Vol.2013-ICS-170 No.4
2013/3/10

シンボリック情報を用いない画像・音声刺激の同時提示性に基づく言語シンボル概念獲得

岩本 悟[1,a)]　荒井 秀一[1]

概要：従来のシンボルに基づいた知識獲得の枠組みでは予め用意されたシンボルを用いるため獲得できる知識が限定される．従って，獲得された知識は特定のタスクに大きく依存してしまう．そこで，我々は同時に提示された画像・音声メディア間に何らかの関係があると考え，事例同士をリンクし構築するネットワークを用いて，事例共通の特徴を取り出すことでタスクに依存しない抽象度の高い知識として”概念”を獲得する枠組みを提案する．

キーワード：概念獲得，学習，シンボルグラウンディング問題

1. はじめに

現在までに行われている知識獲得の研究の多くは，シンボルを用いてより複雑なシンボルの意味を獲得していくアプローチがとられている．このアプローチは，予めシンボルの意味や関係を記述しておくことで，推論が可能となり，人間の論理的思考を模倣することができるため，多くの研究で成果を挙げてきた [1]．メディアの認識・理解を行う際にはこのアプローチをとる場合が多く，画像認識・理解の分野では，シンボルによって物体間の関係の記述や，予め用意したプリミティブな形状により物体を記述することで認識・理解を行う研究が行われてきた [2]．しかし，学習によって新たに知識を獲得しようとする際には，知識が予め用意されたシンボルでしか記述ができないため，獲得できる知識が限定されてしまう問題があった．そこで物体認識の分野では，大量の画像群から抽出した局所特徴量をクラスタリングすることで物体を表現することができるクラスタ群，すなわち，シンボルセットを生成するアプローチをとっている [3]．これらの研究では，生成したシンボルセットの出現頻度を物体の知識として学習することで，対象に依存しない一般物体認識を実現している．しかし，この枠組みではシンボルセットの生成に用いる特徴のみでしか知識を記述することができないため，特定のタスクに依存した知識しか得ることができない．従って，時々刻々と変化する実環境に対応することはできない．

[1] 東京都市大学
Tokyo City University, Tamazutsumi 1-28-1, Setagaya-ku, Tokyo 158–8557, Japan
[a)] iwamoto10@ipl.cs.tcu.ac.jp

このような問題を解決するためには，タスクに応じて獲得した知識の変更・生成を行って知識の再利用が行えるような，抽象度の高い知識が獲得できなければならないと考えた．そこで我々は，知識をシンボルによって直接記述せず，後に定量的解析が可能でシンボルが示す事物そのもののパターンとパターン間の関係を記述することで，計算機に抽象度の高い”言語シンボル概念”を獲得させることができる枠組みを提案する．

2. 言語シンボル概念

一般に概念とは，”経験される多くの事物に共通の内容を取り出し (抽象)，個々の事物のみに属する偶発的な性質を捨てる (捨象) ことによって形成されるもの”である [4]．我々がコミュニケーションに用いている言語シンボルは，概念に結び付けられたラベルのようなものであり，言語シンボルの有無にかかわらず，事物を認識・理解できることから概念に先立って学習されるものではないと考えられる．このことから，我々は概念を実世界に存在する刺激群から共通の特徴を取り出すことで獲得できると考えた．しかし一般に，単一刺激であっても複数の概念が内包されているため，提示された刺激がどのような意味を持つのかはわからない．例えば，リンゴの画像刺激が単体で提示されていても，色の”赤”という概念を示しているのか，形状の”丸い”という概念を示しているのかは特定できない．メディア刺激はあくまで概念を伝達するための媒体であり，刺激そのもので言語シンボル概念を表現することはできないのである．この問題は一般に指示の不可測性 [5] としても知られている．

そこで本研究では，**図 1** に示す画像刺激と音声刺激の異種複数メディアと同時に与えられた刺激は何らかの関係を持つという同時提示性を利用し学習した画像・音声知識群とその間のリンクを言語シンボル概念と定義した．

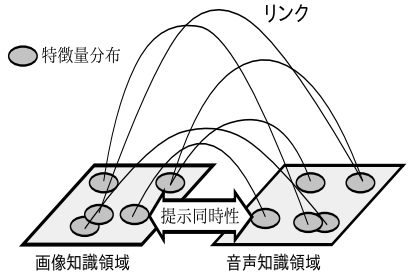

図 1 提示同時性による言語シンボル概念の概略図

3. 言語シンボル概念獲得モデル

2 章で述べた言語シンボル概念を獲得するにあたって，本研究では，画像と音声とその関係を直接記述するのではなく，母が子に絵本を見せるように，画像中の特定の部分の指示とその部位を意味する音声を示し発話することで，計算機自ら知識を学習し，言語シンボル概念を獲得できるような枠組みとする．しかし，このような枠組みを計算機上に実現するにあたり，先験的知識を全く用いずに実現することは不可能である．そこで，言語シンボルの意味に関する知識を一切与えないという条件の下，人間が先天的に持っているであろう以下の 3 つの能力を知識として与えることにした．

