情景と音声言語の混在情報から得た部分空間に基づくタスク推定

Task Identification Using Subspaces Designed by Mixed Information of Scenes and Spoken Words

澤田 心太[*1] 木村 優志[*1] 入部 百合絵[*2] 桂田 浩一[*1] 新田 恒雄[*1]
Shinta Sawada Masashi Kimura Yurie Iribe Kouichi Katsurada Tsuneo Nitta

[*1] 豊橋技術科学大学 大学院工学研究科
Graduate School of Engineering, Toyohashi University of Technology

[*2] 豊橋技術科学大学 情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology

In this paper, we propose a task identification method based on multiple subspaces extracted from mixed media of visual scenes and spoken language. The multiple subspaces are designed by using singular value decomposition (SVD). In the experiments, the frequencies of image objects are measured, as well as counting spoken words. To identify the task, we calculate similarities between an input scene and reference subspaces of different tasks. Experimental results show that the proposed method outperforms the method in which only language information is applied. Moreover, the proposed method achieved accurate performance even if less spoken language information is applied.

1. はじめに

近年，ロボットの研究・開発が盛んになり，様々な課題やその解決方法が議論されている．我々も，人間と共生可能なロボットを目標とし研究を行っている[新田 08]．研究の課題の一つに，多世界の知識獲得とその運用が挙げられる．例えば，「赤い丸」があった場合，それが食卓の上などなら「りんご」，自動車を運転中なら「信号」といった具合に「赤い丸」の意味は変化する．この変化が発生する状況の集合が多世界である．

人間との共生を可能にするには，ロボットが人間と円滑にコミュニケーションする能力と，周囲の状況にあわせて適切な行動をとる能力が必要になる．しかし，環境から得られる発話やオブジェクトの情報は、状況によって意味が変化するため、その情報がどの世界に属するかを扱わなければならない．

本稿では，シーン上から取得した視覚情報と音声情報から遂行されているタスクを推定し、ロボットなどの行動決定に利用することを目的として，多世界の情報に対応可能なタスク推定手法を提案した．タスク推定には，タスク遂行中の画像データと発話データを視覚情報および音声情報として利用する．

以降，2 章で提案手法について述べた後，3 章で提案手法の有効性を示すために実施した評価実験について詳述する．最後に 4 章で本稿をまとめる．

2. タスク推定方法

人間とロボットが共生するには，前述の二つの能力が求められる．この二つの能力を達成するには，ロボットが多世界の知識を適切に利用する必要がある．

このため，提案手法では多世界の知識を扱うために，タスクと情景および音声言語情報の関連性を記述した「タスク行列」を利用してタスク推定を行う．また，タスク推定を行うに当たり，システムが情報を持っている既知のタスクと，推定対象となるタスクのそれぞれについて「タスク内の静止画像」と，「タスク中の発話テキスト」の 2 種類のデータを利用する．図 1 に部分空間に基づくタスク推定手法を示す．最初に，前述の 2 種のデータを基に入力データを作成し，入力データから入力タスクのベクトルを生成する．そして，システムの持つ既知のタスクの情報のベクトルとの類似度を計算し，タスクの推定を行う．以降，提案手法について詳述していく．

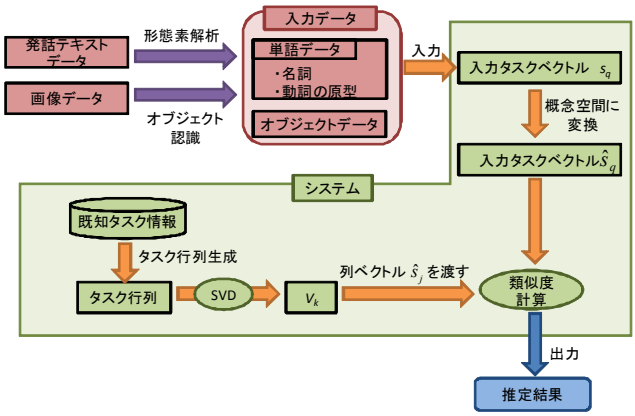

図 1. 部分空間に基づくタスク推定

2.1 タスク行列の生成

本手法では，表 1 のような，行を画像中のオブジェクトや発話中の単語，列をタスク，行列の要素を各々のタスク中でのオブジェクトや単語の出現回数としたタスク行列を生成しタスク推定を行う．

