的に言語獲得を行うことであった．この目的を達成するために，音声に関して音素モデルの構築や，訓練データを用いた学習器の構築を必要としない手法として DTW を用いた．

音声情報処理に用いられる他の代表的な手法に HMM があるが，HMM では一般に多数の訓練データを用いたバッチ学習や，入力音声に応じたトポロジー (状態数などのパラメータ) の決定を必要とするため，本研究の目的には適さないと考えた．[Roy 02, Iwahashi 04] など多くの研究では HMM を用いて音声処理を行っているが，本研究では上記の理由により HMM を用いなかった．

しかし DTW は，HMM と異なり特徴空間上の分布をモデル化出来ないため，結果的に音声の認識精度が低下する．例えば [中川 93] では，言語獲得において連続音声データから単語を抽出する手法に DTW(DP マッチング) を用いた結果，充分な抽出精度が得られなかったと報告されている．これらの背景を踏まえ，今後，事前知識のない状態からの連続音声の頑健な学習・認識，および文法の獲得を実現する上で，DTW に代わる手法の導

入を検討する必要がある．

### §2 視覚情報処理モジュールにおける課題

#### 特徴抽出部における課題

本研究では実験を室内で行った上，実験では黒い背景の上に物体が映っている画像を用いた．しかし日常生活環境では，物体の背景は多様であり，室外であれば日照条件も異なる．この場合，本実験のように物体の形状を正確に切り出すことが困難となる．よって日常生活環境下においても，物体を正確に切り出し，物体の特徴を抽出可能な画像処理手法が必要である．

#### 学習器 (SOINN) における課題

今回の実験で用いた入力視覚情報は簡易なものであり，各クラスタは十分に分離可能であった．しかし一般に，日常生活環境で得られる視覚情報を，SOINN を用いてクラスタに分類することは容易ではない．これは，カメラから得られる画像情報に大量のノイズが含まれたり，クラスタ間に大きなオーバーラップが生じる可能性があるためである．こうした問題にも対処出来るよう，SOINN の機能を拡張する必要がある．

### 5·2 言語獲得における課題

今回の提案システムでは，名詞しか扱うことが出来ない．これより今後，名詞以外の概念 (例えば,「近づける」,「遠ざける」,「またぐ」などの動詞) の獲得も可能とするよう提案手法を拡張する必要がある．また提案システムでは予め分節化した単語を処理対象としており，連続的な発話に対応していない．よってこの点に関しても拡張を行い，連続した発話を単語単位に分節化し，単語の順序から文法の構造を獲得可能なシステムの実現を目指す．

また本研究では，物体は 3 つの属性 (名称，形状，色) のみを有すると仮定し，パートナーもこの 3 つの属性のいずれかについて発話するとした．しかし一般に人間は，物体に対してこの 3 つの属性以外の呼称を用いる場合もある．こうした場合にも的確に対応出来るよう，システムを拡張する必要がある．

これに加え，本研究では 1 つの属性に付き 1 つの単語が対応すると仮定した．しかし現実的には，ある属性とそれを表す単語が 1 対 1 に対応しない場合が想定される．この場合，SOINN 空間の 1 つのクラスタに対し複数の音声を対応させる必要がある．今後，こうした問題にも対応出来るよう，システムを拡張する必要がある．

## 6. 結　　論

本研究では，オンラインでリアルタイムに動作する概念獲得システムを提案した．このシステムでは，音声情報と視覚情報を統合することにより，発達的に言語とその意味を学習する．

実験 1 では，パートナーとのコミュニケーション (指差しによる教示) を通じ，ロボットに色や形状や物体そのものの概念を追加的に学習・獲得させた．実験 2 では，ロボットから人間に積極的に情報を求めることがロボットの学習機能の効率化に役立つことを示した．

今後，5 章で述べた課題を解決し，日常生活環境において発達的に言語獲得を行えるシステム (ヒューマノイドロボット) の実現を目指す．

### 謝　　辞

本研究の実施にあたり NEDO 産業技術研究助成事業から支援を頂きました．記して感謝いたします．

## ◇ 参 考 文 献 ◇

[Elman 93] Elman, J.: Learning and Development in Neural networks: The Importance of Starting Small, *Cognition*, Vol. 48, pp. 71–99 (1993)

[Fritzke 95] Fritzke, B.: A Growing Neural Gas Network Learns Topologies, in *neural information processing systems (NIPS)*, pp. 625–632 (1995)

[Gersho 92] Gersho, A. and Gray, R. M.: *Vector Quantization and Signal Compression*, Kluwer, Boston (1992)

[Gorin 99] Gorin, A. L., Petrovksa-Delacretaz, D., Riccardi, G., and Wright, J.: Learning Spoken Language without Transcriptions, *IEEE Workshop Speech Recognition and Understanding* (1999)

[Harnad 90] Harnad, S.: The Symbol Grounding Problem, *Physica*, Vol. D, No. 42, pp. 335–346 (1990)

[Hsiao 03] Hsiao, K., Mavridis, N., and Roy, D.: Coupling Perception and Simulation: Steps Towards Conversational Robotics, in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (2003)

[Imai 02] Imai, S., Kobayashi, T., Tokuda, K., Masuko, T., and Koishida, K.: Speech Signal Processing Toolkit: SPTK Version 3.0 (2002)

[Iwahashi 03] Iwahashi, N.: Language Acquisition through A Human-robot Interface by Combining Speech, Visual, and Behavioral Information, *Information Sciences*, pp. 109–121 (2003)

