宅内デジタルツインを活用した 環境センサによる生活行動認識手法の検討

菊池尊勝 †1
奈良先端科学技術大学院大学

松井智一 †2
奈良先端科学技術大学院大学

諏訪博彦 †3
奈良先端科学技術大学院大学

安本慶一 †4
奈良先端科学技術大学院大学

1. はじめに

宅内行動認識は，スマートホーム，ヘルスケア，介護支援など幅広い応用が期待される技術であり，住環境における生活の質を向上させるための重要な要素とされている．特に，高齢者の見守りや健康管理の分野では，行動認識技術を活用することで，異常行動の検知や生活習慣の改善支援が可能となる．しかし，従来の行動認識手法は，主にカメラ，マイク， 慣性計測装着（IMU）といったセンサデータを活用することが一般的であり，これらの技術により高い識別精度が実現されている．一方で，これらの手法は，プライバシーの侵害や心理的抵抗といった課題を抱えており，特に家庭環境においては実用化の障壁となることが指摘されている．そのため，プライバシーを保護しつつ行動認識を可能とする手法の開発が求められている．

このような背景から，カメラやマイクを使用せず，温湿度センサや人感センサなどの環境設置型センサ（以下，環境センサとする）を活用して行動認識する手法が注目されている．環境センサは，家庭内の行動を間接的にセンシングし，プライバシーの保護を図ることができるため，家庭環境に適した技術である．しかし，環境センサを用いた行動認識には，以下の３つの課題が存在する．

第一の課題は，環境センサの情報が限定的であるため，単体では高精度な行動識別が難しい点である．環境センサは，行動の直接的な特徴を捉えるのではなく，周囲の環境変化を通じて間接的に行動を推定するため，時系列性が強く，短期間のデータのみでは十分な識別が困難である．

第二の課題は，家庭環境毎にセンサ配置や家具レイアウトが異なるため，モデルの適用範囲が制限され，汎化性能が低下する点である．環境センサの設置場所が異なると，同様の行動であっても取得されるセンサデータに違いが生じるため，学習したモデルが他の家庭環境に適用しづらいという問題がある．

第三の課題は，行動認識モデルの学習に必要な大規模データの収集が困難であり，特に実環境でのデータ取得には高いコストがかかる点である．異なる家庭環境で十分なデータを収集するには時間とリソースが必要であり，居住者の負担も大きい．

本研究では，これらの課題に対処するために，環境センサを活用したマルチモーダル行動認識の統合手法，家庭環境間の違いを吸収可能な汎化モデルの設計，デジタルツインを活用したデータ拡張の検討を行う．

環境センサの情報が限定的である問題に対しては，統一表現空間を構築し，クロスモーダル対照学習を適用することで，環境センサの情報を他のモーダリティと統合し，識別精度を向上させる．また，家庭環境毎のセンサ配置や家具レイアウトの違いによる汎化性能の低下に関しては，家庭特性エンコーダと居住者特性エンコーダを導入し，家庭環境毎の差異を学習することで汎化性能を向上させる．さらに，データ収集のコストが高いという課題に対しては，デジタルツインを活用し，家庭環境を仮想的に再現することで，多様なデータを生成し，学習データの拡張を行う．

本稿の構成は，次のとおりである．2 章では，関連した既存研究を紹介し，本研究の位置付けを明らかにする．3 章では提案するマルチモーダルセンシングを利用した宅内でのマイクロ行動認識汎化モデルの開発手法について説明し，4 章では今後の課題と研究の展望について述べる．5 章では，結論を述べる．

Conference Manuscript Format for Academy of Behavior Transformation by AIoT (BTI)