(1) **発話理解の能力** 受容した刺激を抽象化し，自らの知識と照らし合わせることで認識する能力．
(2) **発話内容の学習能力** 受容した刺激を学習して，知識として蓄える能力．
(3) **発話の能力** 知識から具体的な刺激に具象化し，外部に発話する能力．

以上のことを踏まえ，我々は**図 2** に示す学習の枠組みを提案する．以下に，図中の処理について説明する．

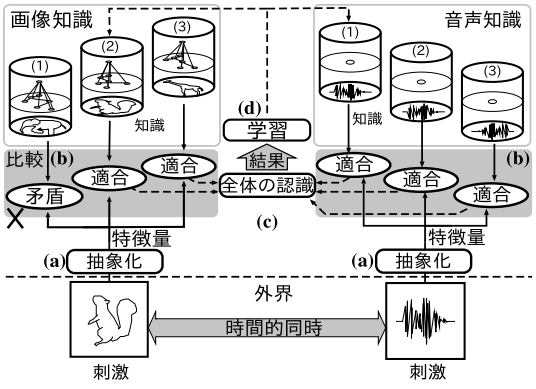

図 2 画像・音声知識学習の流れ

(a) 抽象化

言語シンボル概念は，学習した画像知識や音声知識に存在する共通の特徴を取り出すことで獲得できる．従って，画像・音声知識は言語シンボル概念が表す特徴を刺激から抽出できなければならない．ここで，人間が画像から抽出できる特徴を挙げると，物体に関しては形状，構造，色彩，テクスチャ，動き，3 次元情報等があり，更に物体間の関係を含めたシーン情報もある．これらの特徴を抽出する分析手法は数多く提案されているが，本研究では，画像理解を行う上で重要であると言われている物体の形状，構造特徴のみに関して扱うこととし，この 2 種類の特徴を十分表現し得ると考えられる線画像を用いる．

(b) 刺激の認識

抽象化した刺激と画像・音声知識を比較し，それぞれの知識にどの程度適合するかどうかを表す適合確率を算出する．認識結果を確率で表現することで，リンクの関係を含めた画像・音声刺激を総合的に認識する際に用いる認識結果候補の序列を求めることが可能となるだけではなく，異種メディア間の認識結果を直接比較することが可能となる．適合確率の算出法については，4.1 節で述べる．

しかし，メディア毎に算出された適合確率を用いて序列を定めてしまうと，知識適合確率が高い値を示す知識が複数ある場合，誤認識してしまう可能性がある．そこで，入力刺激と知識全体との適合確率を用いて事後確率を求め，これを知識群 $\mathbf{x}$ に含まれる知識 k と入力との確信度 $B(k|\mathbf{x})$ として，刺激の認識に用いることにした．式 (1) に示す確信度によって，知識全体を考慮した適合度である”確信度”によって序列を定めることが可能で，その結果異種メディアを考慮した言語シンボル概念の認識が可能となる．

$$B(k|\mathbf{x}) = \frac{P(k)}{\sum_{j=1}^{m} P(j)} \tag{1}$$

このとき，入力と知識 k の適合確率が $P(k)$，確信度の算出候補として採用されたカテゴリの総数が m である．

この際，全ての知識と適合しなかった場合には，未知の刺激であるとして，新規知識を生成する．同様に，比較可能な知識が全く存在しない場合にも新規の知識を生成することで，知識が何もない状態からの知識獲得を可能としている．

(c) 全体の認識

刺激を認識することで得た認識結果の候補群を用い，画像・音声間のリンクを考慮して，認識結果の判断を行う．認識結果としてあり得る 4 パターンの判断結果

を**図 3** と以下に示す.

(1) 画像・音声知識共に新規に作成された場合，初めて見聞きした概念であるためこの関係を認識結果とする.

(2) 画像・音声知識のいずれかが新規に作成された場合は，初めて見聞きした刺激と既知の知識の関係を示すものであるから，新規の知識と既知の知識側の最も確信度の高い候補との関係を認識結果とする.

