画像認識技術を活用した新規サービス創出に向けた価値検証プロセスの提案

Value-Verification Process Proposal for Launching New Services using Image Recognition Technology

若井信彦 Nobuhiko Wakai
小塚和紀 Kazuki Kozuka
中田洋平 Yohei Nakata
飯田恵大 Shigehiro Iida

要　旨

行動センシングを含む新規サービス創出はユーザー行動を常に監視する必要があるため，従来のプロトタイピング手法では開発困難である．そこで本稿では，従来のプロトタイピング手法で模擬できない行動認識器を作成し，短期間で実証実験を実施する価値検証プロセスを提案する．既存の特定用途向け認識器を複数組み合わせ，計算コストの小さい一部の認識器のみを学習対象にすることで，少量のデータかつ短期間で行動認識器を学習できる．仮想の新規サービス案に対し提案プロセスを適用することで，一から認識器を学習する従来のプロセスより短期間で実証実験が可能であることを示す．

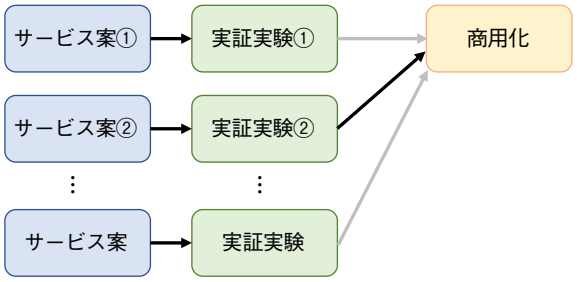
Abstract

Conventional prototyping methods are not suitable for evaluating proof of concept of services using action sensing due to their constantly monitoring users. Therefore, it is difficult to launch new services through repetitive evaluation. We propose a value-verification process for swiftly evaluating proof of concept using simple action recognizers that simulate action that prototyping methods cannot. Our action recognizer can be trained using a small dataset for a short time because the recognizers consist of several specific recognizers and a trained recognizer with a low computational cost. Assuming a virtual new service, we show that the evaluation time in our process is shorter than that in a process that trains a recognizer from scratch.

1. はじめに

新規サービス創出における実証実験は，仮説と検証を実際のユーザー価値で実施することが重要で，また，多くのサービス案を検討し質を高めるため短期間で繰り返し実施する必要がある（**第1図**）．実証実験を迅速に実施するために，ペーパープロトタイプ[1]やオズの魔法使い法[2]などのプロトタイピング手法がある．しかし，これらのプロトタイピング手法が適用できないサービス形態があり，その1つが行動センシングを含むサービスである．行動センシングを含むサービスとは，撮影画像や動画からユーザー行動を認識し，その得られたデータから開発者が意図するサービスをユーザーに提供することであり，実証実験では本番運用と同等のサービスを提供し検証することが必要である．しかし，従来のプロトタイピング手法は，ユーザーの行動を常に監視する必要があり，本番運用と同等のサービスを提供できないため，行動センシングを含む新規サービスを創出することが困難である．

近年，深層学習のブレイクスルーにより動画から人の行動を推論する複雑な認識タスクが可能となった[3][4]．しかし，行動認識の学習には，多くのデータと時間を要し，また，認識する行動によっては，深層学習の構造の設計が必要となり，新規の行動認識は容易ではない．行動センシング向け識別機の作成も同様に容易ではなく，行動センシングを含むサービスの価値検証プロセスの実施は膨大な労力と時間を要することが問題であった．この問題を解決するため，本稿では画像認識による行動センシングを簡易に実現する新規の価値検証プロセスを提案する．提案プロセスは，既存の特定用途向け画像認識技術を組み合わせ，新規の画像認識を簡易に学習する．この簡易な学習により，実証実験を短期間で実施できるようになる．