†1 TAKAMASA KIKUCHI, Nara Institute of Science and Technology
†2 TOMOKAZU MATSUI, Nara Institute of Science and Technology, 理化学研究所革新知能統合研究センター (AIP) RIKEN Center for Advanced Intelligence Project AIP
†3 HIROHIKO SUWA, Nara Institute of Science and Technology, 理化学研究所革新知能統合研究センター (AIP) RIKEN Center for Advanced Intelligence Project AIP
†4 KEIICHI YASUMOTO, Nara Institute of Science and Technology, 理化学研究所革新知能統合研究センター (AIP) RIKEN Center for Advanced Intelligence Project AIP

2. 関連研究

本章では，宅内行動認識に関する既存研究を整理し，本研究がどのような位置付けにあるのかを明確にする．まず，宅内での行動認識技術の基盤となる従来手法を整理する．次にマルチモーダル統合を活用した行動認識研究について述べる．

2.1. 単一モーダリティでの行動認識に関する研究

行動認識の研究は，主にカメラや音声センサ，IMU といったモーダリティを活用する手法が一般的である．これらのデータを用いることで高い認識精度を達成することが可能となるが，それぞれの手法には利点と欠点が存在する．本節では，各モーダリティの従来手法を整理し，それぞれの特徴と課題を明らかにする．

2.1.1. IMU・環境センサを用いた行動認識

IMU ベースの行動認識は，カメラや音声と比較してプライバシーの問題が少なく，ウェアラブルデバイスを用いることで継続的なデータ収集が可能であるため，多くの研究が行われている．[1, 2]

IMU ベースの手法は，モーションデータの連続的な手法が可能である一方で，デバイスの装着を前提とするため，利用者の負担が大きく，デバイスの装着位置によるデータのばらつきが認識制度に影響を与えるという課題や，明確な動作を伴わない特定の行動の認識が他モーダリティと組み合わせないと困難であるという課題がある．

環境センサを利用した行動認識手法は，カメラや音声センサを用いることに抵抗がある家庭環境においては，プライバシーを保護しつつ行動認識を行う手段として注目されている．[3, 4, 5]

環境センサを活用した手法は，取得される情報が間接的であり，行動の発生時刻や持続時間などの精度が他のモーダリティと比較して劣るという課題や，センサ配置に依存しやすく，家具のレイアウト変更やセンサの検出範囲外での行動の認識ができないという課題が存在する．そのため，他のセンサと統合させることで認識精度を向上させる手法が求められる．

2.1.2. カメラ・映像・音声を用いた行動認識

カメラベースの手法は，高い認識精度を実現することが可能であるため，多くの研究が行われている．[6, 7, 8, 9]

カメラベースの手法は，視覚的な情報を直接取得できるため高精度な認識が可能であるが，プライバシーの懸念や心理的な負担が強いため家庭環境への導入には慎重な検討が必要である．また，カメラの設置位置に依存するため，視野外の行動に対する頑健性が低いと言う課題も存在する．

音声ベースの手法は，カメラと比較するとプライバシーの懸念が低く，IMU と違いデバイスを装着の必要がないため，多くの研究が行われている．[10, 11]

音声　ベースの手法は，背景ノイズの影響を受けやすい点や，同じ行動でも異なる音のパターンを持つ可能性があるため，データの一般化が難しいという課題が存在する．また，複数の居住者がいる環境では，個々の行動を認識するための追加の情報が必要となる課題も存在する．

2.2. マルチモーダル統合を活用した行動認識に関する研究

マルチモーダル統合は，異なるセンサから得られる情報を統合することで，単一モーダリティでは得られない詳細な行動情報を取得し，認識精度を向上させる手法である．近年，IMU や映像，音声などのモーダリティを組み合わせた行動認識の研究が進められており，環境センサと統合することでその有効性をさらに高めることが期待されている．本節では，各モーダリティ統合に関する研究を整理する．

映像データを活用したマルチモーダル統合の研究も活発に行われている．Wang ら [12] は，大規模な映像データを活用し，異なるモーダリティの統合を可能とする基盤モデルを提案した．また，Gao ら [13] は，映像とテキスト情報を組み合わせることで，行動の時間的特徴を効果的に捉える手法を提案した．これらの研究は，映像を中心としたマルチモーダル統合が行動認識の精度向上に寄与することを示しているが，プライバシーの観点から家庭環境での実用化には課題が残されている．