(3) 画像・音声各々の確信度が非常に高い場合には，その間にリンクが存在しなくても，既知の知識同士の新たな関係であると考えられるため，これを認識結果とする.

(4) (1) から (3) 以外については，リンクの存在する組み合わせの中で式 (2) に示す総合確信度 B^T が最も高い関係を認識結果とする.

$$B^T = \sqrt{B^I B^S} \tag{2}$$

このとき，B^I は画像確信度,B^S は音声確信度で，これらの相乗平均を総合確信度 B^T とした.

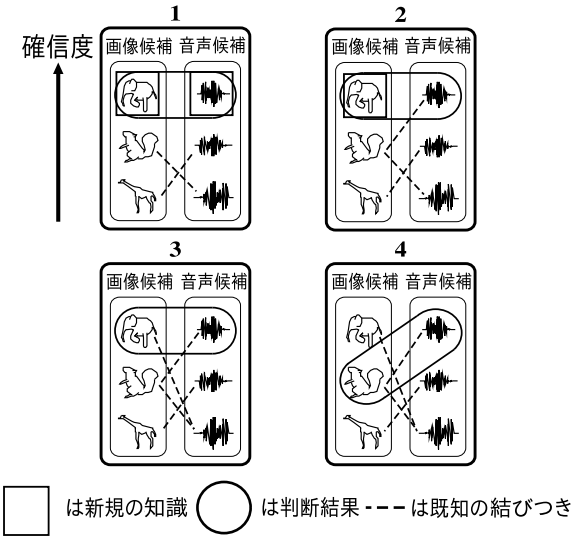

図 3 画像・音声知識間のリンクを考慮した総合的な判断の結果

(d) 学習

認識結果に基づいて，刺激そのものを画像・音声知識に，同時提示された刺激関係を知識間のリンクに強化学習を行う．既存のリンクが存在しない場合には，新たにリンクを張る．このように学習された知識は，次の入力刺激の認識に再利用される.

4. 知識の表現

3 章では，言語シンボル概念を獲得するために知識間の関係を学習する方法について述べた．本章では，実際に刺激として受容した画像・音声刺激をどのように知識として記述し，認識・学習を行うかについて述べる.

本研究では，知識の再利用性の観点から，知識や認識・学習の過程から一切のシンボリック情報を排除し，後に定量的解析が可能な刺激そのものを学習する.

4.1 画像知識

画像知識は，**図 4** のように物体の属性である形状・構造を認識・理解・学習に直接再利用可能な形態で記述する．ここで，人間が視覚から得られる画像を認識する過程を考えると，全体の印象の認識から始まり，明確に認識できない場合に細部の特徴を用いてトップダウンな認識処理を行っている．本研究では，この人間の認識過程に注目し，画像知識を階層的に記述しトップダウンな比較が行えるよう，以下の 4 つの表現により知識を階層的に記述した.

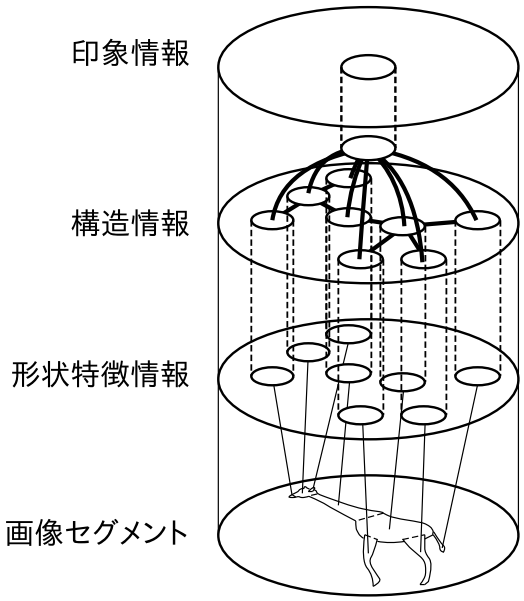

図 4 画像知識の階層構造

4.1.1 印象情報

画像を表現する情報の中でも最も抽象的な情報で，物体の輪郭線の接線の角度の差分から HMM を作成し，物体全体の概形を表現する．入力刺激との比較の際は，入力刺激に含まれる物体輪郭線の接線の角度の差分を用いて Viterbi アルゴリズムによって尤度を計算し，閾値以下であれば次の構造比較に進む．閾値以上であれば，これより先の比較は行われず，認識結果候補から外される.

4.1.2 構造情報

物体を部分領域に分割し，部分領域間の隣接・包含関係によって**図 5** のような構造情報を表現する．部分領域は，分割線を教示することによって学習し，物体の詳細な知識の記述を可能としている．入力刺激との比較の際は，入力刺激画像を知識の領域分割情報を用いて分割し，知識と構造の比較を行う．構造が一致しない場合は，認識結果候補から外される.