第1図　価値検証プロセス
Fig. 1　Value verification process

2. 価値検証の提案プロセス

著者らは，行動センシングを含む新規サービスを短期間で実証実験できる価値検証プロセスを提案する．提案プロセスは認識器の作成に特徴があり，既存のさまざまな画像認識技術で人やものを認識できることに着目し，既存技術を利用して目的とする行動認識器を作成する．なお，ユー

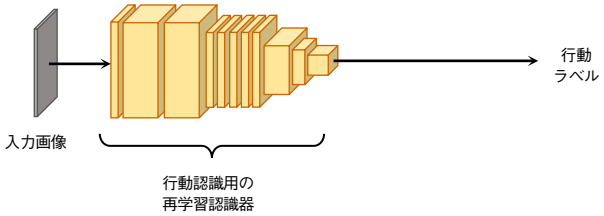
ザーテストは一般的な手順で実施する．

認識する行動はサービスごとに異なるため，行動認識器は，検証するサービスに応じて新規に作成する必要がある．認識器の入力は1枚あるいは時系列の画像であり，出力は行動ラベルである．行動ラベルは，歩くや立つなどの動作を表すラベル，または，認識対象以外全てを含むその他のラベルである．認識器を一から学習するプロセスと提案プロセスの認識器の作成方法を比較して説明する．

2.1 一から認識器を学習するプロセス

認識器を一から学習するプロセスでは，第1に，認識器を学習させるための学習データを作成する．この学習データは，認識対象の行動を含む多数の動画を収集し，各動画，あるいは，動画の時間の区間に対して正解となるラベルを付与する．新規サービスで必要な認識は，既存の学習データで対応できないため，新規に学習データを作成する．第2に，複雑な認識タスクである行動認識には深層学習が必要であり，認識する内容を考慮し，深層学習の構造を設計する．第3に，学習データを用いて認識器を学習する．

第2図に，一から認識器を学習するプロセスの認識器の模式図を示す．単一の行動認識用の再学習認識器を使用し，ディープニューラルネットワークを学習する．Piergiovanniらの手法[5]などの行動認識の学習には，多量のデータによる長時間の学習が必要となる．例えば，行動認識用の大規模動画データセットを用いる場合，UCF101[6]は13000個，Charades[7]は10000個の動画が行動認識の学習データとして使用され，学習には数日必要となる．上記のデータセットなどでディープニューラルネットワークを事前学習させ，作成した学習データで再学習する場合もある．

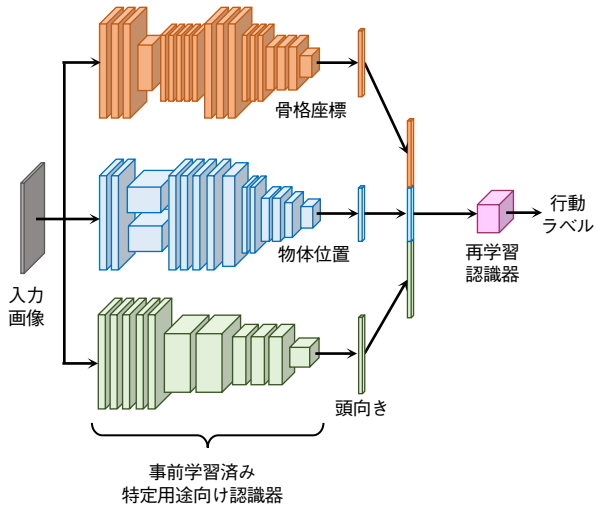


第2図　一から認識器を学習するプロセスの認識器
Fig. 2　Recognizer in the process that trains recognizers from scratch

2.2 提案プロセス

提案プロセスの特徴は，特定用途向け認識器を組み合わせて使用し，短期間で認識器を作成できることである．提案プロセスは，第1に，認識器を一から学習するプロセスと同様に学習データを作成するが，これまでの使用実績から，データ量は認識器を一から学習する場合の1/10～1/100程度である．第2に，提案プロセスにおける認識器を設計する．第3に，1日程度の短時間で認識器を学習する．