IMU と他のモーダリティを統合する手法も多く提案されている．Chatterjee ら [14] は，IMU と音響データを統合することで， IMU 単体では認識が困難であった行動を補完する手法を提案した．この研究は， 身体動作の詳細な情報を取得するための IMU の有用性を示しつつ，追加センサ情報による補完が精度向上に寄与することを示している．しかし，音響データの収集はプライバシーの懸念を伴うため，家庭内での実用化には課題が残されている．

音声データとの統合に関する研究も進められている．Liang ら [11] は，大規模なオンラインビデオデータから音響特徴を学習し，行動認識に適用する手法を提案した．実世界の多様な音響環境を反映したデータを学習することで，より汎用的な音響特徴量の抽出が可能となり，行動認識の適用可能範囲が広がることが示されている．しかし，大規模データを用いることでモデルの表現力は向上するが，学習コストが増大することにより，リアルタイムでの行動認識は困難であるという課題が残されている．

これらの研究に共通する課題として，モーダリティ間の情報補完が不十分である点が挙げられる．従来の手法では，

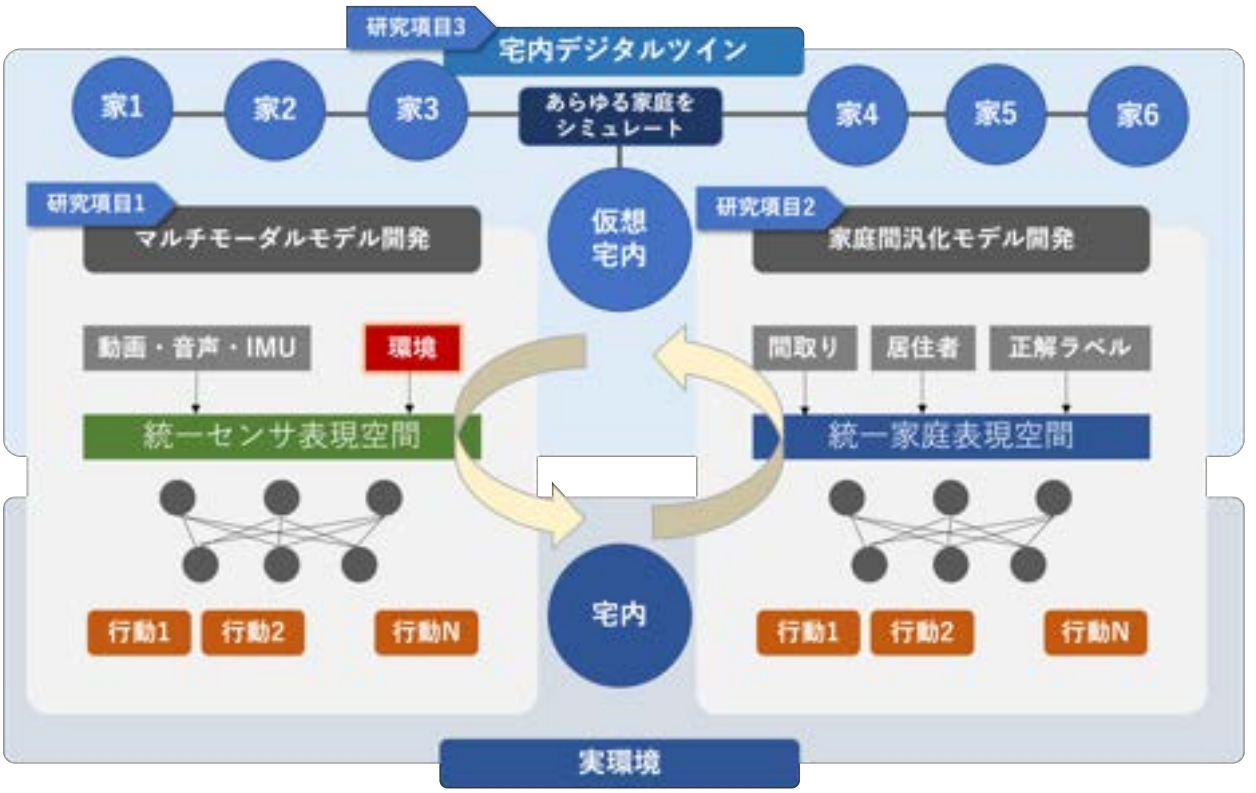

図 1　本研究の全体像

単純な特徴レベルでの統合や，意思決定レベルでの統合が主流であり，環境センサの限界を根本的に補完するには至っていない．また，環境センサを活用したマルチモーダル学習において，統一表現空間を構築する研究が少なく，環境センサの情報を効果的に統合するための新たな手法が求められている．

2.3.　本研究の位置付け

本研究では，マルチモーダルセンサを活用し，環境センサのみの情報から高精度な行動認識を実現する汎用的なモデルを構築することを目的とする．従来の行動認識手法では，IMU，ビデオ，音声といったリッチな情報を含むデータを活用することで高精度な認識が可能となっていたが，それらのモーダリティが常に利用であるとは限らず，特に家庭環境においてはプライバシーの観点からカメラや音声の使用が制限されることが多い．そのため，センサの情報量が限られてる環境でも高精度な行動認識を実現するための手法が求められている．また，家庭毎にセンサの配置や家具のレイアウトが異なることから，学習済みモデルを異なる家庭環境に適用する際の汎化性能にも限界がある．加えて，行動認識モデルの学習には大量のデータが必要となるが，実際の家庭環境でのデータ収集には高いコストがかかるため，十分な学習データを確保することが困難である．

