1G2-OS-21b-3

プロジェクションによるサル認識

Monkey Recognition by Projection

寺田 和憲
Kazunori Terada

岐阜大学工学部
Faculty of Engineering, Gifu University

We define object recognition as believing that the object exists in a location where the existence probability of the internal representation projected on the sensory input space is high. In the present paper, we explain a method in which internal representations of monkeys acquired by deep learning are projected on visual inputs and track them by particle filter. We then discuss whether it can be generalized as the process of projection.

1. はじめに

物理世界で行動するエージェントは外界から情報を受け取り行動を出力する．入力から行動出力に至る過程で情報の取捨選択や圧縮を行ない，世界を分節することは，一様で連続な世界をフレーム問題に陥ることなく，認知資源を節約し，取り扱うことができる有用な戦略である．内部表象とは取捨選択や圧縮によって生成された，エージェント内部で使用される抽象表現である．物事の認識は実際にはエージェント内部で発生するが，情報の提供元の性質として見なされる [鈴木 16]．鈴木は，「氷水に手を入れた時に得られる皮膚からの情報は中枢系に伝わりその事態についての表象を作り出し，「冷たい」という感覚を生み出す．この感覚はシステム内部に生じるのではなく，氷水の中の手に定位される」と例示している．このように内部表象が外部に定位されることをプロジェクション（投射）と呼ぶ [鈴木 16, 小野 16]．本稿では深層学習によって獲得した内部表象を視覚空間に投射し，パーティクルフィルタによって追跡する方法を説明するとともに，プロジェクションのプロセスとして一般化できるかどうかを検討する．

2. プロジェクションによるサル認識

認識においてプロジェクションは必用不可欠である．オックスフォードオンライン辞書によると「認識」とは思考，経験，感覚を通じて知識と理解を獲得する心的な行為もしくは過程とされる．知識および理解の獲得過程においては同一性の判断，すなわち特徴軸の選択が行われる．ある個体のサルを見てサルであることを認識するためには，全てのサル個体に共通する特徴量に注目し，個別のサルが持つ特徴量を無視する必用がある．センサデータから特徴量の取捨選択が行われ，同一性の判定が行われるが，ここで比較される対象はサルに関する内部表象である．すなわち，認識とは現在の入力とすでに持っている抽象概念を比較し，同一かどうかを判定することである．なお，教師あり，教師なしにかかわらず，抽象表現を獲得するプロセスは学習であり，ここでは認識とは呼ばない．

センサ入力とすでに持っている抽象表現との同一性（一致度）はどのように計算されるのだろうか．抽象概念のリストの端から順番に一致度が計算され，リストの最後まで行ったときに最も一致度の高いものが現在の世界の状態として認識されるのであろうか．この方法は，世界が構造を持っていること [*1] を考慮すると非効率である．より効率的な方法は仮説を生成することである．すなわち，あらかじめ一致度を計算するべき抽象表現を準備しておき，それと現在の入力を比較するのである．文脈が認識を促進させることから人はそのような認識を行っていると考えられる [Oliva 07]．さらに，認識において本来個体の適応度を向上させるという目的が仮定される以上，目標に達成に無関係な認識は不要である．何が関係していて何が無関係であるかは文脈によって与えられる．

ここで，ベイズの定理を用いた逐次状態推定方法について説明する．この推定方法はパラメトリックな場合にはカルマンフィルタ，ノンパラメトリックな場合にはパーティクルフィルタとして知られる．問題は，推定するべき対象の状態を $\mathbf{X}$，観測結果を $\mathbf{O}_t = (\mathbf{o}_1, \ldots, \mathbf{o}_t)$ としたときに，$p(\mathbf{X}_t|\mathbf{O}_t)$ を推定することである．この確率はベイズの定理を用いることで，

$$p(\mathbf{X}_t|\mathbf{O}_t) = \kappa p(\mathbf{O}_t|\mathbf{X}_t)p(\mathbf{X}_t|\mathbf{O}_{t-1}) \quad (1)$$