(1) 画像中のオブジェクトの識別手法

提案手法では，タスク中から抽出した画像データよりオブジェクトの判別を行い，オブジェクトの種類毎の出現数をタスク行列に利用している．オブジェクトを判別するため，以下の処理を行う．

i. オブジェクトの領域データの抽出
ii. オブジェクト領域データから画像特徴量を抽出し，オブジェクトの形状特徴量を算出
iii. オブジェクトの各プロトタイプの形状特徴量と上記オブジェクトの形状特徴量のマハラノビス距離を算出
iv. マハラノビス距離よりオブジェクトの種類を判別

以降，各処理の詳細について述べる．

まず，オブジェクト領域の抽出のため，画像データから任意の背景色を除去し，背景色以外の色をオブジェクトの領域として扱う．図2にオブジェクト領域の抽出例を示す．次に，オブジェクトの画像特徴量として以下の五つを抽出する．

連絡先：澤田心太，豊橋技術科学大学，〒441-8580
愛知県豊橋市天伯町雲雀ヶ丘 1-1，0532-44-5735，
sawada@vox.tutkie.tut.ac.jp

- オブジェクト領域の周辺長
- オブジェクト領域の面積
- 曲率の絶対値の平均
- バウンディングボックスの 2 辺の長さ
- Convex Hull の面積

ここで，バウンディングボックスとはオブジェクト領域を囲う最少の矩形であり，Convex Hull はオブジェクト領域を囲う最少の凸図形である．そして，形状特徴量として以下の六つを計算する．

- 面積
- 平均太さ
- 円形度
- 閉方度
- 細長さ
- 曲率の絶対値の平均

面積は画像特徴量のものを流用し，曲率の絶対値の平均については，オブジェクト領域の輪郭から算出する．それ以外の算出方法については文献[画像処理ハンドブック編集委員会 87]を参照されたい．

提案システムでは，各オブジェクトの「プロトタイプ」を学習させ，そのプロトタイプの形状特徴量と抽出オブジェクトの形状特徴量のマハラノビス距離を計算しオブジェクトを判別する．その際，距離が最小となったプロトタイプの種類が抽出オブジェクトの種類となる．

ここで，プロトタイプとは，認知科学において，カテゴリの中心的な内的表象のことを指し，カテゴリ事例の特徴情報を抽象化し統合した単一表象で，事例のもつ平均や頻度，構造などの情報からなるものである[須藤 04, Rosch 76]．

本稿では，プロトタイプを「ある種類のオブジェクトの平均的な形状を持つオブジェクト」，としている．プロトタイプは，同種のオブジェクトを複数システムに提示し，その形状特徴量を平均することによって得ている．

(2) 発話からの単語データの抽出

提案手法では，タスク遂行中の発話を書き起こし，テキストデータ化したものを利用している．発話テキストに対して MeCab を用いて形態素解析し，その中から名詞と動詞のみを取り出す．これらの単語をタスク行列の単語要素として入力する．

2.2 特異値分解(SVD)

提案手法では，タスク行列を利用することによりタスク推定を行う．このタスク行列内の各タスクに関する情報を得るために，タスク行列に対して SVD を行う．

SVD は，行列を分解する手法の一つである．具体的には，ある行列を三つの行列に分解する手法である．例えば，$m \times n$ の行列 A を分解すると，

$$A = USV^T \quad (1)$$

となる．ここで，U は $m \times r$, S は $r \times r$, V は $n \times r$ の行列で，$r = \min(m, n)$である．また，U, V の各列ベクトルはそれぞれ左特異ベクトル，右特異ベクトルと呼ばれる．また，S は対角行列で，その対角成分を特異値という．また，SVD によって分解された行列は，上位 k 個の特異値とそれに対応する特異ベクトルを用いて，階数 k において最小二乗誤差で近似を得ることが出来る．