本研究は，これらの課題を解決するために，環境センサの情報を用いた高精度な行動認識を実現し，異なる家庭環境への適応性能を向上させるとともに，デジタルツインを活用したデータ拡張による学習効率の向上を図る．この目的を達成するために，本研究では，マルチモーダル統合の最適化，家庭環境間の適応，デジタルツインによるデータ拡張の３つの要素を組み合わせたアプローチを提案する．

3.　提案手法

本研究の全体像を，図 1 に示す．本研究は，環境センサを活用した行動認識のフレームワークを提案する．本章では，マルチモーダル統合，家庭間汎化，およびデジタルツインによるデータ拡張の 3 つの観点から，課題を解決するためのアプローチを検討する．

3.1.　マルチモーダル学習に向けたセンサデータ統合

クロスモーダル対照学習と知識蒸留を組み合わせることで，統一表現空間を構築するマルチモーダル学習フレームワークを提案する．本手法は，ビデオ，音声，IMU，および環境データを統一的なベクトル空間にマッピングし，モダ

表 1　各モーダリティに適用するエンコーダ

モダリティ	エンコーディング手法
環境センサ	1D-CNN, GRU
IMU	LSTM, CNN
画像	ResNet, ViT
音声	Transformer, Wav2Vec

リティ間の整合性を確保しつつ，モーダリティの欠損に対して頑健な特徴学習を実現することを目的とする．

従来のマルチモーダル学習では，画像や音声といった主要なモーダリティに関しては統一表現空間を学習する手法が提案されている [15] が，環境センサを統一表現空間に統合する手法は少ない．本研究では，教師モデルを用いた統一表現空間の学習と，クロスモーダル対照学習を組み合わせることで，異なるモーダリティ間の関係を強化し，環境センサのみでも高精度な推論を可能にする．

3.1.1.　特徴抽出

本手法では，環境センサ， IMU ，画像，音声データを統一表現空間に適応させるために適切なエンコーダを適用し，それぞれのモーダリティから高次元特徴を抽出する．それぞれのセンサに適したエンコーディング手法を表 1 に示す．

