4Q−6 音声と画像の対応付けに基づく概念獲得アルゴリズム

古部好計†, 中川聖一†, 中西宏文‡
†豊橋技術科学大学, ‡愛知教育大学

1 まえがき

近年になって、言語学習の研究において、言語入力と非言語入力を同時に与えることにより、効率的に言語、および、その意味的な内容を学習するシステムがSiskindらによって提案されてきた[1]。しかし、これらは非言語入力が一階述語論理の形式であり、実際に非言語的な情報を入力として与えているわけではない。

本研究では、実際に音声と画像を与えることにより、その音声が意味する画像の内容を概念として獲得するシステムを作成した。以前開発したシステム[2]では、音声や画像の特徴抽出ミスに関して、あまり考慮していなかった。本稿では、新しいアルゴリズムについて述べる。本システムは、画像処理部、音声処理部、概念学習部の三つの主要な要素からなる。本システムは、初期知識を全く持たない状態から学習を始める。入力情報として、画像と、画像中の図形を表す音声の対を次々に入力していく過程で、概念を獲得する。なお、本研究が対象としている領域は、二次元図形画像である。そして獲得対象は、図形の形状、大きさ、色、そして、位置についての概念である。

2 画像処理[2]

画像処理部では、ビデオカメラから入力した画像を処理し、上記の概念を獲得するために必要なパラメータを抽出する。具体的には、図形の形状を表すパラメータとしてその図形の輪郭が持つ屈曲点の数を抽出する。大きさについては図形の面積をパラメータとする。色については図形のグレーレベルを、そして、位置については図形の座標をパラメータとして求める。

3 音声処理

音声処理部では、マイクから入力した音声から10ms毎に10次元のメルケプストラム係数を抽出する。音声同士のマッチングにはDPマッチングにより行う。DPマッチングにより二つの音声から共通区間を抽出する際の基準としては、DPパスの線形性、平均距離、そして、距離の最大値を用いた。また、辞書に登録された音声を入力音声からスポッティングする機能もある。

4 概念獲得

画像処理部と音声処理部から得た情報により、学習部で概念を獲得する。学習部の基本的な操作は、二つの概念の一般化である。二つの概念の間の共通部を抽出し新たに辞書に加える処理(一般化操作)を繰り返すことにより、階層的な構造の辞書を生成する。その階層構造の中から、最終的に出現回数の多いものが概念として獲得される。また、辞書操作として、出現回数の少ない辞書を削除する、つまり、忘却を仮定することで、入力の誤りに対処できる。

以下にアルゴリズムを簡略に記す。ただし、次の三辞書を定義しておく。

入力辞書 ... 入力の音声と画像をそのまま概念として登録する辞書である。入力辞書は、それに続くある一定回数の入力の間だけ維持され、その後削除される。

(階層)辞書 ... 階層構造の辞書であり、学習アルゴリズムが獲得した概念を格納する。ふたつの辞書の概念の関係が、一般的—特殊、であれば、そのふたつの辞書の間にリンクを張る。

一時辞書 ... 入力同士のマッチングにより獲得された共通部を概念として登録する辞書である。これは、その辞書を獲得するために用いた入力辞書がすべて削除された時点で削除される、一時的な辞書である。

<アルゴリズム>

1° 音声画像を入力し、入力辞書に加える。

2° (階層)辞書を検索し、入力中に辞書の概念が含まれていれば、その辞書を強化する。一時辞書の概念が含まれていれば、その一時辞書を辞書に変更する。辞書の強化は、辞書の出現回数を表すパラメータを増加させることで行なう。

3° 辞書と入力の間で一般化処理を行ない、冗長度の低い辞書を作成する。そして、元の辞書と新たにできた辞書との間にリンクを張る。

4° 3°の処理で、階層辞書のリンク関係が不足する場合があるので、リンク関係を補充する。

5° 入力と以前の入力とでマッチングをとり、共通区間を抽出する。そしてその共通区間を一時辞書に登録する。

6° 辞書の中で、出現回数の低いものなどを削除する。

7° 1°へ。

このアルゴリズムの第一の特徴は、辞書が階層構造を作りあげることである。正しい概念も誤った概念も同様に階層構造に現れる。しかし、通常は誤った概念の出現回数が正しい概念の出現回数より少ないと仮定すると、階層構造中の辞書の概念の出現回数が多いも

An Conceptual Acquisition Algorithm based on mapping between auditory and visual information
Yoshikazu KOBU†, Seiichi NAKAGAWA†, Hirobumi NAKANISHI‡
†TOYOHASHI University of Technology, ‡AICHI University of Education

のを概念として抽出することができる。二番目の特徴は、二つの入力の間でマッチングを取り、新たな辞書を作る時にその共通部のみを利用することである。その理由は、共通部の抽出で誤りが生じた場合、辞書全体に与える影響が大きいからである。そのため、本抽出アルゴリズムでは抽出すべき箇所がすべて正しく抽出されなくても、抽出されたものは必ず共通区間になるようしきい値を設定している。

5 実験および実験結果

本システムは大量の入力データを要するので、システムの評価実験として、とりあえず、音声の代わりに誤りを含むテキスト文を入力とするシミュレーションを行った。実験では、100文の入力を一セットとし、システムがどのような概念を獲得したのかを観察する。そして、その一セットの入力に含まれる概念の数と、獲得された正しい概念の数を比較することで評価した。