$$A_k = U_k S_k V_k^T \quad (2)$$

このような近似により，次元圧縮が可能となる．

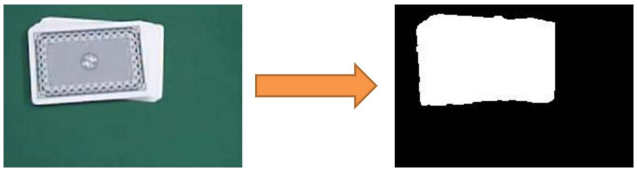

図 2. オブジェクト領域抽出例

表 1. タスク行列の例

	タスク 1	タスク 2	…	タスク n
オブジェクト 1	0	1	…	2
⋮	⋮	⋮	⋮	⋮
オブジェクト i	1	1	…	0
単語 1	2	0	…	1
⋮	⋮	⋮	⋮	⋮
単語 j	1	1	…	1

SVD により A を分解して得ることのできる U の行ベクトルと V^T の列ベクトルは，それぞれ A の行要素と列要素に関する情報を含有している．U の行ベクトル同士，もしくは V^T の列ベクトル同士を比較することにより行要素同士または列要素同士の関連性を調べることができる．また，ベクトルの存在する空間が同じならば，U の行ベクトルまたは V^T の列ベクトルと任意のベクトルを比較することができる．提案手法では，タスクの情報が必要なため，列要素の情報が含まれている(2) 式の V_k^T を利用する．

2.3 タスク間の類似度判定

推定対象タスクが，どのタスクであるのかを判断するためには，タスク間の類似度が必要となる．本稿では，推定対象タスクの情報をベクトルとし，既知のタスクのベクトルとのコサインを類似度として使用する．この類似度は，値が 1 に近いほど二つのシーンが類似しており，-1 に近いほど二つのシーンが似ていないことを表す．

類似度は，V_k^T の各列ベクトルと推定対象タスクの情報のベクトルとのコサインを計算することで得られる．しかし，そのままでは V_k^T の列ベクトルと推定対象タスク情報のベクトルの空間が異なる．そこで，推定対象情報のベクトルを以下の(3)式で V_k^T の列ベクトルと同じ空間に変形し，計算を行う．

$$\hat{q} = S_k^{-1} U_k^T q \quad (3)$$

ここで q は入力情報のベクトルである．また，$\hat{q}$ は変換後のベクトルである．

この変換を適用した後に，V_k^T の列ベクトルとのコサイン尺度を計算し，最も類似度が高くなる V_k^T の列ベクトルのタスクが推定結果となる．

3. 実 験

複数のタスクを対象としたタスク推定を行い，提案手法の有効性を評価する．

3.1 実験条件

今回対象とするタスクは，テーブル上で行われるポーカー，大富豪，ブラックジャック，回り将棋，詰将棋，切り絵の 6 種とした．被験者は 1 グループを 3 名とした 4 グループで計 12 名である．そして，各グループがテーブル上でタスクを遂行している場面を録画・録音したデータを実験データとして利用した．画像データには録画した動画から未知のオブジェクトが存在せず，オブジェクト同士が重なっていない(但しカードの山などを除く)シーンを抽出し，これを画像データとして実験で利用する．また，発話テキストデータには，録音した音声データを人が書き下したものを利用する．実験に使用したオブジェクトは，トランプ，将棋の駒，折り紙，はさみの 4 種であり，これらのオブジェクトのプロトタイプを予め学習させておいた．被験者はタスク遂行中に自由に発話を行う．4 組のデータの内一つを評価用データとして抜き出し，残りを学習用データとする交差確認法(CV 法)を用いて推定結果を評価する．評価用のタスクと同種のタスクを推定さ

れた場合を正解とした。SVD での近似に使用する階数 k は事前実験の結果，6 とした．

また，どの程度の発話量があれば推定可能なのかを調べるため，推定対象タスクの発話データの量を 10%から 100%まで 10%刻みで変化させ，推定対象タスクの画像を使用する場合と使用しない場合の各々に対してタスク推定を行った．

3.2 実験結果

画像有りの場合の実験結果を図 3，5，7，9，11，13 に，画像なしの場合の結果を図 4，6，8，10，12，14 にそれぞれ示す．これらの図より，画像なしの場合では，一部のタスクを除いて全発話を使用しても正解を得られていないことが分かる．一方，画像有りの場合には，図に示されるように，すべての場合で正解タスクの値が他のタスクの値よりも大きくなっていることが分かる．

