1J1-OS13-10

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の改良

Task Identification Using Latent Semantic Analysis of Visual Scenes and Spoken Language

澤田心太[*1] 木村優志[*1] 桂田浩一[*1] 新田恒雄[*1]
Shinta SAWADA Masashi KIMURA 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

In this paper, we propose a task identification method on multiple subspaces extracted from multimodal information of image objects and utterances appeared in the task scene. The multiple subspaces are obtained by singular value decomposition (SVD), or latent semantic analysis (LSA). In the experiments, the frequencies of image object appearances and the words in the utterances are extracted. To identify the task, an input scene and reference subspaces of different tasks are compared. Experimental results show that the proposed method outperforms the method in which only single modality information is applied. Moreover, the proposed method achieved accurate performance even if less spoken information is applied. The effect of TF-IDF as a preprocessor of LSA is also investigated.

1. はじめに

人間とロボットとの共生社会を実現するためには，様々なタスクやドメインにおいて両者がマルチモーダルに対話を行うことが不可欠である．様々なタスクやドメインが混在する状況において，ロボットはまず，現在進行している協業すべきタスクを認識する必要がある．例えば，赤くて丸いものが認識されたとき，その解釈は状況によって様々である．それが道路で認識された場合には，交差点を渡ってはいけないことを意味する．またある状況では，機械の電源が入っている状態を意味することもある．そのため，赤い丸に関する発話は状況によって変化することになる．

我々が何らかのタスクを遂行している場合，タスク遂行に必要なオブジェクトを確認し，関連する発話を聞き取っている．タスクに関するマルチモーダル情報から状況を理解することで，適切な支援を行うことができる．本報告では，発話と画像シーンのマルチモーダル情報に対して潜在意味解析（LSA: Latent Semantic Analysis）を適用し，タスクを推定する手法を提案する．これまで潜在意味解析は，文書分類などのテキスト解析において，大きな成功を収めてきた[1]．近年，LSA をテキスト情報だけでなく画素ヒストグラムのような画像特徴とともに用いる手法が提案されている[2]．本稿では，タスク遂行中に表れる画像オブジェクトと発話の情報から，LSA を用いて部分空間を構築し，タスク推定に用いる．

以下ではまず，2節で提案手法について説明した後，3節でタスク推定の精度に関する実験を行う．最後に，4節で本報告のまとめを行う．

2. LSA を用いたタスク推定手法

以下では，我々の提案する LSA を用いたタスク推定手法について説明する．提案手法の概要を図1に示す．

2.1 タスクのベクトル空間表現

我々は，他人が行っている作業を見たり，話していることを聞いたりする中で，彼らがどのようなタスクを遂行しているのか分かる．これは，共同作業中には，タスクに関連する話をしたり，タスクに関連する物体を操作することによると考えられる．そこで，以下ではタスク遂行中の発話と，出現するオブジェクトを用いてタスクを表現することを検討する．

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.cs.tut.ac.jp

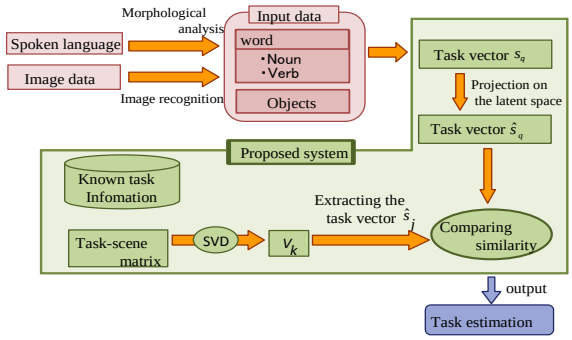

図 1 LSA を用いたタスク推定手法

文書分類などのテキスト解析研究では，文書をベクトル空間上に表現する手法が用いられてきた．これらの手法では，単語頻度を要素とするベクトル文章を表現する．これと同様，タスク遂行中の発話に含まれる単語と，作業中に使われたオブジェクトとその数を要素にすることで，タスクをベクトル空間上に表現する．以下，発話中の単語とオブジェクトをあわせてタームと呼び，タスク中に表れるタームのベクトルをタスクシーンと呼ぶ．タスクシーン t は，(1)式のように表される．

$$t = (w_1 \ w_2 \ \cdots \ w_M \ o_1 \ o_2 \ \cdots \ o_K)^T \quad (1)$$

ここで，w_i はコーパス中の i 番目(i=1, 2, …,M)の単語の正規化した頻度，o_j は j 番目(j=1, 2, …,K)のオブジェクトの正規化頻度である．

