1B3-03

自己増殖型ニューラルネットを用いたヒューマノイドロボット上の発達的シンボルグラウンディング

Developmental Word Grounding through a Growing Neural Network with a Humanoid Robot

小島量[*1] 長谷川修[*1*2]
Ryo Kojima Osamu Hasegawa

[*1]東京工業大学 Tokyo Institute of Technology
[*2]科学技術振興機構さきがけ研究 21 PRESTO JST

This paper presents an unsupervised approach of integrating speech and visual information without using any prepared data, which enables a humanoid robot to learn words with their meanings. The approach is different from most other existing approaches in that it learns online from audio-visual input, rather than from stationary data provided in advance. In addition, it is capable of learning incrementally which is considered to be indispensable to lifelong learning. A noise-robust self-organized growing neural network 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 with the least information for subsequent learning. Experimental results show that it makes the learning process more efficient.

1. はじめに

近年，パーソナル 2 足歩行ロボットを様々な企業が発表するなど，実世界で人間と共存することを志向したロボットに関する研究が非常に盛んに行われている．環境から必要な情報を抽出し状況に応じて適切な処理を行うコンピュータやロボットは，近い将来に日常生活の場において身近な存在になるであろう．そしてその際，対話という手段がコンピュータと人間を繋ぐ主要なインターフェースになると考えられるが，言語を用いた自然な対話を行うためには，物事に対する共通の理解と認識が必要不可欠である．同じ言葉でも，その言葉の意味する物事が自己と他者の間で違ってくると，最早対話は成り立たない．そのためコンピュータと自然な対話を行うためには，物事の概念に関する理解を人間とコンピュータの間で一致させることが必要であると言える．

このような背景の下で，言語情報と映像情報の入力から，各単語が一体どのような意味を持つのかを説明するための言語認識モデルをコンピュータ上に構築する研究がいくつか行われている [Roy 00, Yu 04, Zhang 03]．また，ロボットと実世界で共存していくためには，ロボット自身が複雑な実世界環境において学習し続けることが必要不可欠であると考えられる．そのためロボットの学習アルゴリズムとしては，追加的かつリアルタイムに動作すること以外にも，他者とコミュニケーションを取りながら必要な物事だけ獲得していく形態が望まれる．ロボットを用い，人間とのインタラクションを通して物体の動きに関する概念を一から獲得する Iwahashi のシステムはこの要求を適えていると言えるが，前もって音声情報と映像情報の組み合わせをシステムに与える必要がある [Iwahashi 04]．提案システムは，視聴覚情報の統合を行う上で，

1. 学習データを前もって用意する必要は無く，知識が全く無い状態から一から物体の属性に関する概念 (色の概念，形状の概念，及び色と形状から得られる物体そのものの概念) を，リアルタイムに連続して獲得していく
2. 獲得した知識の確からしさをシステムが認識することが出来る

連絡先: 小島量，ryo@hi.pi.titech.ac.jp

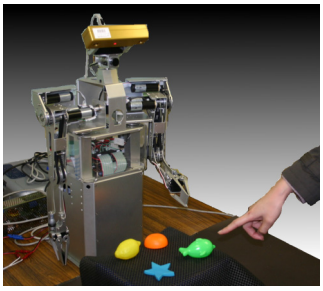

図 1: 対話的に学習するヒューマノイドロボット

3. 追加的な学習を可能にするため，ANG(Adaptive Neural Gas)[Shen 04] を利用している

という点で，他の手法とは異なっている．また，システムを実際にヒューマノイドロボットに搭載し，ロボットの身体性を利用したインタラクティブな学習を行っている．さらに，知識の獲得に対する積極性を利用することで学習の効率を高めることに成功した．

2. 提案システム

本研究ではアールラボ社製のヒューマノイドロボット H3 を利用した．このロボットは基本性能として片手で 5 自由度，首周りが 2 自由度の計 12 自由度を保持している．また，ステレオ入力が可能な画像入力装置を備えており，物体の空間上の相対位置等を測定できる．これらをサーバーとなる PC と接続することにより，音声の入出力も可能となり，視聴覚や身体性を用いたロボットの学習を実現している．図 1 は実験者が物体を指でさすと，ロボットもその物体に視線を向けて同様に指差しする様子を示している．このような協同注意の機能をロボットに組み込むことにより，より対話的な学習システムが実現できると考えられる．