各モダリティの特徴は，それぞれのエンコーダを通じて抽出され，統一表現空間に変換される．

3.1.2.　統一表現空間の学習

環境センサの情報のみを用いて高精度な行動認識を実現するため，マルチモーダル統合を行い，統一表現空間を構築する．統一表現空間の学習においては，以下の 2 つの手法を統合し，異なるモーダリティ間の相関を考慮した特徴表現を学習する．

1. **クロスモーダル対照学習**
   クロスモーダル対照学習では，各モーダリティの特徴を統一表現空間にマッピングし，異なるモーダリティ間の整合性を向上させることを目的とする．具体的には，同じサンプルに対する異なるモーダリティの埋め込みが統一表現空間内で近接するように学習し，異なるサンプルの埋め込みとは十分な距離を保つことで，モーダリティ間の相互関係を強化する．これにより，統一表現空間内で異なるモーダリティ間の特徴が整合性を持ち，欠損モーダリティに対しても補完的な情報を提供できる．
2. **知識蒸留**
   知識蒸留では，学習時にすべてのモーダリティを用いて教師モデルを学習し，推論時には環境センサのみを利用する生徒モデルを構築する．教師モデルは統一表現空間を学習し，生徒モデルはその表現を模倣することで，欠損モーダリティの影響を軽減する．蒸留損失には，教師モデルと生徒モデルの出力分布の差を最小化するために Kullback-Leibler (KL) ダイバージェンスを適用し，生徒モデルが教師モデルの確率分布を再現できるように最適化する．これにより，生徒モデルは環境センサのみを用いる場合でも，統一表現空間を通じて他モーダリティの情報を活用し，識別精度を維持できる．

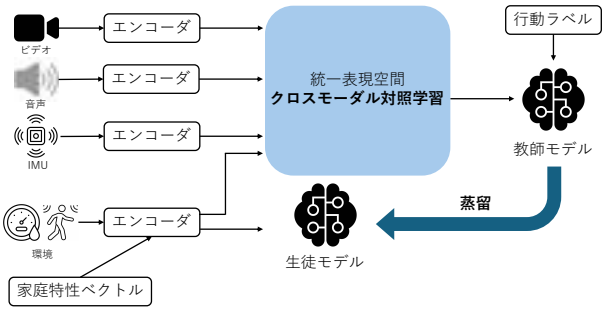

図 2　マルチモーダル学習の全体像

3.1.3.　マルチモーダル学習の全体像

マルチモーダル学習の全体像を図 2 に示す．それぞれのセンサデータに対してエンコーディングを行い，統一表現空間にマッピングを行う．環境センサは情報量が少ないため，家庭特性ベクトルを追加して，情報量を補完する．その後，蒸留を用いて環境センサのみで認識を行うモデルの構築を行う．以下に，手法のフローを示す．

1. **教師モデルの学習**
   すべてのモーダリティの特徴を統一表現空間に投影し，クロスモーダル対照学習を適用することで，統一的な表現を学習する．
2. **生徒モデルの学習**
   モーダリティを部分的に削減した状態で学習し，統一表現空間を通じて教師モデルの知識を模倣する．
3. **知識蒸留の適用**
   生徒モデルが教師モデルの出力を再現するように蒸留損失を適用し，モーダリティ欠損時の影響を最小限に抑える．

3.2.　家庭環境間の汎化モデル

家庭毎の環境特性や居住者の行動パターンは，行動認識モデルの汎化性能に影響を与える．間取りやセンサ配置の違いは，同じ行動に対するセンサの応答を変化させ，居住者

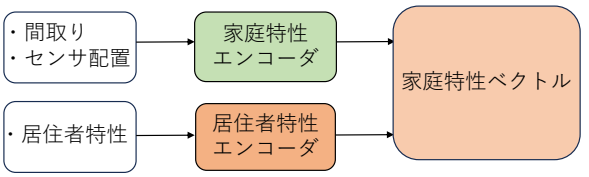

図 3　家庭環境間の汎化モデルの全体像

の人数や生活リズムの違いは行動パターンのばらつきを生む．これにより，学習済みモデルの適応性が制限され，認識精度が低下する．

家庭環境間の汎化モデルの全体像を図 3 に示す．間取りやセンサ配置などの環境要因を学習する家庭特性エンコーダと，居住者の行動特性を考慮する居住者特性エンコーダを統合することで，環境間の違いを補正する手法について検討を行う．エンコーダで得られるデータで環境センサの情報量を補完することで，環境センサのみを用いた行動認識の汎化性能を向上させる．