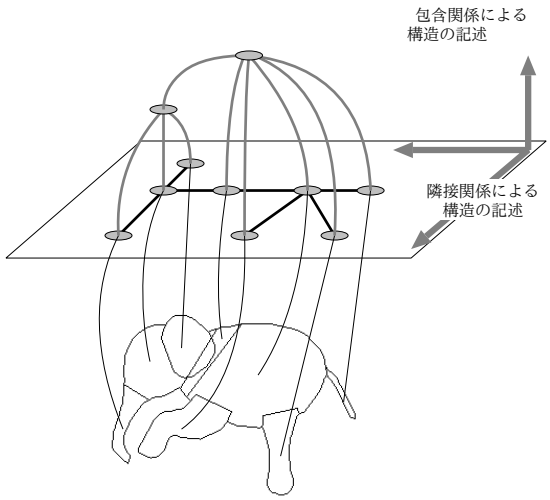

図 5 構造の記述の例

4.1.3 形状特徴情報

部分領域毎に**図 6** に示す画像処理で一般的に用いられる特徴量である，面積・骨格長・幅・周囲長・最外郭距離の 5 つの特徴量の多次元正規分布によって物体の詳細な形状を表現する．入力刺激との比較の際は，部分領域毎に知識側の分布の重心と入力刺激の特徴量とのマハラノビスの平方距離から知識適合確率を算出する．

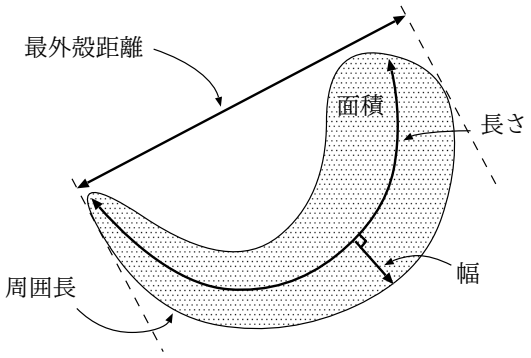

図 6 部分領域を表す特徴量

式 (3) で定義されている，マハラノビスの平方距離 D^2 は χ^2 分布に従うことが知られている．画像情報は 5 次元の形状特徴であるため，自由度 5 の χ^2 分布に従う．そこで，画像の知識適合確率として，χ^2 分布の上側確率を用いることにする．

$$D^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \tag{3}$$

このとき，$\mathbf{x}$ は入力刺激から抽出した 5 次元の特徴量，μ は知識側の部分領域の特徴量分布の重心，Σ は分散共分散行列である．画像知識全体との適合確率は，これら部分領域との適合確率の相乗平均で表すことにした．

4.1.4 セグメント画像

入力画像刺激を線分・円弧・角・分岐点の 4 つのセグメントによって表現することによって作成し，印象・構造・形状情報の生成に用いる．

4.2 音声知識

言語シンボル概念は，単一メディアの刺激のみからは獲得することができないため，本研究では，音声刺激も用いる．音声刺激からは，LPC ケプストラム 16 次元・パワー・ピッチ周波数・有声無声音の割合をフレーム毎に算出し，HMM を作成することで知識を表現する．

5. 言語シンボル概念の獲得実験及び考察

本節では，ここまで述べてきた言語シンボル概念獲得のための枠組みを用いて，本研究の目的である言語シンボル概念が獲得できているかどうかを評価する．本稿では，256 カテゴリ 30607 枚の画像群からなる Caltech-256[6] より，dolphin，hummingbird，leopards，syringe の 4 カテゴリから 5 枚ずつ，合計 20 枚の学習を行った．本研究では，形状・構造を示す言語シンボルの概念を獲得することが目的であるため，これらの画像から物体の輪郭抽出した**図 7** に示すような線画像を用いる．また，形状と領域の連結関係が重要であるため，他の領域によって隠蔽されることによって，その形状や連結関係が不明確になる領域は物体から除いた．これらの画像を概念が未知な言語シンボルの音声とともに，**図 8** に示す領域に対応させ教示した．なお，本実験では形状を表現する入力刺激を与えているが，この基準は実際の尺度を考慮したものではなく，画像中で教示者が主観的に感じた部分に対して教示したものである．