第3図に，提案プロセスの認識器の模式図を示す．提案プロセスの認識器は複数の事前学習済み特定用途向け認識器と，1個の再学習認識器から構成される．特定用途向け認識器は，例えば，人の骨格推定[8]，物体検出[9]，頭向き推定[10]であり，それぞれ，骨格座標，物体位置，頭向きを出力するように事前学習済みである．なお，頭向き推定などで入力に顔領域のみの画像が必要な場合，Dengらの手法[11]などの顔検出の認識結果を用いる．特定用途向け認識器の推論結果を結合したベクトルを再学習認識器に入力する．特定用途向け認識器の出力形式がベクトルでない場合，固定長のベクトルに変換することで，複数の認識器の出力を1つの固定長ベクトルに結合できる．



第3図　提案プロセスの認識器
Fig. 3　Recognizer in our proposed process

学習対象である認識器は，行動ラベルを出力する直前の計算コストの小さい認識器のみであり，短時間で学習可能である．また，特定用途向け認識器が出力する骨格座標や物体位置などの特徴量は，低次元だが行動認識において有用な情報を含む．なぜならば，行動ごとに人の姿勢などが異なり，骨格座標などから作成したベクトルは行動の特徴を表現できる．したがって，再学習認識器の入力は低次元のベクトルであり，浅いニューラルネットワーク，あるいは，サポートベクターマシン[12]やXGBoost[13]などの機械学習手法が利用できる．これらの機械学習手法の学習時間はディープニューラルネットワークに比べて短いため，交差検証法[14]で認識精度が最良の手法を選択しても良い．

特定用途向け認識器の選択方法について説明する．認識器の組み合わせは複数存在するが，認識対象の行動の特徴から認識器を選択できる．例えば，体全体の動きがある行

動の場合，骨格推定と頭向き推定が有効である．一方，行動の特徴が手の動きの場合，手検出[15]などの手や指を認識する認識器を選択する．また，行動に特定の物体が関係する場合，物体検出を用いる．それぞれの特定用途向け認識器の出力が行動認識に有効か否かは，学習結果を評価するまで不明である．しかし，認識器の学習過程で，行動認識に有効な特徴量が自動的に重視されるため，行動認識に有効と推測される特定用途向け認識器を複数選択すれば良い．

提案する価値検証プロセスは，既存の特定用途向け認識器を利用することで，短時間で行動認識器を学習し，実証実験を効率的に実施できる．

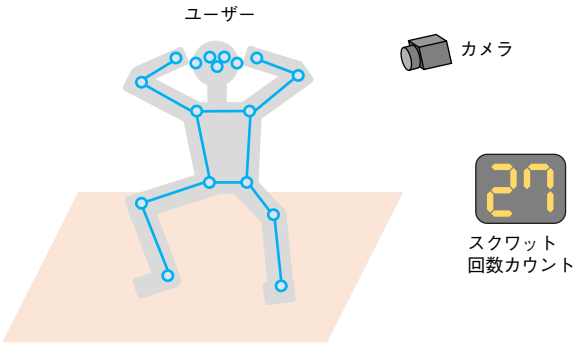
3. 価値検証における提案プロセスの適用例

仮想的な新規サービス案として，宅内でスクワット回数を認識するサービスを例として，提案する価値検証プロセスが短期間で実証実験を実施できることを示す．このサービスを実証実験のユーザーが体験するためには，スクワットの認識器が必要である．

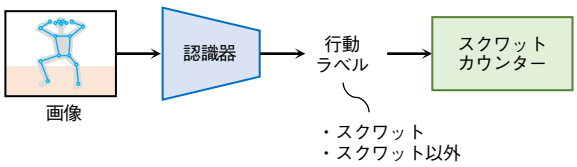
第4図に，スクワット回数を数えるサービスの模式図を示す．宅内に設置したカメラでユーザーを撮影し，ユーザーの操作を必要とせずに自動的にスクワットの回数を数える．したがって，サービス稼働時は，ユーザーが意識的に運動を記録する必要がない．ここで，骨格推定が認識するユーザーの体の主要関節と，眼・耳・鼻の例を水色の線と円で示している．

