2024 年度卒業研究概要

動作時間範囲の自動検出機能を有するアノテーションツール

大谷 紀子 研究室
2172026　岡村 和哉

１．背景と目的

認知科学分野における行動分析では，人間の動きや行動パターンが認知プロセスや心理状態にどのように関連しているかを分析するために，動画や音声への注釈付けが行われている．注釈付けの際，ELAN などのアノテーションツールが使用されることが多いが，注釈付けは手作業で行われている．また，注釈付けには動作の詳細な記述と的確な時間範囲の設定が求められるため，多大な時間と労力が必要となる．

自動で注釈付けする方法として，Google Research が公開する Vid2seq や Streaming Dense Video Captioning 等の利用が挙げられるが，いずれも動画内でどのようなことが起きているか，人物が何をしているかなど，大まかな動作に対してのみ注釈付けする．したがって，腕の動きや顔の向きの変化など，細かな動作への注釈付けが必要となる認知科学分野での活用には適していない．

本研究では，注釈付け作業者の負担軽減および作業時間短縮を目的とする．「時間範囲の設定」と「動作の説明」という注釈付けの 2 つの工程のうち，作業者への負担がより大きい「時間範囲の設定」を対象として，動作時間範囲の自動検出機能を有するアノテーションツールを開発する．

2. ツール概要

本ツールでは，3 次元の姿勢推定の結果を基に，ユーザが指定した人物と関節に応じて動作の時間範囲を自動検出する．まず，ユーザは動画内に映る人物が 1 人か複数人かを選択する．複数人が選択された場合，2 次元の姿勢推定モデルである

図 1 人物・関節選択画面

AlphaPose[1]を用いて動画内に映る人数を特定する．その後，ユーザは図 1 のようなインタフェースを通じ，着目したい内容ごとに任意の人物と任意の関節を選択する．ユーザが選択できる関節は右肩，右ひじ，左肩，左ひじ，右腰，右ひざ，左腰，左ひざ，背中，首の計 10 か所ある．関節は複数選択が可能であり，複数選択した場合，それぞれの関節の動作時間範囲が合成される．

AlphaPose により生成した 2 次元の姿勢推定データを用いて，MotionBERT[2]により 3 次元の姿勢を推定する．MotionBERT では，体を 17 のキーポイントで表現しており，動画の各フレームにおいて各キーポイントの 3 次元座標を出力する．座標のデータから本ツールで指定が可能な 10 か所について，それぞれ隣接する 2 点との間で外積を計算し，外積の値の変化を時間軸方向に分析することで動作範囲を推定する．さらに，各キーポイントで x, y, z 軸方向それぞれの値の最大・最小の差を計算し，差が大きい軸を最大 2 つ選定する．選定した軸について，フレームごとに次のフレームとの値の差が一定の閾値を超えた場合，「動作あり候補」と判定する．さらに，「動作あり候補」が一

定以上のフレーム数連続している場合，連続するフレーム全体を「動作あり」とみなす．最終的に「動作あり」となった範囲が動作範囲として検出される．検出された動作範囲に応じて本ツール上に注釈が生成される．最後に，ユーザは必要に応じて注釈範囲の変更や動作説明テキストを書き入れ，注釈付けを完了させる．

3. 評価実験

ELAN を熟知した相互行為分析の専門家 1 名を被験者とし，評価実験を実施した．2 人の人物があっち向いてホイをしている動画を対象に，それぞれの人物に対して顔の向きの変化と腕の動きについて注釈を付けさせる．まず，ELAN を使用した注釈付けをさせ，その後，本ツールを使用して注釈付けをさせた．両ツールでの注釈付けが完了した後，アンケートと半構造化インタビューを実施した．アンケートでは，5 段階評価を用いて本ツールの使いやすさ（1=使いにくい，5=使いやすい）や，負担軽減度合い（1=軽減されなかった，5=軽減された）などについて調査した．半構造化インタビューでは，本ツールの良かった点や改善点，追加機能の要望などについて尋ねた．

アンケートの結果，負担軽減度合いを問う 5 段階評価では，5 の評価を受けた．また，半構造化インタビューの結果，「本ツールの良かった点は何か」という質問に対し，顔の向きの変化に関する「上を向く」，「正面を向く」などの注釈の時間範囲はほとんど手を付ける必要がなかったとの肯定的な意見が得られた．加えて，「追加してほしい機能は何か」という質問に対して，普段からアノテーションツールを使用する人の多くが利用する，指定範囲を繰り返し再生できる部分再生機能が欲しいとの回答があった．ELAN での作業時間と本ツールでの作業時間を比較した場合，ELAN での作業時間は 24 分 53 秒で，本ツールでの作業時間は 41 分 45 秒となり，本ツール使用時のほうが長くなった．しかし，本ツールでの作業時間には，作業者が作業していない動作時間範囲の検出処理時間が含まれるため，実質的な作業時間は 19 分 41 秒となり，ELAN での作業時間よりも短くなっている．

また，両手の動きに注釈がつけられている動画に対して，本ツールを用いて動作時間範囲を検出し，時間範囲のずれを調査した．既存の注釈では，「ものを持つ」という動作が，持っている時間全体に含まれていたが，本ツールで検出された動作範囲は，持ち始める部分だけで，持って静止している範囲は検出されなかった．

4. 考察

今回の評価実験では，約 17 秒の動画を使用したが，認知科学分野で分析の対象となる動画は数分の場合もあり，長い動画であるほど作業時間は多く短縮されると考えられる．したがって，本ツールによって作業者の負担を軽減できるといえる．3 次元の姿勢推定に MotionBERT だけを使用しているため，手首や指の動きなどを検出することが困難である．認知科学分野では，指の動きに着目することが多いため，他のモデルを併用することによる動作範囲検出の対象となるキーポイントの増加や検出精度向上が望まれている．また，静止している動作の範囲には注釈がつかないことがあるため，目的や場面に応じた使い方が求められる．本ツールの機能面では，部分再生機能の追加が必要だと考える．

参考文献

[1] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, C. Lu, "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.7157-7173, 2022.

[2] W.Zhu, X.Ma, Z.Liu, L.Liu, W.Wu, Y.wang, "MotionBERT: A Unified Perspective on Learning Human Motion Representations", Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.15085-15099, 2023.