のように求めることができる．ここで κ は正規化定数である．また，$p(\mathbf{O}_t|\mathbf{X}_t)$ は尤度関数や観測モデル，知識と呼ばれる．時刻 t における事前確率 $p(\mathbf{X}_t|\mathbf{O}_{t-1})$ はマルコフ性を仮定することによって，時刻 $t-1$ の事後確率と状態遷移モデル $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ を用いて，

$$p(\mathbf{X}_t|\mathbf{O}_{t-1}) = \int p(\mathbf{X}_t|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{O}_{t-1})d\mathbf{X}_{t-1} \quad (2)$$

のように計算できる．我々は式 1 において事前確率が仮定されることをプロジェクションと考える．また，尤度の計算は仮定された状態と現在の入力と比較であると考える．

我々はディープラーニングとパーティクルフィルタを用いて動画像中のサルの個体追跡を行っている [上野 17]．本手法では，サルの認識を画像中の矩形領域 $\mathbf{X} = (x, y, w, h)$ の同定問題と考える (図 1 参照)．この方法では，まずあらかじめ，畳み込みニューラルネットワークを用いてサルの全身画像の抽象表現を学習する．この学習においては，正事例として約 1 万 3 千枚のサルの全身画像，負事例として草，枝・落ち葉，雪，

連絡先: 寺田和憲，岐阜大学工学部電気電子・情報工学科，〒501-1193 岐阜市柳戸 1-1，terada@ai.info.gifu-u.ac.jp

*1 我々は，世界は一様で連続であるという前提のもとに認識を議論している，すなわち認識主体の存在にかかわらず世界が構造を持っているとは考えていないが，認識主体が分節した世界の構造がある程度持続性を持ち，主体間で共有されていることから，「世界が構造を持つ」と記述する．

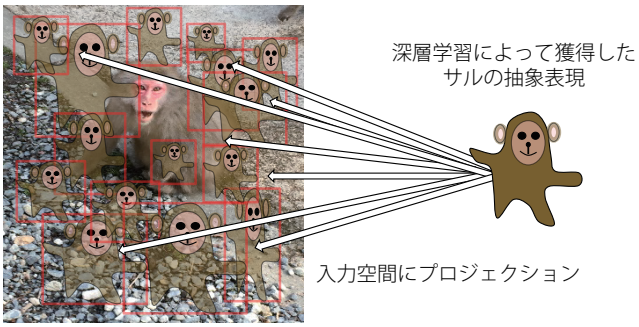

図 1: プロジェクションによるサル認識．画像入力空間中に深層学習によって獲得したサルの抽象表現を仮説としてプロジェクションし，その仮説の正しさに基づいてサルの存在領域を同定する．

アスファルト，岩，川，砂利，サルの顔のみ，サルの体の一部の画像 5 万 5 千枚を用いた．認識においては，この抽象表現を拡大・縮小し，画像入力空間に投影し，尤度を計算する．実際には矩形領域を切出し，学習済みの畳み込みニューラルネットワークに入力し，出力層の値からサルらしさの確率を求め，尤度値とした．認識の初期では，矩形は $\mathbf{X}$ の全空間に分布するが，事後確率が更新されるに従って，実際にサルが存在する領域に集中する．我々は実験によって頑健なサルの追跡が可能であることを確認した．また，同様に畳み込みニューラルネットワークによって予め学習されたサル個体の顔の抽象表現を入力画像空間に投影し，逐次ベイズ更新を用いることで，個体の存在確率の分布が推定可能であることを確認した．

推定するべき状態 $\mathbf{X}$ はサル認識の場合のは画像入力空間中のサルの重心座標と大きさであった．より一般的には，$\mathbf{X}$ は認識対象が存在するかどうか，存在するとすればどこに存在するかについての情報である．あるいは識別クラスと言ってもよい．$p(\mathbf{O}|\mathbf{X})$ は対象が状態 $\mathbf{X}$ にあるときにどのような感覚入力 $\mathbf{O}$ が得られるかについての知識である．別の言い方をすると，$p(\mathbf{O}|\mathbf{X})$ はセンサ空間と概念空間を接続する知識である．この知識は通常教師あり，教師なしにかかわらず，統計学習によって獲得される．事前確率を考慮しなければ，事後確率は知識のみに依存する．

