2B2-03

視聴覚情報に基づく概念構造の対話的獲得

Interactive acquisition of conceptual structure based on audio-visual information

山田 大輔[*1] Daisuke Yamada
長谷川 修[*2] Osamu Hasegawa

[*1]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 Computational Intelligence and Systems Science, Tokyo Institute of Technology

[*2]東京工業大学大学院理工学研究科像情報工学研究施設 / 科学技術振興機構 さきがけ研究21
Imaging Science and Engineering Lab., Tokyo Institute of Technology / PRESTO, JST

This paper proposes an algorithm for interactive acquisition of conceptual structure about the objects in the real world based on audio-visual information. The system employs an interaction with a user to acquire conceptual structure. It guesses the conceptual structure by the given audio-visual data, and makes questions to a user about ambiguous points. By this, efficient interactive learning is achieved.

1. 研究の目的と背景

近年、人間とコミュニケーションを取りながら、日常生活をサポートする人間共存型ロボットの研究が盛んである [Roy 02, 岩橋 03]。こうしたロボットにおいては、ロボットが使用される環境の想定は困難であり、事前知識や自律行動の埋め込みができないことが多い。そこで状況や文脈に応じてユーザから対話的に教示を受け、学習を進めるアプローチが有効となる。

ここで、一般に実環境中の対象に関する概念には階層構造があり、ロボットもそうした構造を把握できなければ学習した概念知識の有効活用は期待できない。また構造に関する情報をすべて受動的に教示により学習するのでは、学習の効率が悪い。

そこで本稿では、実環境中の対象の概念に関して視聴覚を通じた教示を受け、入力された視聴覚情報を相互に参照することにより、概念に含まれる階層構造を獲得するためのアルゴリズムを提案する。この際、システムはある程度の教示データから階層構造の推定し、推定結果の不確定な箇所から優先的にユーザに質問を発することによって、効率的な知識構造の獲得を図る。なお今回の実験では、システム構成の都合上、教示音声については手書き文字画像を用いてこれに代え、提案アルゴリズムの有効性を検証した。

2. 提案アルゴリズム

2.1 提案アルゴリズムの概要

まず学習対象の自律的な観察およびユーザからの教示を通じ、対象の「①画像」および「②画像＋教示音声のセット」(ペンの画像と「ペン」or「文房具」等。教示音声は実際には手書き文字画像。以下同様)を得る。ついで①の画像をクラスタリングする。ここで得た各画像クラスの平均特徴ベクトルと、②の画像の特徴ベクトルを比較する。比較の結果、最も類似した①の画像クラスに②の教示音声を割り振る。この後、すべての教示音声をクラスタリングする。

以上により、表1に示すような画像クラスと音声クラスとの初期の対応関係を得る。表1の例では、画像データは6クラスに、音声は7クラスに分かれている。また例えば、画像クラス Visual 1 に該当する音声クラス Audio A のデータ数は1となる。

この表を行方向に見たとき、数値の多い箇所は、その音声クラスと画像クラスが対応する可能性が高いと考えられる。一方、数値が表全体の平均値前後を持つものは、その音声クラスと対応する画像クラスの関係が曖昧(ノイズも加味するため)と判断できる。そうした箇所を、ユーザへの質問箇所とする。

以上のように表中の数値が「大きい or 小さい」箇所ほど情報の確度が高いとし、それぞれ相応の確信度を与えて知識構造の推定結果の信頼性の情報として利用する。

表1: 画像クラス, 音声クラスの対応例

Visual \ Audio	A	B	C	D	E	F	G
1	1	1	7	0	1	5	6
2	5	1	2	5	0	4	4
3	1	0	5	1	0	3	4
4	0	6	1	3	1	4	4
5	1	6	2	5	0	3	5
6	1	1	0	0	5	0	5

表1の対応関係を推定した後、各音声クラスが対象の個別の名称(ペンの画像に対し「ペン」等)であるのか、あるいは階層としてより上位の総称であるのか(同、「文房具」)を推定する。画像クラスと音声クラスが一対一に高い確信度で対応していれば、個別の名称と判定する。一つの音声クラスに複数の画像クラスが対応している場合は、先に述べた画像のクラスタリング時に得た、画像クラス間の「距離」を参照する。ここで画像クラス間が近ければ、それらの画像クラスを統合する。

以上により、(1)音声クラスが画像クラスの個別の名称を指すとするもの、(2)複数の画像クラスを統合して一つの音声クラスに対応させ、個別の名称とするもの、(3)一つの音声と複数の画像が対応しているとするもの、の三通りの分類を得る。(1)、(2)を階層構造の最下層とし、これらと(3)(より上位の層と考えられる)の対応関係を求めることにより、全体の構造を推定する。

2.2 提案アルゴリズムの有効性の検証

連絡先:山田大輔, 〒226-8503 横浜市緑区長津田町 4259 東京工業大学像情報工学研究施設 R2-52 長谷川研究室, Tel: 045-924-5180, Fax: 045-924-5175, yamada@isl.titech.ac.jp

図 1 に、提案アルゴリズムの有効性の検証に用いた概念の階層構造を示す。2 種類のクリップを用意し、それらの個別の名称を「 クリップα」、「 クリップβ」、それらの総称を「 クリップ」とした。別にペン、コップを用意し、それらの個別の名称を「 ペン」、「 コップ」とした。また「 クリップ」と「 ペン」の総称を「 文房具」とし、コップを含めた全ての対象の総称を「 物体」とした。