3.2.1.　家庭特性エンコーダ

家庭環境の違いを考慮し，行動認識モデルの汎化性能を向上させるために，家庭特性エンコーダを構築する．このエンコーダは，家庭の間取り，センサ配置の情報を統合し，環境の違いを補正する．以下に，各要素のエンコーディング方法を示す．

1. **間取り情報のエンコーディング**
   各部屋の用途をワンホットエンコーディングし，それを埋め込みベクトルとして学習する．さらに，部屋間の関係性を考慮するために，GNN を用いて部屋間の距離や隣接関係を学習し，空間的な構造情報を組み込む．
2. **センサ配置の補正**
   センサ配置の違いを補正するために，Transformer ベースの位置埋め込みを適用し，家庭毎のセンサ配置の影響を軽減する．

3.2.2.　居住者特性エンコーダ

家庭環境だけでなく，居住者の行動特性も行動認識モデルの汎化性能に影響を与える．そのため，居住者の人数や生活リズムの違いを考慮するために，居住者特性エンコーダを導入する．このエンコーダは，家庭毎の居住者数や行動パターンを学習し，センサデータへの影響を補正することで，モデルの適応能力を向上させる．

具体的には，居住者数を埋め込みベクトル化し，これを環境センサデータの特徴と結合してエンコーダへ入力する．この手法により，単身世帯と大家族世帯における行動パターンの違いをモデルが適応的に学習し，汎化性能を向上させることが可能となる．

3.3.　デジタルツインを活用したデータ拡張

行動認識モデルの汎化性能を向上させるためには，多様な家庭環境に適応可能なデータを学習することが不可欠である．本節では，宅内デジタルツインを活用し，家庭環境を仮想的に再現することで，データ拡張と適応学習を可能にする手法についての検討を行う．

3.3.1.　シミュレーション環境によるセンサデータ生成

家庭環境毎の違いを再現し，モデルの適応性能を向上させるために，宅内の物理環境を仮想的にシミュレーションし，センサデータを合成する．本手法では，家庭の 3D モデルを構築し，環境センサの変化や家具の配置変更を仮想的にシミュレーションすることで，実環境でのデータ収集の負担を軽減する．家庭の空間情報は，間取り図を基に 3D 構造に変換することで生成されるが，LiDAR や RGB-D カメラを用いたスキャン技術を活用することでより精密な環境再現が可能となる．この仮想空間上で，異なる家庭環境を想定した家具配置の変更やセンサの設置場所の違いをシミュレーションし，センサデータへの影響を評価する．

センサデータの生成には，物理ベースシミュレーションを適用し，環境センサデータの変動を再現する．さらにエージェントベースモデリングを活用し，仮想的な居住者の行動をシミュレーションすることで，人感センサの反応を人工的に生成する．これにより，異なる家庭環境におけるセンサデータの変化を事前に学習し，モデルの汎化性能を向上させることができる．

3.3.2.　合成データを活用したデータ拡張

デジタルツイン環境で生成されたセンサデータをそのまま学習に使用するのではなく，実際の家庭環境で収集されたデータと統合し，モデルの精度を向上させる．合成データと実データの統合を行う際に，ドメイン適応技術を活用し，異なる環境間でのデータ分布の違いを補正する．また，シミュレーション環境で生成されたセンサデータの特性が実データと異なる場合，それを補正するために，スタイル変換技術を適用し，合成データを実データに近づける．このプロセスにより，合成データの信頼性が向上し，学習データとして活用可能な質の高いデータを収取できる．また，センサノイズのシミュレーションを行うことで，実際の家庭環境で発生する測定誤差を再現し，より実データに近いデータを生成する．

