第 23 回日本バーチャルリアリティ学会大会論文集（2018 年 9 月）

VR 技術を用いたマルチモーダル アノテーションツールの提案

Proposal of Multimodal Annotation Tool Using VR Technology

亀山悟 [1)]，山崎寛 [1)]，佐々木実 [1)]，坂根裕 [1)]

1) 株式会社エクサウィザーズ　（〒105-0013 東京都港区浜松町 1-6-15 VORT I 3 階, satoru.kameyama@exwzd.com）

概要： マルチモーダルセンサデータに対して，時系列アノテーションを付与するシステムを開発した．複数のセンサデータを仮想空間上で統合することで，自由な視点および時間からデータ全体を分析しアノテーション入力できる．本研究では，分析対象として介護者および被介護者の動作データから，両者の対話に関するアノテーション入力を行いシステムの有効性を確認した．

キーワード：**マルチモーダル，機械学習，アノテーション，モデリング**

1. はじめに

近年，計算機の処理能力向上やアルゴリズム改善に伴い，深層学習を活用したさまざまな技術が実用化されている．筆者らは，介護現場における介護者と被介護者の様子から，どのようなケアが行われているか判断し，その振る舞いを評価できる技術の開発を進めている．ケアの種類や振る舞いを判断するには，介護者や被介護者の動き，周囲の状況変化など，複数対象を観察する必要がある．それゆえ，現場に設置した多数のカメラやセンサ群が処理対象となる．

ただし，入力データの次元数が高くなると学習効率が著しく低下するため，どのセンサから何を抽出したいか，入力と出力を制限することで次元数を下げるといった，人手による作業が必要となることが多い．本研究では，多種多様なセンサを対象とする複雑な状況理解モデル開発において，上記作業を支援するツールを実現する．

本稿では，複数のセンサデータを三次元空間上に展開し，同期しながら内容確認しつつ，アノテーション入力できるツールを構築した．設計および実装の詳細について述べる．

2. 多種センサを対象とした深層学習モデルの開発

筆者らが研究開発を進めている，介護現場における状況理解モデルの開発とその応用を図 1 に示す．現場に設置したカメラやセンサ群から得たデータを，処理関数（図中 F_1, F_2, …, F_N）で処理し，事前に定義したシンボルに変換する．シンボルは複数統合し，時系列データとして管理する．シンボルパターンをイベントとして抽出し，特定サービスを実行するイメージとなる．この，特定サービスを実行することをアクションと定義する．

Satoru KAMEYAMA, Hiroshi YAMASAKI, Minoru SASAKI and Yutaka SAKANE

図 2 に，本研究で開発している，センサデータからシンボルを生成するアノテーションツールの概要を示す．

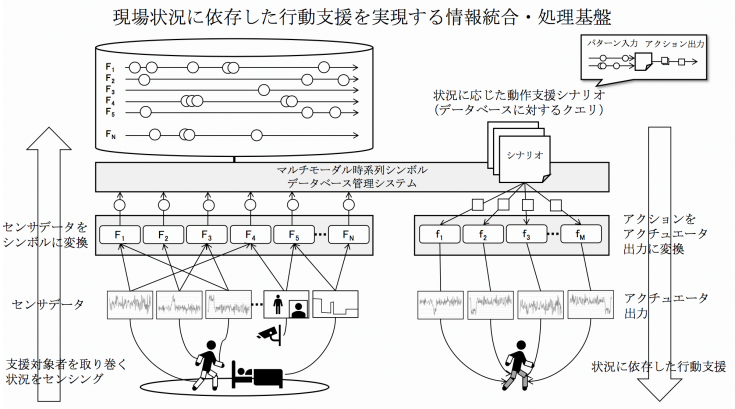

図 1：システム全体像

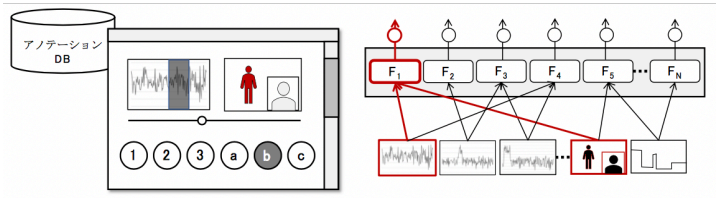

図 2：センサデータからシンボル出力する AI のための アノテーションツール（本稿範囲）

3. 関連研究

動画へのアノテーションツールに関する研究やシステム[1,2]は複数存在するが，画像や音声などの複数のセンサ情報を並べた上でのアノテーションツールは存在していない．また，3D データに対して有効なアノテーションツールも存在していない．

4. 実装

4.1 機能

提案するアノテーションツールは，以下の機能を持つ．

1. メニュー画面から必要なデータを選んで表示
2. データの表示位置や角度，スケール調整
3. データの再生，一時停止，停止，シーク
4. アノテーション画面表示のオン，オフ
5. アノテーションの付与
6. アノテーションラベルの追加，削除
7. アノテーション結果の保存