(1) 画像中のオブジェクトの識別法

オブジェクトの頻度 oj を得るには，タスク中の画像からオブジェクトを認識し，オブジェクトの種類毎に出現数を数える必要がある．以下に処理手順を示す．

i. オブジェクトの領域データを抽出する
ii. オブジェクト領域データから画像を抽出し，オブジェクトの形状特徴を算出する
iii. オブジェクトの各プロトタイプと，上記オブジェクトの形状特徴のマハラノビス距離を算出する
iv. 距離値からオブジェクトの種類を判別する
v. オブジェクトを種類毎に集計する

まず，オブジェクトの存在する領域を抽出するため，画像データから任意の背景色を除去する．次に，背景色以外の部分がオブジェクトの領域となる．図 2 に領域の抽出例を示す．次に，オブジェクトの画像特徴として以下の五つを抽出する．

- オブジェクト領域の周辺長: l
- オブジェクト領域の面積: s
- 曲率の絶対値の平均: c
- バウンディングボックスの 2 辺の長さ:長辺 e_1，短辺 e_2
- Convex Hull の面積: h

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．次に，形状特徴として以下の六つを計算する[3]．

- 面積 : s
- 平均太さ: $2s/l$
- 円形度 : $4\pi s/l^2$
- Convex Hull の面積に占める割合: s/h
- 細長さ: e_1/e_2
- 曲率の絶対値の平均: c

これらの特徴を使って認識対象オブジェクトと，各プロトタイプの平均形状とのマハラノビス距離を計算する．距離が最も近いオブジェクトを，認識結果とする．これをプロトタイプ毎に数えたものをタスク行列のオブジェクト要素として用いる．

(2) 発話からの単語抽出

本稿では，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用いた．

2.2 潜在意味解析

タスクシーンのコーパスは，タームを行，タスクシーンを列とする，タスクシーン行列によって表現できる．このタスク行列内の各タスクに関する情報を得るため，タスク行列に対して特異値分解 (SVD: Singular Value Decomposition) を行う．SVD は，行列を分解する手法の一つである．例えば，$m\times n$ の行列 A を分解すると，

$$A = USV^T \qquad (2)$$

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からは，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，近似行列$\tilde{A}$を得ることが出来る．

$$A \cong \tilde{A} = U_k S_k V_k^T \qquad (3)$$

SVD により A を分解して得た U の行ベクトルと V 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，タスクを同定するため，列要素の情報が含まれている(3) 式の V_k を利用する．

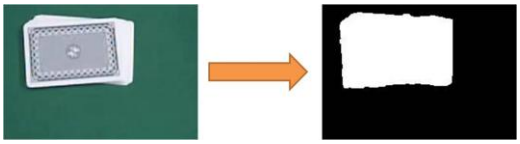

図 2　オブジェクト領域の抽出

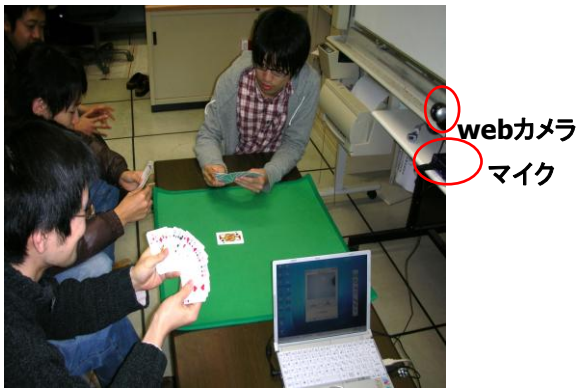

図 3　実験風景

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には，タスク間の類似度を定義しなければならない．本報告では，推定対象タスクと既知タスクのベクトルとのコサイン尺度を用いる．既知タスクのベクトルは，V_k^T の各列のベクトルから得ることができる．しかし，V_k^T の列ベクトルと推定対象タスクシーンのベクトル t では次元数が異なるため，推定対象情報のベクトルを(4)式で V_k^T と同じ空間に写像する．

$$\hat{t} = S^{-1}U_t^T \qquad (4)$$

ここで，$\hat{t}$は写像によって得たタスクシーンベクトルを表す．この変換後に，$\hat{t}$と 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 の列ベクトルに対応するタスクを推定結果とする．