3.3.3.　フィードバックループを活用した適応学習

モデルの汎化性能をさらに向上させるためには，家庭毎の特性を考慮し，モデルの動的な適応が必要である．そこで，リアルタイムフィードバックを用いた適応学習の仕組みを導入する．具体的には，行動認識結果を居住者が確認し，誤

認識が発生した場合には修正を可能とする機能を導入する．修正されたデータををデジタルツイン環境にフィードバックし，環境特性に適応したモデル更新を行う．

適応学習の実現には，転移学習を適用し，各家庭において特定の行動認識精度が向上するように微調整を行う．これにより，家庭毎の行動特性の違いを反映し，異なる環境でも高精度な行動認識を維持することが可能となる．

4. まとめと今後の展望

本研究では，環境センサを活用した行動認識手法を提案し，家庭環境間の汎化性能向上およびデジタルツインを用いたデータ拡張の可能性についての検討を行った．従来の行動認識手法では，カメラや IMU を主に活用するアプローチが一般的であり，環境センサ単体での高精度な認識手法の検討は十分に進んでいなかった．本研究では，クロスモーダル対照学習と知識蒸留を組み合わせることで，環境センサの情報のみから統一表現空間を構築し，行動認識の精度向上を目指した．また，家庭特性エンコーダと居住者特性エンコーダを導入することで，家庭毎の特性を考慮しながら汎化性能を向上させる手法を検討した．さらに，デジタルツインを活用し，仮想環境を用いたデータ拡張と適応学習の可能性についても検討を行った．

しかし，本研究は Position Paper としての議論を主としており，提案手法の定量的な評価は行っていない．提案手法の実装および評価を通じて，環境センサを用いた行動認識の実現可能性を検証し，各アプローチの有効性を定量的に評価することが今後の課題となる．また、デジタルツインを活用したデータ拡張に関しても，仮想環境で生成されたデータと実データの特性の違いを完全に解消することは難しく，特にセンサノイズや居住者の行動特性の再現性に課題が残る．

提案手法の適用範囲を拡張し，より多彩な家庭環境に対応するためには，いくつかの技術的課題を解決する必要がある．まず，家庭環境の多様性に関する問題が挙げられる．本研究では，家庭特性エンコーダを導入することで，家庭毎のセンサ配置や家具の違いを考慮したが，時間とともに変化する動的な環境の影響については十分に考慮できていない．長期的な居住環境の変化や一時的な家具の移動に対応するためには，モデルの動的な更新手法の導入が必要である．また，デジタルツインを活用したデータ拡張の精度向上も課題として挙げられる．仮想環境で生成されたデータと実データの特性の違いを完全に解消することは難しく，特にセンサノイズや居住者の行動特性の再現性に課題が残る．デジタルツイン環境の実データとの整合性を向上させるためには，合成データの適応的な補正手法の導入が求められる．さらに，実環境への適用に関しては，データ取得条件やセンサノイズの影響によるモデルの精度の変動が課題として挙げられる．提案手法では，異なる家庭環境に適応可能なモデルを構築したが，長期間にわたるモデルの維持や更新には更なる工夫が必要である．特に，居住者の行動フィードバックを活用した適応学習を行う際には，利用者に過度な負担をかけることなく，継続的なモデル更新を実現するための設計が求められる．

今後の研究においては，提案手法の更なる汎用向上と実環境への適用を目指し，いくつかの方向性が考えられる．まず，多様な家庭環境への適応性能を高めるために，より大規模なデータセットの収集と活用が必要である．本研究では，異なる家庭環境間のデータを統合するための手法を提案したが，より多様なデータを用いた学習により，異なる文化圏や生活習慣も適応可能なモデルを構築することが求められる．また，家庭毎のカスタマイズ学習を行うことで，個々の居住者の行動特性をより正確に反映した認識モデルを実現することが可能となる．さらに，環境センサだけでなく，音声データによるゲーミフィケーションを活用しおたアノーテーションインターフェースを統合することで，継続的なフィードバックに対する負担を軽減しつつ，認識モデルの精度の向上を図る．