提案システムは図 2 のように，ロボットと人間と物体の三者の関係によって成り立っている．ここで重要なのはロボット

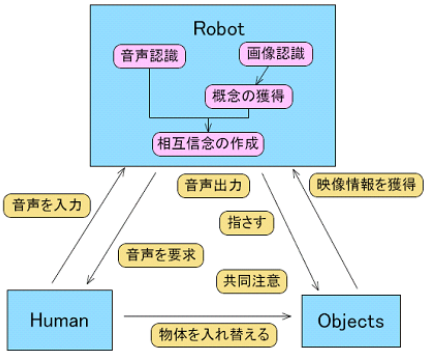

図 2: システムの全体像

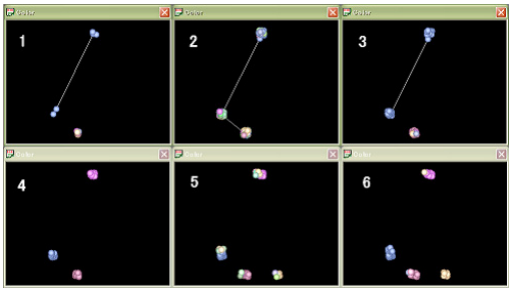

図 3: 概念の形成 (初めは三種類の色を見せ，5 で色を一つ追加した)

も物体の存在を認識していることである．そして，これら三者が，相互に影響を与えながら学習が以下のように進行していく．

まず，実験者はロボットの目の前に物体をいくつか置く．ロボットはステレオカメラを用いて映像情報から物体を切り出し，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物体そのものを意味する 11 次元のベクトルを抽出する．本システムはリアルタイムに稼動し，一秒間に各物体あたり 10 から 20 個のベクトルがそれぞれ算出される．これらのベクトルを，ANG を用いてクラスタリングを行う．ANG は，ニューロンが自己増殖しながら入力ベクトルを連続的に近似し分類することにより，追加的にクラス数を増やしていくことが可能であり，教師なしクラスタリング手法の一つと見なすことができる．またノイズに対する耐性を有しており，例えば実験の途中で対象物体を入れ替える際に一瞬手がカメラに写ってもそれをノイズと認識するため，連続的な実験が可能である．ここで，クラスタリングされた入力ベクトルは “概念”と捉えることができる．このようにして人間からの一切の入力なしに，視覚情報からのロボットによる概念の自己組織的形成が行われる．なお, 図 3 は，初めにそれぞれ色が異なる 3 種類の物体を見せ，しばらくしてから新しい色を持つ物体を一つ追加したときの，時間の経過と共に概念の形成が進んでいく様子を表したものである．図中の 4 の段階で三つの概念が形成され，図中の 5,6 では追加的に四つ目の概念の形成がなされているのが確認できる．

映像情報から得られる概念と音声が結合すれば，人間とロボットとの間に共通の “理解”が生まれると考えられる．実験者は図 1 のようにロボットと向き合って特定の物体に対して音声を一つずつ入力していく．音声情報のクラスタリングは，元の音声情報をベクトル量子化し，得られたコードに対して動的計画法を用いることで実現される．また，本システムはベクトル量子化したコードから逆に音声を生成する機能を備えており，この機能を用いることによりロボットに発話をさせることが可能になる．

3. 視聴覚情報の統合

すべての概念は “確信度”という値によって過去に入力した音声との結合の度合いが定められる．また，ある物体が持ちうる三つの音声 (色，形状，及び物体そのものに関する音声) の確からしさを示す “既知度”は，この確信度をもとに計算される．ロボットは人間が指でさしている物体に関する音声を一方的に聞いて学習するだけではなく，ある物体に対して音声が入力された際に，目の前にある物体全部の既知度の増加分が最大となるような，そのような物体を手や視線を向けることにより選択し，人間に向かって情報を要求することができる．このような “知識欲”を持ったロボットとインタラクティブに学習を進めることにより，円滑に効率よく概念と結び付いた言語の獲得が進むものと考えられる．本章ではこれらのシステムを実現するアルゴリズムについて説明する．