3.3 考 察

図 4，6，8，10，12，14 より，画像を使用しない場合，大富豪と周り将棋、詰将棋は正しくタスクを推定することができたが，それ以外の三つは，発話をすべて使用しても正しく推定できなかった．間違った三つの内，ポーカーとブラックジャックの二つはほとんど大富豪と推定され，切り絵は詰将棋と推定された．これは，それぞれに共通する言葉が原因と考えられる．つまり，ポーカーやブラックジャック，大富豪では，カードの数字が大きな意味を持つからである．特に大富豪ではカードの数字は直接カードの強さとなるので，数字の持つ意味合いが大きくなり，出現頻度も増える．このため，ポーカーやブラックジャックが大富豪と誤推定されたと考えられる．そして，切り絵のタスクでは，グループの一人が他の二人に切り方と折り方を教えるが，この時に場所を指す「ここ」や「そこ」などのような言葉が見られた．これらの単語は，詰将棋のタスクでどこに駒を動かすかを話し合っている時などにも出現している．こちらの場合，この指示語が影響して誤推定されたと思われる．

図 3，5，7，9，11，13 より，発話の量が少ない場合でも，画像がある場合は安定して正しくタスクを推定出来ていることがわかる．このことから，画像から正しくオブジェクトが認識できる場合，オブジェクトのデータを利用することによってより短い時間でタスクを推定することが可能であると言える．また，画像を用いることによって，使用するオブジェクトが共通しているタスクが上位になり，そうでないタスクから分離され，かつ使用オブジェクトが共通しているタスク群の中でも，正解タスクとそれ以外がはっきりと分離されていることが分かる．使用しているオブジェクトが同じタスク同士でも，使用されているオブジェクトの個数が違えば，同様の効果が得られると考えられる．このことから，単語と画像を組み合わせることにより，使用するオブジェクトや単語が共通している場合にも，誤った推定を防ぐ効果が得られると言える．ただし，今回の実験では，遂行されるタスクと，タスク中で使用されるオブジェクトの種類が多くはないので，今後はより大規模な実験を行い提案手法の汎用性を検証していく予定である．

4. まとめ

本稿では，タスク中の発話と静止画像から部分空間に基づいたタスク推定手法を検討した．そして，その有効性を確認するための評価実験を行った．その結果，タスク遂行中の発話のみで推定を行った場合と，タスク遂行中の発話と静止画像を併用して推定を行った場合におけるタスク推定精度を比較したところ，両手法に有意な差を確認することができた．また，発話量が少ない場合においても，本手法は正確にタスクを推定できていることが分かった．このことから，本手法は，発話のみからタスクを推定する手法よりも正確で，有用であることが確認された．

今後の課題としては，提案手法を適用できるタスクの拡張があげられる．本稿では，ポーカー，大富豪，ブラックジャック，回り将棋，詰将棋，切り絵といったゲーム的なタスクを実験に使用した．このようなゲーム的なタスクだけではなく，一般的・日常的な状況に対応できるようにするためにも，本提案手法を適用できる例を拡張していき，より多様なタスクの推定を行えるようにすべきであろう．

参考文献

[新田 08] 新田恒雄: 知的エージェントとその言語発達に関する研究フレームワーク，人工知能学会，3E3-1，2008

[画像処理ハンドブック編集委員会 87] 画像処理ハンドブック編集委員会: 画像処理ハンドブック，昭晃堂，1987

[須藤 04] 須藤珠水，茂木健一郎: 言語獲得期における語意学習とカテゴリー認知のメカニズム，信学技報 SIS2004-4 Vol.104 No.144，電子情報通信学会，pp17-22，2004

[Rosch 76] Rosch，E.，Mervis，C. B.，et.al: Basic objects in natural categories，Cognitive Psychology，pp382-439，1976