[Iwahashi 04] Iwahashi, N.: Active and Unsupervised Learning of Spoken Words through A Multimodal Interface, in *Proc. 13th IEEE Workshop Robot and Human Interactive Communication*, pp. 437–442 (2004)

[Myers 81] Myers, C. S. and Rabiner, L. R.: A Comparative Study of Several Dynamic Time-warping Algorithms for Connected Word Recognition, *The Bell System Technical Journal*, Vol. 60, No. 7, pp. 1389–1409 (1981)

[Oates 00] Oates, T., Eyler-Walker, Z., and Cohen, P. R.: Toward Natural Language Interfaces for Robotic Agents: Grounding Linguistic Meaning in Sensors, in *Proceedings of the Fourth International Conference on Autonomous Agents*, pp. 227–228 (2000)

[Piaget 56] Piaget, J. and Inhelder, B.: *The Child's Conception of Space*, Routledge and Kegan Paul, London (1956)

[Regier 96] Regier, T.: *The Human Semantic Potential: Spatial Language and Constrained Connectionism*, MIT Press, MA (1996)

[Roy 02] Roy, D. and Pentland, A.: Learning Words from Sights and Sounds: A Computational Model, *Cognitive Science*, Vol. 26, No. 1, pp. 113–146 (2002)

[Roy 04] Roy, D., Hsiao, K., and Mavridis, N.: Mental Imagery for a Conversational Robot, *IEEE Transactions On Systems, Man, And Cybernetics-PART B: Cybernetics*, Vol. 34, No. 3, pp. 1374–1383 (2004)

[Shen 06] Shen, F. and Hasegawa, O.: An Incremental Network for On-line Unsupervised Classification and Topology Learning, *Neural Networks*, Vol. 19, No. 1, pp. 90–106

(2006)
[Siskind 96] Siskind, J.: A Computational Study of Cross-situational Techniques for Learning Word-to-meaning Mappings, *Cognition*, Vol. 61, No. 1-2, pp. 1–38 (1996)
[Siskind 00] Siskind, J.: Learning Word-to-Meaning Mappings, in Broeder, P. and Murre, J. eds., *Models of Language Acquisition: Inductive and Deductive Approaches*, pp. 121–153, Oxford University Press (2000)
[Steels 97] Steels, L. and Vogt, P.: Grounding Adaptive Language Games in Robotic Agents, in Harvey, I. and Husbands, P. eds., *ECAL97*, Cambridge, MA (1997), MIT Press
[Steels 02] Steels, L., Kaplan, F., McIntyre, A., and Van Looveren, J.: Crucial factors in the origins of word-meaning, in Wray, A. ed., *The Transition to Language*, chapter 12, pp. 252–271, Oxford University Press, Oxford, UK (2002)
[Steels 03] Steels, L. and Baillie, J.-C.: Shared Grounding of Event Descriptions by Autonomous Robots, *Robotics and Autonomous Systems*, Vol. 43, No. 2-3, pp. 163–173 (2003)
[Thrun 95] Thrun, S. and Mitchell, T.: Learning One More Thing, in *Proceedings of IJCAI*, Montreal (1995)
[Vogt 05] Vogt, P.: The Emergence of Compositional Structures in Perceptually Grounded Language Games, *Artificial Intelligence*, Vol. 167, No. 1-2, pp. 206–242 (2005)
[Wachsmuth 00] Wachsmuth, S., Socher, G., Brandt-Pook, H., Kummert, F., and Sagerer, G.: Integration of Vision and Speech Understanding Using Bayesian Networks, *Videre: A Journal of Computer Vision Research*, Vol. 1, No. 4, pp. 61–83 (2000)
[Weng 01] Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E.: Autonomous Mental Development by Robots and Animals, *Science*, Vol. 291, No. 5504, pp. 599–600 (2001)
[Yu 04] Yu, C. and Ballard, D.: On the Integration of Grounding Language and Learning Objects, in *Nineteenth National Conference on Artificial Intelligence (AAAI '04)* (2004)
[中川 93] 中川聖一, 中西宏文, 古部好計, 板橋光義：視聴覚情報の統合化に基づく概念の獲得, 人工知能学会論文誌, Vol. 8, No. 4, pp. 449–508 (1993)

〔担当委員：佐藤 理史〕

2006 年 10 月 18 日　受理

## 著 者 紹 介

**岡田　将吾**

2003 年横浜国立大学工学部知能物理工学科卒業，2005 年 東京工業大学大学院総合理工学研究科　知能システム科学専攻修士課程修了．現在，同大学博士課程在学中．

**賀　　小淵**

2002 年浙江大学 情報工学科卒業，2006 年 東京工業大学大学院総合理工学研究科　知能システム科学専攻修士課程修了．人工知能，ヒューマノイドロボット，言語獲得に関する研究に従事，現在，日本電気（株）勤務．

**小島　量**

2005 年 東京工業大学工学部情報工学科　計算工学専攻卒業，2005 年 東京工業大学大学院総合理工学研究科知能システム科学専攻修士課程修了．現在，みずほ証券（株）勤務．

**長谷川　修**（正会員）

1993 年東京大学大学院電子工学専攻博士課程修了，博士（工学），同年電子技術総合研究所入所，1999 年から 1 年間米国カーネギーメロン大学客員研究員，2001 年産業技術総合研究所主任研究員，2002 年 東京工業大学像情報工学研究施設助教授，JST さきがけ研究 21 研究者（兼任）．