3.1 確信度，既知度の定義

ANG を構成する各ニューロンは確信度 $\mathbf{c}$ という値によって各音声との結合具合が決定される．i 番目のニューロンの確信度 $\mathbf{c}_i$ は入力音声のクラス数が j 個あるとき，

$$\mathbf{c}_i = \{d_{i1}, d_{i2}, ..., d_{ij}\} \tag{1}$$

で定義される．$d_{ik}(1 \leq k \leq j)$ は 0 以上の実数値をとり，この値によって具体的に各音声との結合具合が定められる．d_{ik} の値が大きければ大きいほど，このニューロンは k 番目のクラスに所属する音声と結合の具合が強いと言える．また，未知の音声が入力されるたびに，$\mathbf{c}$ の要素の数は一つずつ増加していく．d_{ik} の初期値は 0 である．さらに，各概念に対しても確信度 $\mathbf{cc}$ を持たせることで，概念と音声の結合を図る．i 番目のニューロンを n_i，p 番目の概念 C_p に所属するニューロンの数を a_p とすると，概念 C_p の確信度 $\mathbf{cc}_p$ は，

$$\mathbf{cc}_p = \frac{1}{a_p} \sum_{n_i \in C_p} \mathbf{c}_i \tag{2}$$

$$= \{\tilde{d}_{p1}, \tilde{d}_{p2}, ..., \tilde{d}_{pj}\} \tag{3}$$

で表される．

また，各概念の既知度 k_p を次式で定義する．

$$k_p = \frac{\max(\mathbf{cc}_p)^2}{\sum_i (\tilde{d}_{pi})^2} \tag{4}$$

そして，物体の既知度 ok は，その物体の持つ色の概念の既知度，形状の概念の既知度，及び物体そのものの概念の既知度から求められる．

3.2 音声を入力した際の確信度の変化

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る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)．まず最初に，物体から色，形状，及び物体そのもの

を意味する三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概念を最近傍決定則を用いてそれぞれ求める．このとき，各空間内に概念が一つも存在しなかったらその属性に関しては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ン n の確信度の k 番目の要素を次式のように 1 だけ増やす．

$$d_{ik} := d_{ik} + 1 \ (\forall i\{n_i \in C\}) \tag{5}$$

3.3 確信度の伝播

もし，確信度がニューロン間で移動しないと，次の瞬間に確信度を持つニューロンが消滅して音声情報が失われる可能性がある．また，ニューロンの確信度を概念内で平均化すると，概念同士が一瞬でくっついて離れた際に音声情報が混同される恐れがある．そこで本システムにおいては，確信度がニューロン間を相互に少しずつ伝播していくモデルを考える．i 番目のニューロンを n_i，i 番目のニューロンに隣接するニューロンの集合を N_i，i 番目のニューロンに隣接するニューロンの数を a_i とする．また時刻 t における i 番目のニューロンの確信度を $\mathbf{c}_{i(t)}$，α を伝播の速度を表す定数とすると，伝播の式は，

$$\mathbf{c}_{i(t+1)} = (1 - \alpha a_i)\mathbf{c}_{i(t)} + \alpha \sum_{n_j \in N_i} \mathbf{c}_{j(t)} \tag{6}$$

で表される．実験では $\alpha = 0.01$ に設定した．

3.4 知識欲を利用した概念獲得

ロボットが積極的に学習に参加すればより効率的に学習が捗ると考えられる．提案システムでは，目の前にある各物体に対して既知度を測定し，音声を入力する物体 o_i を式 (7) に基づいてロボットに選択させるようにした．ただし，音声を s，音声のクラス数を j，i 番目の物体の既知度を ok_i とする．また，E は期待値，Δ は変化量を意味する関数である．言葉で説明すると，「ある物体に対して音声が入力された際に，目の前にある物体全部の既知度の増加分が最大となるような，そのような物体を選択する」という意味になる．