5.1 入力データ

シミュレーション用のデータとして、以下のような画像パラメータとテキストの対を100文入力した。

```
    ( 4 , 1 , 1 , 2 )   "さんかくけいが あります"
    ( 4 , 1 , 2 , 1 )   "くろい さんかくけいが あります"
*   ( 1 , 1 , 2 , 3 )   "くろい しかくけいが あります"
    ( 4 , 1 , 1 , 4 )   "うえに さんかくけいが あります"
                ⋮
```

*：誤りを含んだデータであることを示す。

画像パラメータは、左から順に、(屈曲点数+1, 大きさ, グレーレベル, 位置) となっている。大きさ、グレーレベル、位置、のパラメータは、それぞれの数値を量子化したものである。大きさについては、中=1、大=2、小=3となっている。グレーレベルは、白=1、黒=2となる。そして位置については、真中=1、右=2、左=3、上=4、下=5、となるように量子化した。

テキストは、平仮名の日本語を入力した。なお、テキストは、文節ごとに区切って入力した。テキストの半角スペースは、文節間の区切りである。

5.2 獲得概念

入力データの集計と、その入力より獲得された概念辞書を表1、2に示す。

このデータをまとめると、次のようになる。
入力中の (獲得することを希望する) 概念の総数：11
獲得された辞書の総数：22
希望する概念を含む辞書の数：9 / 22 = 0.41
正しい辞書の出現回数の総数/入力の出現回数の総数：123 / 220 = 0.56

100文の入力で、獲得を希望する概念がほぼ辞書として得られている。また、入力中の概念の約半数が辞書形成に活用されている。いくつかのバリエーションを変えた入力データについて実験したが、これとほぼ同様な結果が得られた。以上の結果から、このアルゴリズムに、入力データを多数与えることにより、概念が獲得できると思われる。

但し、全辞書のうちの約半数が不要な概念を含むものであるので、これらをより効率的に削減するストラテジーを考える必要がある。

表1 入力に存在する概念の集計

概念		出現回数	暗黙**	誤り
(1 , − , − , −)	"まる"	34	0	2
(4 , − , − , −)	"さんかくけい"	36	0	0
(5 , − , − , −)	"しかくけい"	25	0	3
(− , − , 1 , −)	"しろい"	14	36	2
(− , − , 2 , −)	"くろい"	33	14	1
(− , − , − , 1)	"まんなかに"	6	16	1
(− , − , − , 2)	"みぎに"	15	9	0
(− , − , − , 3)	"ひだりに"	13	6	0
(− , − , − , 4)	"うえに"	13	5	2
(− , − , − , 5)	"したに"	7	7	0
(− , − , 1 , −)	""	0	46	2
(− , − , 2 , −)	"おおきな"	15	16	0
(− , − , 3 , −)	"ちいさな"	9	11	1
		220	166	14

**：画像パラメータは存在するがテキストが存在しない。

表2 獲得された概念

	概念		出現回数	正しい辞書
LEVEL = 2***				
	(1 , 1 , 2 , −)	"まるが"	12	
	(− , − , − , 1)	"かくけいが"	9	
•	(− , − , 1 , −)	"しろい"	5	5
	(− , 1 , 2 , −)	"ろい"	4	
•	(5 , − , − , −)	"しかくけいが"	22	22
	(− , − , − , 3)	"かくけいが"	4	
•	(− , 2 , − , −)	"おおきな"	4	4
	(− , 1 , 1 , −)	"かくけいが"	4	
•	(− , 3 , − , −)	"ちいさな"	13	13
•	(− , − , − , 3)	"ひだりに"	1	1
	(1 , 2 , 2 , −)	"まるが"	3	
	(− , 2 , − , 2)	"みぎに"	1	
	(1 , − , 2 , 2)	"まるが"	1	
LEVEL = 3				
	(1 , − , 2 , −)	"まるが"	3	
•	(4 , − , − , −)	"さんかくけいが"	21	21
	(− , 2 , − , −)	"かくけいが"	18	
	(− , − , 1 , −)	"かくけいが"	20	
•	(− , − , 2 , −)	"くろい"	25	25
•	(− , − , − , 2)	"みぎに"	14	14
	(− , − , − , −)	"あります"	49	
	(1 , − , − , −)	"ろい"	12	
LEVEL = 4				
•	(1 , − , − , −)	"まるが"	18	18
				123

***：辞書の階層的なレベル。0、1は、入力辞書、一時辞書に割り当てているので、2からはじまっている。

6 むすび

結論として、画像と音声を対応付けながら学習を行う事により、ある概念に対応する音と画像特徴が対応付けられ、計算機で自動学習を行う可能性を示した。

今後の課題としては、実際に音声データを用いて評価することである。また位置や大きさの概念は、絶対的な概念のみしか取り扱っていないが、今後は相対概念に対処しなければならない。そして、動作の概念を獲得することも今後の課題である。

参考文献

[1] Jeffrey Mark Siskind. Acquiring core meanings of words. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp.143-156, 1990.

[2] 古部, 中西, 辰巳, 中川, 音声と画像の対応づけに基づく概念の獲得, 人工知能学会全国大会論文集, 1991.