スクワット時の膝の曲げ伸ばしにより，下半身の骨格の画像座標が大きく変化する．また，歩くなどの姿勢と異なるスクワット特有の脚の開きがあるため，骨格座標からスクワットを認識できる．

第5図に，スクワット回数を数えるサービス処理フローを示す．カメラで撮影した画像を提案プロセスの認識器に



第4図　スクワット回数を数えるサービスの模式図
Fig. 4　Concept diagram of the squat counting service

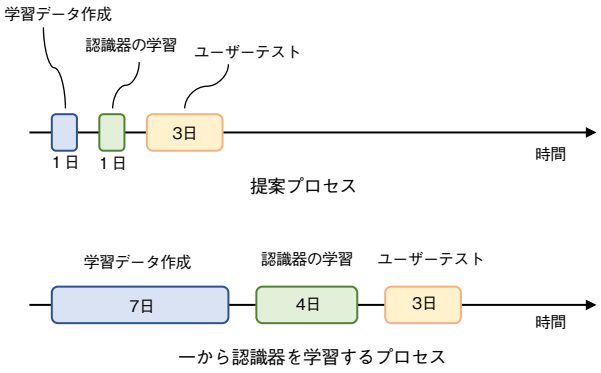


第5図　スクワット回数を数える処理フロー
Fig. 5　Process flow of the squat counting service

入力することで，スクワットかスクワット以外の行動ラベルを得る．この行動ラベルを時系列処理するスクワットカウンターでスクワット回数を数える．スクワット1回の膝の曲げ伸ばしを撮影した複数枚の画像全てに対して認識器が行動ラベルを出力するため，行動ラベルがスクワットである個数は実際のスクワット回数より多い．そのため，スクワットカウンターの時系列処理は，1回のスクワットの膝の曲げ伸ばしの間に出力される複数のスクワットのラベルをまとめて1回のスクワットとして数える．また，1日あたりの回数や，1週間あたりの回数として記録する．

3.1　実証実験の作業項目と日数

サービス考案後，実証実験の主要な作業項目は，①学習データの作成，②認識器の学習，③ユーザーテストである．**第6図**に，提案プロセスと一から認識器を学習するプロセスの作業日数を示す．各作業に必要な日数は，これまでの使用実績を踏まえ，著者らが実施する場合の概算日数を見積もる．提案プロセスの学習データの作成と認識器の学習に必要な日数はどちらも1日である．一方，一から認識器を学習するプロセスの場合，学習データの作成に7日，認識器の学習に4日必要である．両プロセスともに，ユーザーテストは3日である．



第6図　提案プロセスと一から認識器を学習するプロセスの作業日数
Fig. 6　Number of days of the proposed process and the process that trains recognizers from scratch

作業日数の詳細について，学習データ作成と認識器の学習を順に説明する．学習データ作成に関して，提案プロセスの学習に必要なデータは少量であり，スクワットとスクワット以外を認識する易しい認識タスクの場合は20個程度の動画，複数の行動を認識する場合は50個程度の動画が学習データとして必要である．これらの少量の学習データ作成は1日で実施可能である．一方，一から認識器を学習するプロセスの場合，約1000個以上の動画が必要となる．データセット[6][7]の10000個以上の動画で事前学習をした場合でも，**第2図**で示すディープニューラルネットワークを学習するためには，多数のデータが必要である．