$$i \quad = \quad \arg_i \max \sum_l \mathrm{E}(\Delta(ok_l|o_i, s)) \tag{7}$$

4. 実験

4.1 追加学習に関する実験

本実験では，色が 9 種類，形状が 8 種類の合計 72 種類の物体を利用してこれらの物体の概念を順にロボットに獲得させていき，同時に音声を入力することで，提案システムにおいてリアルタイムで追加的なシンボルグラウンディング学習が可能であることを示す．実験に用いる物体の一例を図 4 に示す．実験ではまず最初に，72 種類の物体から色が 4 種類，形状が 3 種類の合計 12 種類の物体を選択し，順に四つずつ物体をロボットに見せ，概念の形成がなされるのに十分な時間が経過したら，色と形状と物体そのものを意味する音声を各物体に対して順に入力していく．この際，音声を入力する物体は人間が指差しして指定する．物体を何回か入れ替え，12 個すべての物体に対して音声の入力が済んだら，次に音声の結合精度を測る．これは，再びロボットの目の前に四つずつ物体を置き，そのすべての物体に対して，色と形状と物体そのものを意味する音声情報が正しく結合されているかを，発話という形でロボットに出力させて測定する．ここで，本来色を意味する音声を形状の音声として出力してもこれは間違いであるとみなす．また，

図 4: 追加学習に関する実験で用いた物体の例

間違った音声を出力したら，その物体に対して再び正しい音声を入力し直す．音声の再入力が終わったら物体の数を色が 5 種類，形状が 4 種類の計 20 種類に増やし，追加した 8 種類の物体に対してのみ最初の 12 種類の物体に対して行ったのと同様の方法で音声の入力を行う．そして，今度は 20 種類すべての物体に対して音声の結合精度を測り，同様に間違った出力を行った物体に対してのみ音声の再入力を行う．以降，順に物体の数を色 6 種類形状 5 種類の計 30 種類，色 7 種類形状 6 種類の計 42 種類，色 8 種類形状 7 種類の計 56 種類，色 9 種類形状 8 種類の計 72 種類と増やしていき，同様の測定を行った．

4.2 追加学習に関する実験の結果

実験の結果を表 1 に示す．表中の値は各回の測定において，色，形状，及び物体そのものを意味する音声が正しく出力出来ている確率を百分率で表している．物体の色と形状においては，確実に音声との統合がなされているのが認められる．また，物体そのものについても高い確率で音声との統合がなされているのが確認できる．このように提案システムでは，認識の精度を落とさずに追加的な学習が可能であると考えられるが，当実験は初めから終わりまで連続して行われたということも別に強調しておきたい．本システムは，様々なノイズ (代表的なものとしては物体を入れ換える際の人間の手が考えられる) がロボットの視界に入っているのにも係わらず，正確に動作し続けることができる．これは必要な情報と不要な情報を自動的に分別する能力，言わば情報の取捨選択能力であり，こうした能力は今後実世界で人間と共存するロボットを製作する上で必要不可欠な能力であると著者らは考える．

4.3 知識欲の効果に関する実験

3.4 で述べた “知識欲” の効果性に関する実験を行った．“知識欲” を利用することで学習効率がどれくらい向上するかを測定する．使用した物体は色が 4 種類，形状が 4 種類の合計 16 種類で，これらの物体の内からランダムで 4 種類の物体を選択しロボットの目の前に置く．そして，ロボットの目の前に置いた四つの物体から，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．

A) 既知度に基づいてロボットに選択させる

B) ランダムに選択させる

なお，物体を既知度に基づいてロボットに選択させるときは，対象の物体をロボットが指差しするようにした．音声を入力する物体の選択が済んだら，次にその物体の色，形状，及び物体そのものを意味する音声を入力する．そして音声の入力が完了したら，四つの物体を一旦ロボットの視界の外に戻す．以上，4 種類の物体をランダムに選択するところから戻すところまでの操作を 1 サイクルと考え，全部で 16 サイクル操作を

	一回目	二回目	三回目	四回目	五回目	六回目
物体の数	12 個	20 個	30 個	42 個	56 個	72 個
色	100	100	100	100	100	100
形	100	100	100	100	100	100
物体	100	95.00	100	95.24	98.21	95.83