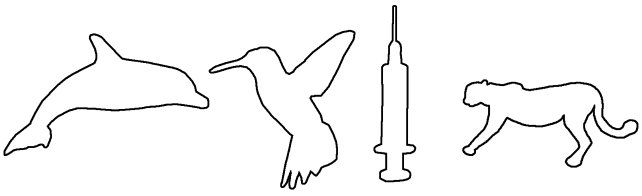

図 7 入力画像の例

5.1 形状に関する概念の評価法

知識内において，ある音声知識と関連付けられている画像領域集合は，その音声が示す形状概念を表していると考えた．しかし，人間が形状を評価するとき，**表 ??**のように必ず何らかの基準に対して行う．つまり，たとえ同一形状概念であっても基準が変化することによって異なった評価を下すと考えられる．従って，音声知識が示す概念情報は基準を規定しない限り，一意に求めることは困難であると考えた．そこで，本研究では基準となる領域の集合から

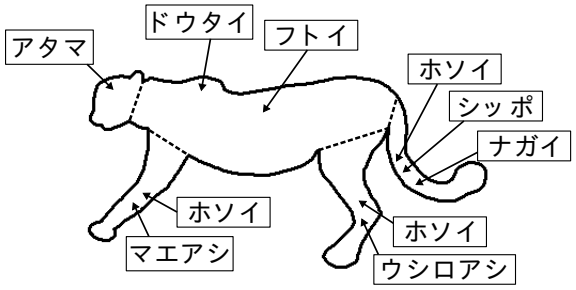

図 8 教示した画像・音声の例

表 1 形状概念とその基準
Table 1 Shape Concept and Standard.

基準	形状表現
一般的に	長い
耳としては	長い
ウサギの耳としては	長い

形成される分布によって，対象となる概念の領域の集合から形成される分布を正規化することにより得られる形状の特異性を表す分布によって音声が示す概念の形状的特徴を表す．具体的には，原点からその分布の重心までのマハラノビス距離を形状依存度とすることで獲得された概念の評価を行う．

5.2 形状を示す概念の獲得実験

実験結果として，構築された各音声が表現する形状概念の重心位置と，形状依存度を示すマハラノビス距離を各形状特徴量毎に**表 ??**に示す．

”クチバシ”の形状依存度が高いが，これは”クチバシ”が示す領域がどれも同じような形状をしているためである．

以上より，獲得した知識を再利用し，形状概念を獲得することが可能であることを確認し，モデルの妥当性を示した．

(実験に用いる画像 20 枚全てを学習し，教示したシンボルに結びつけられた分布群を統合し，統合した分布の重心とその分布の特異性を示す．)

6. おわりに

本研究では，同時に提示された画像・音声メディア間に何らかの関係があると考え，事例同士をリンクすることで構築したネットワークを用いて，事例共通の特徴を取り出すことでタスクに依存しない抽象度の高い知識として”概念”を獲得する枠組みを提案した．この枠組みが，概念獲得に関する研究分野で未だに解き明かされていない，概念獲得の初期段階におけるシンボルグラウンディング問題に対する一解釈を与える概念獲得モデルであり，言語に関する知識が全く存在しない場合にも言語シンボルの学習が可能であることを実験的に示した．

参考文献

[1] 森 英悟 , 荒木 健治 , 宮永 喜一 , 栃内 香次, ”自然言語：意味構造対応ルールの獲得と適用”, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション 95(321), pp.17-22, 1995-10-20.
[2] David Waltz, ”Understanding Line Drawings of Scenes with Shadows”, The Psychology of Computer Vision, pp.19-91,1975.
[3] Hao Zhang, ”SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition”, Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, vol.2, pp.2126-2136, 2006.
[4] 新村：”広辞苑 第五版”，岩波書店
[5] W.V.O.Quine，大出・宮館訳：”ことばと対象”，勁草書房
[6] G. Griffin, A. Holub, and P. Perona. Caltech 256 object category dataset. Technical Report UCB/CSD-04-1366, California Institute of Technology, 2007.

表 2 形状概念とその基準
Table 2 Shape Concept and Standard.

音声知識	面積	長さ	幅	最外郭距離	周囲長	形状依存度
アタマ	-0.175	0.079	-0.033	0.223	0.003	0.314
チイサイ	-1.713	0.344	0.368	-0.616	0.044	21.261
ミジカイ	-0.684	0.137	0.310	0.106	-0.459	4.188
ナガイ	0.991	-0.069	-0.176	-0.053	-0.024	4.596
クチバシ	-0.418	-1.126	0.004	0.742	0.158	110.581
ドウタイ	0.282	0.697	-0.742	0.152	0.344	8.004
ホソイ	-0.712	-0.259	0.175	-0.024	-0.138	1.765