認識器の学習に関して，提案プロセスの学習は，**第3図**で示した一部の認識器のみである．また，再学習認識器は計算コストの小さい手法を用いるため，1日で学習可能である．一方，一から認識器を学習するプロセスの場合，学習データが多く，再学習認識器の計算コストも大きいため，4日程度必要となる．なお，学習時間は使用する認識器の構造と計算機の処理性能に依存するため，典型的な構造を著者らがパソコンで計算する場合の見積もりである．ユーザーテストは対象人数や実施回数によって日数が変わるため，少数のユーザーテストを短期間で実施する場合を想定し，提案プロセスと一から認識器を学習するプロセスで共通の3日とする．

したがって，提案プロセスの合計作業日数は5日であり，一から認識器を学習するプロセスの14日より短い．

3.2 提案プロセスの簡易な認識器の有効性

ユーザーのサービス体験が実証実験の評価において重要であり，実際に動作する提案プロセスの認識器は検証に有効である．

実証実験において，提案プロセスの簡易な認識器を利用できることを説明する．提案プロセスの認識器は少量の学習データで簡易に作成するため，未検知や誤検知を生じる．しかし，実証実験において，誤認識率0％（認識精度100％）である認識器は必ずしも必要ではない．仮に，スクワットの回数の誤差が10％あった場合でも，平日と休日の回数の違いの傾向は判明する．また，スクワットを実施した日と，実施しない日の区別もできる．このように，認識対象であるスクワットの回数において誤りは存在するが，サービスに着目した場合，提案プロセスの簡易な認識器を実証実験に使用できる．

また，上述のスクワットの回数を数えるサービスに限らず，ユーザーインターフェースの設計にも提案プロセスの認識器は利用できる．例えば，スクワットを検知した時とそれ以外の時で，ユーザーに提示する画面を変える場合を想定する．この場合，ユーザーの動作に応じてリアルタイムに提示内容が切り替わることが重要であり，動作するサービスとして機能することが必要である．したがって，簡易な認識器を用いることで，ユーザーインターフェースの設計が容易となる．

4. 提案プロセスの適用範囲

提案プロセスの特徴と制約について述べる．**第1表**に学習と推論の特徴の比較を示す．提案プロセスは推論計算コスト（認識に必要な計算量）や認識精度を犠牲にすることで，少量データによる短時間の学習を実現する．処理性能の高いパソコンなどを使用可能なため，実証実験で推論計算コストが大きいことは問題ない．

第1表 学習と推論の特徴の比較
Table 1 Comparison of training and prediction characteristics

	一から認識器を作成するプロセス	提案プロセス
学習時間	× 長時間	○ 短時間
学習データ量	× 多量	○ 少量
学習データ費用	× 高い	○ 安い
学習計算コスト	× 大きい	○ 小さい
推論計算コスト	○ 小さい	× 大きい
認識精度	○ 高い	× 低い

特定用途向け認識器を組み合わせて簡易に認識器を作成するため，認識が困難な行動がある．第1に，使用する機器などが物体検出で認識できない場合である．例えば，空気清浄機を操作する行動の場合，空気清浄機は一般的な物体検出器で認識できない．第2に，認識内容が詳細な行動の場合である．例えば，パソコンで書類を作成する行動の場合，パソコンで使用しているアプリケーションを特定することは困難である．第3に，行動が多様な場合である．例えば，休息をとる行動の場合，椅子に座る・ベッドで寝る・ソファで横になるなどの多くの行動が休息をとるに該当し，姿勢や物体位置で学習することは困難である．第4に，行動にユーザーの意図が伴う場合である．例えば，誕生日を祝う行動の場合，動作に直接関係しない誕生日は認識できない．

認識精度の制約について説明する．認識する行動や学習データに依存するため，提案プロセスの認識器の精度を学習前に見積もることは困難である．また，提案プロセスは一部の認識器のみを少量のデータで学習するため，学習データ[6][7]と同等のデータ量で学習した認識器より精度が低い可能性が高い．したがって，誤検知や未検知が許容できない実証実験には適用できない．

上述の制約を踏まえ，提案プロセスが有効な典型的行動は，標準的な動作の姿勢があり，行動に関連するものが物体検出で認識できる行動である．日常的に繰り返す動作などは，提案プロセスが有効な典型的行動である．