表 1: 音声結合率 (%) 物体の数のうちどれだけの割合で正しい音声を出力できたのかを表している

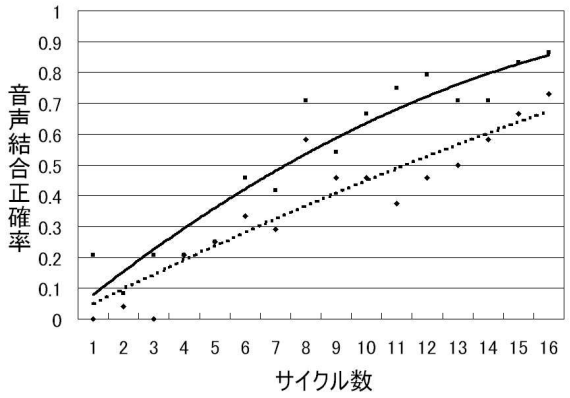

図 5: 音声結合正確率

行った．また，既知度及び音声結合正確率の測定を，ロボットの目の前に置いた四つの物体に対して音声を入力する前に毎回行った．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．なおここでの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，および物体そのものを意味する音声をロボットが発話という形で正確に出力できた確率のことをいう．

4.4 知識欲の効果に関する実験の結果

実験で測定した音声結合正確率の変化の具合を図 5 に表す．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 5 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大幅な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．さらに，ここで重要なのは，ロボットが教示してもらいたい物体を指差しする行為自体が，人間とロボットにおいて教示者と被教示者という立場の違いはあるにせよ，双方向の対等なコミュニケーションを可能にしているという点である．実験でも利用されたが，本システムには，人間が指差しした物体に関する音声をロボットが発話という形で出力したり，人間が発話した音声を持つ物体をロボットが指差ししたりする機能を備えている．しかしながらこれらの機能においては，常に人間側からのアクションが先にあり，ロボット側から積極的に行動するということはない．本システムでは，そこにロボットによる指差しの機能を持たせることでロボットにも積極性を備えさせ，“より自然な”学習を行うことを試みた．人間の幼児の発達段階における指差しの重要性は指摘されているが，本論文では今後のロボットの知能学習において，提案システムに見られるような人間とロボットの双方向的なコミュニケーションの重要性を主張しておきたい．

5. おわりに

本稿では，人間とロボットの間で概念の共有化を行うことを目的とした，新たな学習システムを提案した．提案手法は完全に追加的でなおかつリアルタイムに動作し，さらにヒューマノイドロボットの身体性を活かしたインタラクティブな学習メカニズムを持つ．そのため，実世界に存在する物体の概念を自律的かつ追加的に獲得していき，これと言語を結び付けることによって，人間との間に共通の認識を得ることが可能となる．実験では，実際に色や形状や物体そのものの概念を，人間とのコミュニケーションを通じてヒューマノイドロボットに獲得させた．また，ロボットから人間に積極的に情報を求めることが学習効率の大幅な向上に役立つことも示した．

今後の課題としては，

1. 動き等の色や形状以外の属性に関する概念の獲得
2. 文法の獲得と利用
3. 概念の持つ構造の発見と利用
4. 美しさといった定量的ではない概念の獲得

などが挙げられる．

参考文献

[Iwahashi 04] Iwahashi, N.: Active and unsupervised learning of spoken words through a multimodal interface, in *Proceedings of 13th IEEE Workshop Robot and Human Interface Communication*, pp. 437–442 (2004)

[Roy 00] Roy, D.: Integration of speech and vision using mutual information, in *Proceedings of Internationl Conference on Acoustics, Speech and Signal Processing*, pp. 2369–2372 (2000)

[Shen 04] Shen, F. and Hasegawa, O.: A growing neural net-work for online unsupervised learning, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 8, No. 2, pp. 121–129 (2004)

[Yu 04] Yu, C. and Ballard, D.: On the integrarion of grounding language and learning objects, in *Nineteenth National Conference on Artificial Intelligence(AAAI '04)* (2004)

[Zhang 03] Zhang, Y. and Weng, J.: Conjunctive visual and auditory development via real-time dialigue, in *Proceedings of the Third International Workshop on Epigenetic Robotics(EpiRob2003)*, pp. 147–154 (2003)