まず上記の対象の画像データをクラスタリングし、先述のように 6 クラスを得た（ 表1最左列）。ここで今回のデータでは、ペンの画像が Visual 1,3 の2クラスに、またクリップβの画像が Visual 4, 5 の2クラスに誤分類された。

ユーザの初期教示により得た「 画像＋教示音声」の画像の特徴ベクトルを、クラスタリングにより得た各画像クラスの特徴と比較し、最も近い画像クラスに教示音声を割振った。この後、教示音声をクラスタリングした。以上の結果、表1の対応関係を得た。

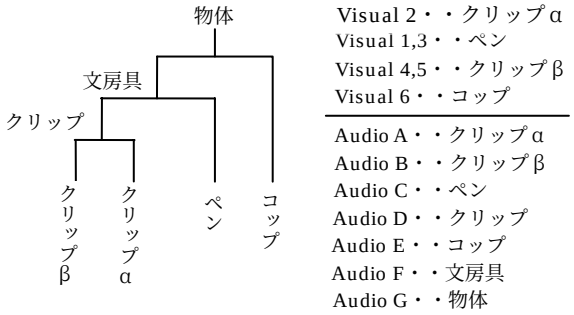

図 1: 検証データの知識の階層構造（左）と，その画像・音声データのクラスタリング結果（右）

表1において、各項目の数値の平均値は 2.6 であった。これと列方向（ 音声基準）に各項目の数値を比較し、例えば Audio A は、Visual 2 に対応する可能性が高く、Visual 1, 3, 4, 5, 6 に対応する可能性は低いとした。Audio D においては、Visual 2, 5 に対応する可能性が高いが、Visual 4 は判断が難しいとした。提案アルゴリズムでは、こうした平均値に対する各項目の数値を確信度とし、最終的な評価に用いている。以上から、対応の可能性に応じた Audio, Visual の関係表を新たに得た（ 表2 ，可能性高: 白、低: 黒、曖昧: 灰色にて表現）。

表2: 画像と教示音声間の対応関係の確信度

Audio

Visual	A	B	C	D	E	F	G
1	■	■	□	■	■	□	□
2	□	■	▒	□	■	□	□
3	■	■	□	■	■	▒	□
4	■	□	■	▒	■	□	□
5	■	□	▒	□	■	▒	□
6	■	■	■	■	□	■	□

表 2 を作成後、灰色の部分（ 曖昧な部分）を質問（ 対話）により解消させた。すなわち、表 2 の灰色の部分を白か黒に確定させた。例えば Audio D は Visual 4 の名称でもあるため（ 図1参照）、質問により灰色の部分は白となった。

表2の曖昧性を解消後、その音声が個別の名称であるか、総称であるかを調べた。Audio A は Visual 2 のみが白色になっていることから、Visual 2 の個別の名称であると判定した。Audio B には Visual 4, 5 が該当したが、Visual 4, 5 の両クラス間の距離を参照すると非常に近いので統合し、Audio B は Visual 4･ 5 の個別の名称と判定した。Audio E は Visual 6 と一対一に確信度高く対応しているため、Audio E を Visual 6 の個別の名称と判定した。

以上の「 Audio A - Visual 2」、「 Audio B - Visual 4･ 5」、「 Audio E - Visual 6」を構造の最下層とし、より上位の構造の推定を行った。具体的には、一つの音声に複数の画像が対応しているものを、これら最下層の対応関係を用いて書き換えた。表3にその結果を示す。表3において、個別の名称（ 最下層）にのみ配色しており、これにより、より上位の層は複数の配色の組み合わせとなる。すなわち、「 Audio D は Audio A と Audio B の上位構造」、「 Audio F は Audio C と Audio D の上位構造」、また「 Audio G は、Audio E と Audio F の上位構造」となった。

表3: 検証データの階層構造の推定結果（ 1 ）

Audio

Visual	A	B	C	D	E	F	G
1			■			■	■
2	■			■		■	■
3			■			■	■
4,5		■		■		■	■
6					■		■

図2に、表3中に含まれる階層構造を樹形図形式にて示す。

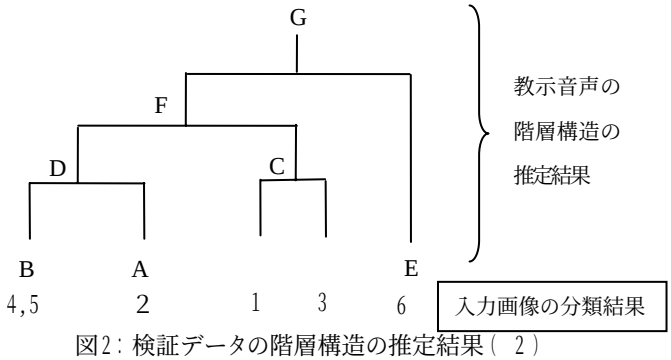

図2: 検証データの階層構造の推定結果（ 2 ）

3. 考察と課題

図1と図2を比較すると、ペンに対応する部分が分離した。この種のエラーは、本研究では曖昧性を含むパターン情報からの概念構造の推定を行っているために生じたものであるが、今後、こうした箇所のより正しい推測（ 修正）を可能とするアルゴリズムの検討を進める。また今回のデータでは、音声データの誤分類は生じなかったが、画像と音声の双方のデータのクラスタリングが正しく行われなかった場合についても検討する。

[Roy 02] Deb Roy: “A Trainable Spoken Language Understanding System for Visual Object Selection”, Proc. of the Int’l Conference of Spoken Language Processing, (2002)

[岩橋 03] 岩橋直人: “ロボットによる言語獲得”, 人工知能学会誌, vol.18, no.1, pp.49-58, (2003)