5. まとめ

既存の特定用途向け認識器を組み合わせることで，サービスの実証実験に必要な行動認識器を作成するプロセスを提案した．行動認識を対象としたサービス創出プロセスについて述べたが，物体検出などの認識器は人以外のものを認識できる．したがって，行動認識に限定せず，空間センシングを含むサービス創出に適用できる可能性がある．今後，提案プロセスの適用範囲を広げ，さまざまなセンシングを含むサービス創出を目指す．

参 考 文 献

[1] A. Lancaster et. al., “Paper prototyping: The fast and easy way to design and refine user interfaces,” IEEE Transactions on Professional Communication, vol. 47, pp. 335-336, 2004.

[2] C. Hummel et al., “Meaningful gesture for human computer interaction: Beyond hand postures,” IEEE Computer Society Press, pp. 591-596, 1998.

[3] F. C. Heilbron et al., “ActivityNet: A large-scale video benchmark for human activity understanding,” in Proc. Conference on Computer Vision and Pattern Recognition, Boston, Jun. 2015, pp. 961-970.

[4] C. Gu et al., “AVA: A video dataset of spatio-temporally localized atomic visual actions,” in Proc. Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun. 2018, pp. 6047-6056.

[5] A. Piergiovanni et al., “Representation flow for action recognition,” in Proc. Conference on Computer Vision and Pattern Recognition, Long Beach, Jun. 2019, pp. 9945-9953.

[6] K. Soomro et. al., “UCF101: A dataset of 101 human actions classes from videos in the wild,” CRCV-TR-12-01, University of Central Florida, Florida, Nov. 2012.

[7] G. A. Sigurdsson et al., “Hollywood in homes: Crowdsourcing data collection for activity understanding,” in Proc. European Conference on Computer Vision, Amsterdam, Oct. 2016.

[8] G. Papandreou et al., “PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model,” in Proc. European Conference on Computer Vision, Munich, Sep. 2018.

[9] W. Liu et al., “SSD: Single shot multibox detector,” in Proc.European Conference on Computer Vision, Amsterdam, Oct. 2016.

[10] N. Ruiz et al., “Fine-grained head pose estimation without keypoints,” in Proc. Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, Jun. 2018, pp. 2187-2196.

[11] J. Deng et al., “RetinaFace: Single-shot multi-level face localisation in the wild,” in Proc. Conference on Computer Vision and Pattern Recognition, virtual, Jun. 2020, pp. 5203-5212.

[12] V. Vapnik et. al., “Pattern recognition using generalized portrait method,” Automation and Remote Control, vol. 24, pp. 774-780, 1963.

[13] T. Chen et al., “XGBoost: A scalable tree boosting system,” in Proc. International Conference on Knowledge Discovery and Data Mining, Porto, Aug. 2016, pp. 785-794.

[14] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proc. International Joint Conference on Artificial Intelligence, Montreal, 1995, vol. 2, pp. 1137-1143.

[15] S. Narasimhaswamy et al., “Contextual attention for hand detection in the wild,” in Proc. International Conference on Computer Vision, Long Beach, Oct. 2019, pp. 9567-9576.

執筆者紹介

若井　信彦　Nobuhiko Wakai
プラットフォーム本部 くらし基盤技術センター
Lifestyle Foundational Technology Center, Platform Div.
博士（科学）

小塚　和紀　Kazuki Kozuka
テクノロジー本部 デジタル・AI技術センター
Digital and AI Technology Center, Technology Div.
博士（工学）

中田　洋平　Yohei Nakata
テクノロジー本部 デジタル・AI技術センター
Digital and AI Technology Center, Technology Div.
博士（工学）

飯田　恵大　Shigehiro Iida
プラットフォーム本部 くらし基盤技術センター
Lifestyle Foundational Technology Center, Platform Div.