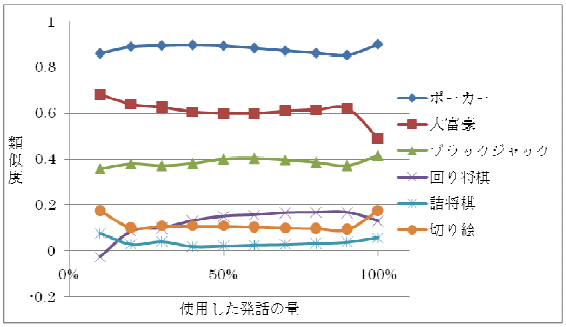

図 3. ポーカーを入力した場合の類似度(画像有り)

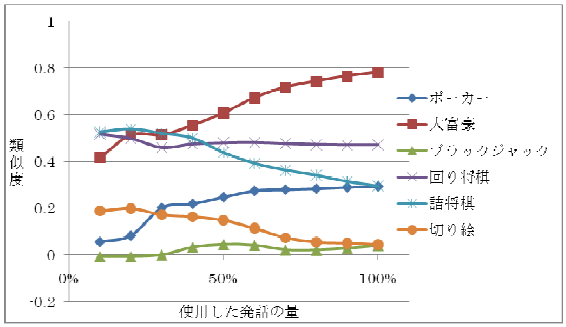

図 4. ポーカーを入力した場合の類似度(画像無し)

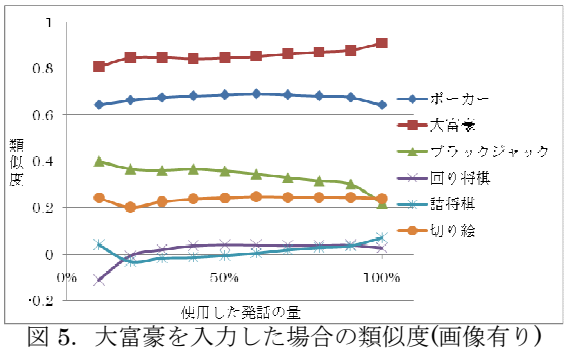

図 5. 大富豪を入力した場合の類似度(画像有り)

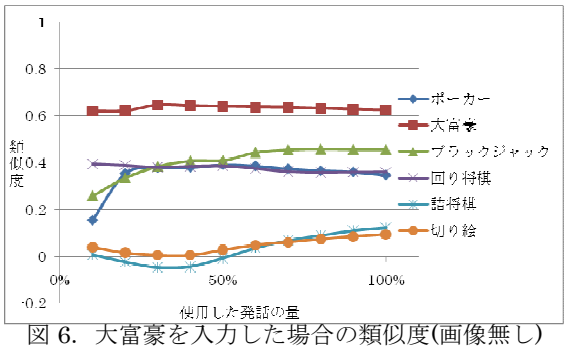

図 6. 大富豪を入力した場合の類似度(画像無し)

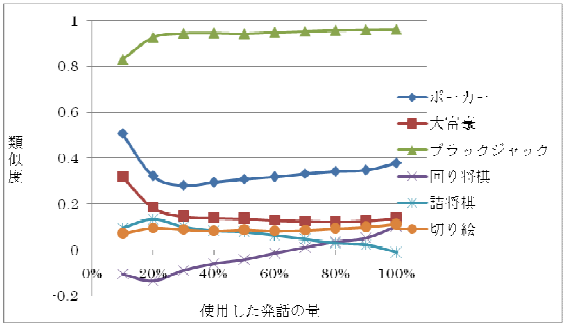

図 7. ブラックジャックを入力した場合の類似度(画像有り)

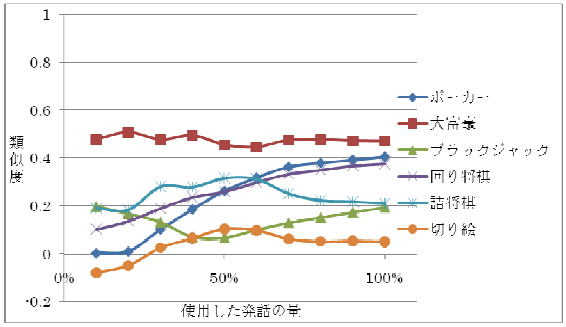

図 8. ブラックジャックを入力した場合の類似度(画像無し)