3. 実験

3.1 実験条件

提案するタスク推定性能を調べるために評価実験を行った．被験者は三人一組の4グループ，計12名である．被験者は，タスク遂行中に自由に発話する．実験対象のタスクは，トランプを用いる三種類のタスクと将棋駒を用いる二種類のタスク，計五種類である．このうち，トランプを用いるタスクは，「ポーカー」，「大富豪」，「ブラックジャック」，将棋駒を用いるタスクは，「まわり将棋」と「詰将棋」である．画像オブジェクトは画像認識によって得た．一方，発話単語の頻度は，発話音声を人手によって書き起こして得ている．予備実験から，タスク行列の近似に用いる SVD の次元数は6とした．評価は，各組から1種類のタスクを評価データとし，残りをタスク行列作成に用いる一つ抜き法によって行った．実験では，どの程度の発話量がタスク推定に必要かを調べるために，発話の量を 0%から 100%まで 10%刻みで変化させた．図3に実験風景を示す．

3.2 結果と考察

タスク行列に LSA を適応した後の左特異ベクトルの要素を表1に示す．接頭辞に Obj がついているものは画像オブジェクトを表している．第一左特異ベクトルには，全タスクに共通して表れるタームが表れていることが分かる．同様に，第二にはおおむね将棋に関するタームが，第三にはおおむねブラックジャックやポーカーに関するタームが，第4には大富豪に関するタームが表れていることが分かる．

ポーカーを推定対象としたときの実験結果を図に示す．図 4 は TF-IDF を適用しない場合を，また図 5 は TF-IDF を適用した場合の結果を示している．同様に，ブラックジャックを対象と

した際の結果を，図 6（TF-IDF なし）と図 7（TF-IDF あり）に示す．それぞれのグラフの縦軸は推定対象タスクとの類似度を，横軸は推定に用いた発話量を表している．ポーカーを推定対象とした場合は，図 4 と図 5 から TF-IDF を用いた場合，用いない場合ともにポーカーの類似度が最も高くなり，正しく推定できた．

ブラックジャックを対象とした実験では，TF-IDF を適用してなかった図 6 では発話量が 100 %になっても，ポーカーと区別できていない．一方，TF-IDF を用いた図 7 では，少ない発話でも両者を区別できていることが分かる．ブラックジャックを対象とした場合に TF-IDF を適用しないと区別できない理由としては，まず，両者がともにトランプを対象としたタスクであることが挙げられる．両方のタスクには共通してトランプが画像オブジェクトとして出現する．次に，両タスクには共通して数字に関する発話が多いことが上げられる．これは，同じく数字に関する発話が多い「周り将棋」の類似度が，発話量が増えるに従って増えることからも分かる．TF-IDF を適用すると，数字に関する発話のように，複数のタスクに横断して表れるものは，IDF の値が小さくなり，悪影響が小さく抑えられている．

4. まとめ

本報告では，情景と音声言語のマルチモーダル情報から，LSA に基づきタスク固有の部分空間を抽出することで，タスクを推定する手法を提案した．また，実験から，TF-IDF を適用することで，高いタスク推定性能を得られることを示した．

今後の課題としては以下が挙げられる．今回の実験では対象タスクが5つで，各タスクに対して 3 つの学習データセットしか用いていないなど，小規模な実験に留まっている．より多くのタスクに対して本手法の有効性を検証する必要があろう．タスクベクトルの作成には，タスク中の発話の書き起こしが必要なため多くの労力が必要となった．大規模な評価を行うためにはこれの自動化が必要である

参考文献

[1] J. Bellegarda : “Exploiting latent semantic information in statistical languagemodeling,”, Proceedings of the IEEE vol.88-8, 2000.