本研究で提案した行動認識手法は，スマートホームシステム，健康管理，介護支援，支援モニタリングなど幅広い応用が期待される．今後の研究では，実環境での適用を視野に入れた実証実験を行い，提案手法の有効性を明確に示すとともに，より実用的なシステムの開発に向けた研究を進めていく必要がある．

参考文献

[1] Alevizaki, A., Pham, N. and Trigoni, N.: Hierarchical activity recognition with smartwatch IMU, *Proceedings of the 24th International Conference on Distributed Computing and Networking*, pp. 48–57 (2023).

[2] Thakur, D., Guzzo, A. and Fortino, G.: Attention-based multihead deep learning framework for online activity monitoring with smartwatch sensors, *IEEE Internet of Things Journal*, Vol. 10, No. 20, pp. 17746–17754 (2023).

[3] Matsui, T., Onishi, K., Misaki, S., Fujimoto, M., Suwa, H. and Yasumoto, K.: Salon: Simplified sensing system for activity of daily living in ordinary home, *Sensors*, Vol. 20, No. 17, p. 4895 (2020).

[4] Fujiwara, M., Kashimoto, Y., Fujimoto, M., Suwa, H., Arakawa, Y. and Yasumoto, K.: Implementation and evaluation of analog-pir-sensor-based activity recognition, *SICE Journal of Control, Measurement, and System Integration*, Vol. 10, No. 5, pp. 385–392 (2017).

[5] Guan, Q., Li, C., Qin, L. and Wang, G.: Daily activity recognition using pyroelectric infrared sensors and reference structures, *IEEE Sensors Journal*, Vol. 19, No. 5,

pp. 1645–1652 (2018).

[6] Tan, T.-H., Gochoo, M., Huang, S.-C., Liu, Y.-H., Liu, S.-H. and Huang, Y.-F.: Multi-resident activity recognition in a smart home using RGB activity image and DCNN, *IEEE Sensors Journal*, Vol. 18, No. 23, pp. 9718–9727 (2018).

[7] Gaidon, A., Harchaoui, Z. and Schmid, C.: Temporal localization of actions with actoms, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 35, No. 11, pp. 2782–2795 (2013).

[8] Choi, W., Chen, J. and Yoon, J.: PWS-DVC: Enhancing Weakly Supervised Dense Video Captioning With Pretraining Approach, *IEEE Access*, Vol. 11, pp. 128162–128174 (2023).

[9] Cho, S., Kim, Y., Jang, J. and Hwang, I.: AI-to-Human Actuation: Boosting Unmodified AI's Robustness by Proactively Inducing Favorable Human Sensing Conditions, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 7, No. 1, pp. 1–32 (2023).

[10] Wu, J., Harrison, C., Bigham, J. P. and Laput, G.: Automated class discovery and one-shot interactions for acoustic activity recognition, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14 (2020).

[11] Liang, D. and Thomaz, E.: Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 3, No. 1, pp. 1–18 (2019).

[12] Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y. et al.: Internvideo2: Scaling foundation models for multimodal video understanding, *European Conference on Computer Vision*, Springer, pp. 396–416 (2024).

[13] Gao, J., Sun, C., Yang, Z. and Nevatia, R.: Tall: Temporal activity localization via language query, *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275 (2017).

[14] Chatterjee, S., Chakma, A., Gangopadhyay, A., Roy, N., Mitra, B. and Chakraborty, S.: LASO: Exploiting locomotive and acoustic signatures over the edge to annotate IMU data for human activity recognition, *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 333–342 (2020).

[15] Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D. and Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455 (2024).