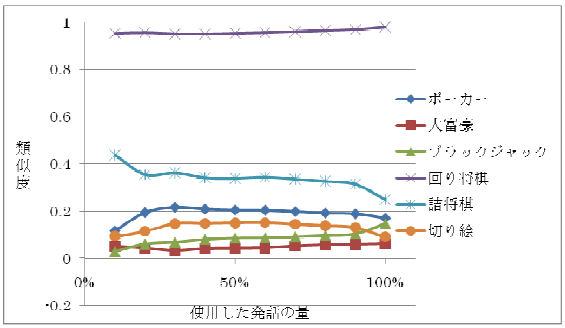

図 9. 回り将棋を入力した場合の類似度(画像有り)

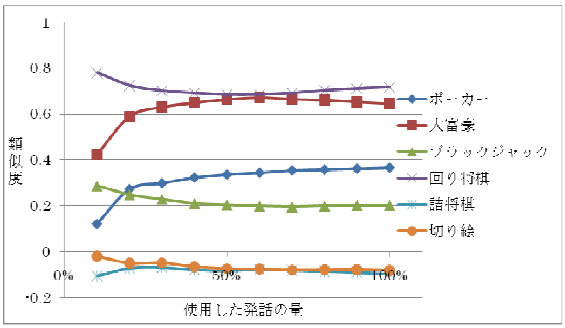

図 10. 回り将棋を入力した場合の類似度(画像無し)

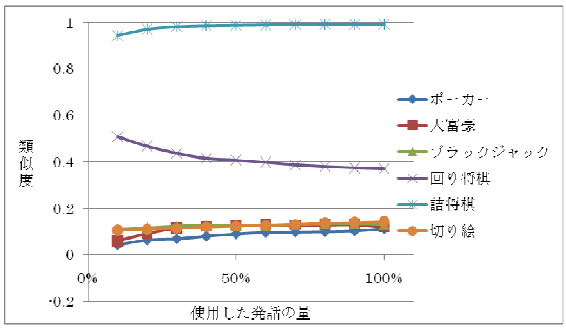

図 11. 詰将棋を入力した場合の類似度(画像有り)

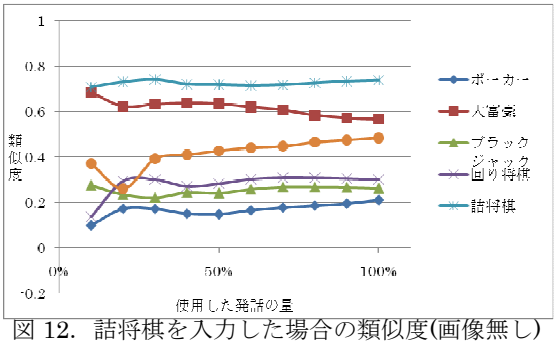

図 12. 詰将棋を入力した場合の類似度(画像無し)

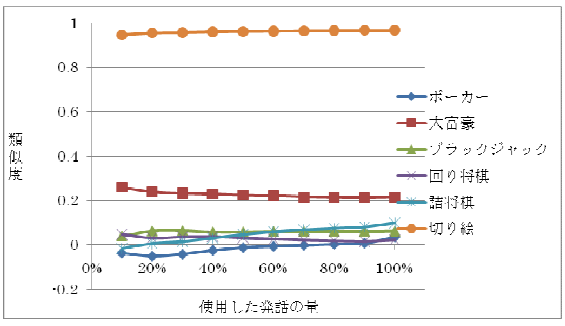

図 13. 切り絵を入力した場合の類似度(画像有り)

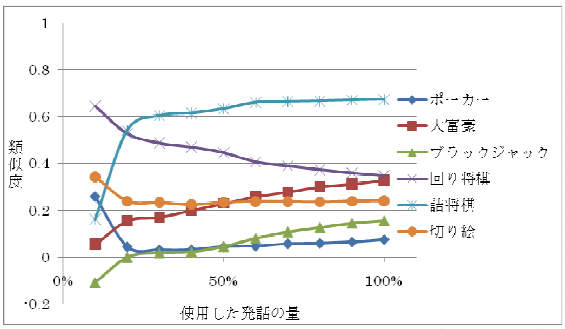

図 14. 切り絵を入力した場合の類似度(画像無し)