4.2 特徴

提案するアノテーションツールは，複数のセンサ情報をひとつの空間に集約する．その様子を図 3 に示す．この 2 つの動画では，天井に設置した全天球カメラと，介護者のメガネに設置したカメラで撮影した動画を並べたものである．これら 2 つの動画を同時に閲覧しながら，アノテーションが出来る．そのため，センサ間の関係を情報に含めたアノテーションが可能となる．

さらに，VR を活用することで 3D データのアノテーション付与も直観的に行える．次項にて，実際に VR を活用したアノテーションツールの使用例を述べる．

5. 実際の使用例

5.1 対象

対象として介護訓練用 VR アプリケーションを選んだ．ここでは，介護訓練用 VR アプリケーションで介護する者を訓練者，その様子を見てフィードバックを行う高度技術者を指導者と呼ぶことにする．訓練者に取り付けた複数のセンサをセンサデータとして利用する．センサは頭部・両手に取り付け，位置情報(x,y,z)および向き情報$(\theta_x, \theta_y, \theta_z)$の計 6 自由度(6 degrees of freedom; 6DoF)の情報を取得することができる．介護に必要な現場の状況を表したものをシンボル，指導者が指導した内容をアクションとしてアノテーションを実施した．

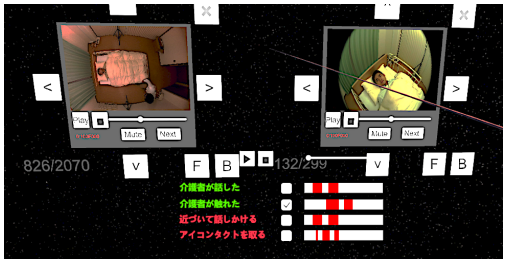

図 3: 複数動画に対するアノテーションツール画面

5.2 使用方法

以下の手順にてアノテーションを実施した．

まず，介護訓練用 VR を通して，訓練者が行った介護動作をセンシングする．

指導者はアノテーションツールを起動し，先ほどセンシングしたデータを並べて表示し，閲覧する．

指導者は，指導に必要なシンボルとアクションのラベルを追加する．その後，データをシークしながらシンボルとアクションをアノテーションする．

5.3 使用結果

上記方法により実施した結果，表 1 のデータを元に，表 1 のシンボル，アクションが得られた．また，アノテーションを完了した時のアノテーションツール画面を図 4 に示す．この画面上の顔と手だけで表示されたものは，訓練者が実施した介護行動を再生したものである．この行動を，画面上にある操作バーにより再生，一時停止，シークなどを繰り返して指導者は閲覧する．緑文字で書かれた文字がシンボルで，赤文字で書かれた文字がアクションを表す．それぞれの隣に灰色バーがあり，赤で塗った部分が，該当シンボルまたはアクションが発生したフレームに対応する．アノテーションツールは，この赤で塗られたフレームを 1，それ以外を 0 として，図 5 のような csv を出力する．

上記 csv を元に機械学習を行い，シンボルを出力する AI と，アクションを出力する AI を開発することが出来る．

表 1: データとアノテーション例

属性	内容
データ	介護者の頭と両手の 6DoF/ 被介護者の位置/ 音声データ
シンボル	介護者が話した/ 介護者が触れた
アクション	近づいて話しかける/ アイコンタクトをとる

図 4: 介護訓練用 VR アプリケーションに対するアノテーションツール画面

FrameNum	5760			
SymbolNum	2			
ActionNum	2			
Frame	介護者が話した	介護者が触れた	近づいて話しかける	アイコンタクトを取る
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
434	1	0	1	0
435	1	0	1	0
436	1	0	1	0

図 5: アノテーション結果 csv

6. アノテーションツールの有効性検証

介護訓練用 VR アプリケーションに対する，アノテーションを実施することで，これまで困難であった 3DVR データに対するアノテーションが可能となることを確認した．

今後は，さらに多くの動画データや音声データ，センサデータを組み合わせて同一空間に並べた上でのアノテーションを実施することで，より汎用性の高いツールとなるように開発を進める．

7. むすび

本研究では，マルチモーダルセンサデータに時系列アノテーションを付与するシステムを開発した．複数のセンサデータを仮想空間上で統合することで，自由な視点および時間からデータ全体を分析しアノテーション入力できる．本研究では，分析対象として介護者および被介護者の動作データから，両者の対話に関するアノテーション入力を行いシステムの有効性を確認した．

謝辞

本研究は，総合科学技術・イノベーション会議が主導する革新的研究開発推進プログラム(ImPACT)の一環として実施したものです．

参考文献

[1] Carl Vondrick, Donald Patterson, Deva Ramanan. "Efficiently Scaling Up Crowdsourced Video Annotation" *International Journal of Computer Vision* (IJCV). June 2012.

[2] David Acuna and Huan Ling and Amlan Kar and Sanja Fidler. "Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++" CVPR2018.