[2] R. Zhao and W. I. Grosky: “Narrowing the semantic gap Improved text-based Web document retrieval using visual features,”, IEEE Transactions on Multimedia. Vol. 4 no.2 pp. 189-200, 2002.

[3] 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.

表 1 左特異ベクトルの要素

順位	第一	第二	第三	第四
1	ある	Obj_将棋駒	Obj_トランプ	パス
2	これ	取る	ヒット	枚
3	する	飛車	Obj_将棋駒	取る
4	てる	イチ	イチ	ハチギリ
5	なる	ゼロ	スタンド	出す

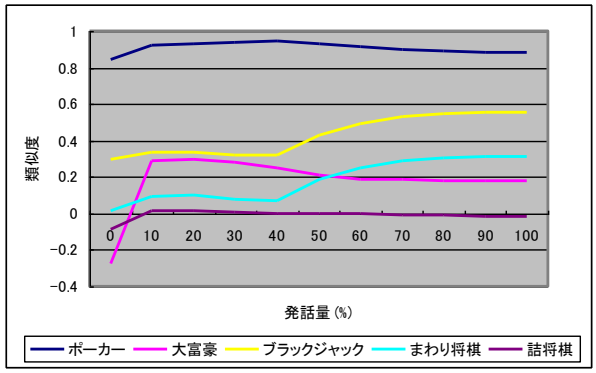

図 4.「ポーカー」を対象とするタスク：(TF-IDF なし)

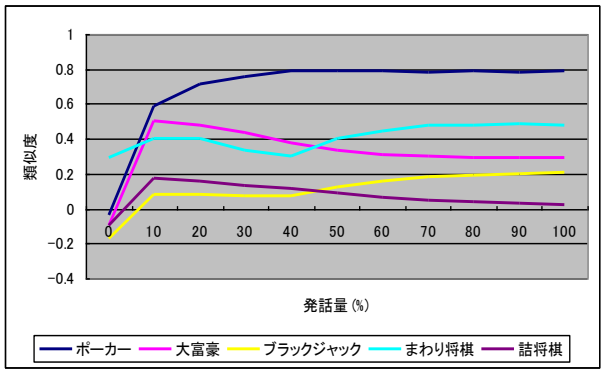

図 5.「ポーカー」を対象とするタスク：(TF-IDF あり)

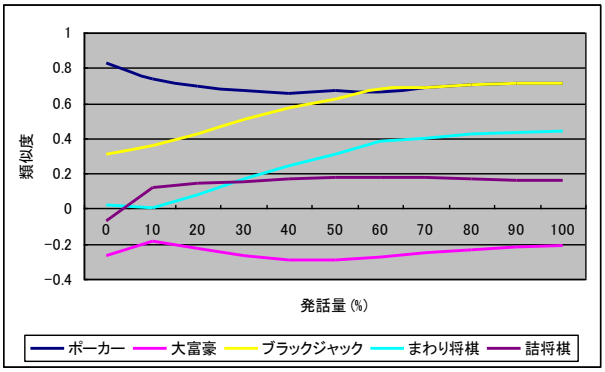

図 6 .「ブラックジャック」を対象とするタスク：(TF-IDF なし)

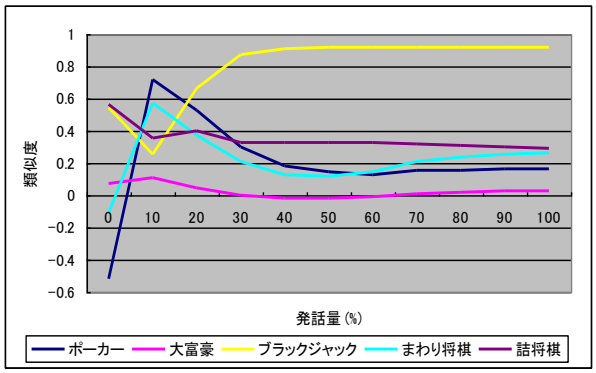

図 7 .「ブラックジャック」を対象とするタスク：（TF-IDF あり）