先に述べたように，我々は事前確率を考慮することをプロジェクションであると考える．図 1 に示すように，パーティクルフィルタの実装では，事前確率の分布に従ってパーティクルがセンサ入力空間中にばら撒かれ，パーティクルそれぞれについて尤度が計算される．これは，サルの状態を仮説として生成し，その仮説の正しさに基づいてサルの存在を確信することと言える．

誤投射 [鈴木 16] の例として挙げられるラバーバンド錯覚もベイズの定理に基づく状態推定によって説明できる可能性がある．定位に直接関与するセンサは視覚 $\mathbf{o_v}$ と自己受容感覚 $\mathbf{o_p}$ である．手の定位はそれらの情報が与えられたときに $p(\mathbf{x}|\mathbf{o_v}, \mathbf{o_p})$ を推定する問題である．ラバーバンド錯覚条件下におけるある時刻の入力が $\mathbf{o}_{\mathbf{v}t}$，$\mathbf{o}_{\mathbf{p}_t}$ であったときに，事後確率は，すでに持っている位置とセンサ入力の関係についての知識である $p(\mathbf{o_v}|\mathbf{x})$ と $p(\mathbf{o_p}|\mathbf{x})$ に基づいて計算され，二峰性の分布となる．二つのセンサが異なる位置を出力した場合には，異なる位置として知覚されるのではなく，統合されて単一の位置として知覚される．また，確信度の高い（分散が小さい）方のセンサが信頼される [Ernst 02]．ラバーバンド錯覚では，ゴムの手と自分の手が同時にタップされることによって，視覚の確信度が向上し，その結果，視覚の位置に定位が偏るものと考えられる．このときに重要なのは，定位にほとんど影響のない皮膚感覚を使って視覚の確信度を向上させることである．同様の考えは Samad らによって提案されている [Samad 15]．

3. おわりに

鈴木が「プロジェクションがあまりに当たり前に行われるために，その存在自体に気づかない」[鈴木 16] と記述するように，プロジェクションは認識プロセスに深く組み入れられているために，これまで明示的に議論されることはなかった．本稿では，パーティクルフィルタの実装において，実際に内部表象がセンサ空間に投射されることにヒントを得て，ベイズの枠組みによるプロジェクションの説明可能性を示した．その中で，誤投射の一つであるラバーハンド錯覚についての説明は試みたが，雪山などの極限状態において実在しない人物が知覚されるというサードマンなどの虚投射についての説明は行っていない．今後は雪山においてサードマンを発生させる条件の実験的検討をするなどしてプロジェクション科学の発展に貢献したい．

謝辞

サル認識器の開発においてご協力頂きました大阪大学の山田一憲先生，上野将敬氏，岐阜大学の林英誉氏，加畑亮輔氏に感謝致します．本研究は JSPS 科研費 16K12757 の助成を受けたものです．

参考文献

[Ernst 02] Ernst, M. O. and Banks, M. S.: Humans integrate visual and haptic information in a statistically optimal fashion, *Nature*, Vol. 415, pp. 429–433 (2002)

[Oliva 07] Oliva, A. and Torralba, A.: The role of context in object recognition, *Trends in Cognitive Sciences*, Vol. 11, No. 12, pp. 520 – 527 (2007)

[Samad 15] Samad, M., Chung, A. J., and Shams, L.: Perception of Body Ownership Is Driven by Bayesian Sensory Inference, *PLOS ONE*, Vol. 10, No. 2, pp. 1–23 (2015)

[小野 16] 小野 哲雄：「プロジェクションサイエンス」の視点からの認知的メカニズムのモデル論的理解, 日本認知科学会第 33 回大会発表論文集, pp. 26–30 (2016)

[上野 17] 上野 将敬, 寺田 和憲, 加畑 亮輔, 林 英誉, 山田 一憲：ディープラーニングとパーティクルフィルタを用いた動画像中のニホンザルの個体追跡, 第 64 回日本生態学会大会 (2017)

[鈴木 16] 鈴木 宏昭：プロジェクション科学の展望, 日本認知科学会第 33 回大会発表論文集, pp. 20–25 (2016)