H-005

人物姿勢推定を用いた動画結合手法
Human Pose Estimation-Based Video Fusion

沈 展帆[1)]　五十川 麻理子[2)]
Zhanfan Shen　Mariko Isogawa

1　抄録

近年，スマートフォンや小型カメラの性能向上およびその普及によって，自身や友人，家族が映っている動画を撮影し，自らが楽しむだけでなく，SNS に投稿したり他者と共有したりするユーザが増加している．その際，一般ユーザにとっては一度の撮影で所望の動画を撮ることが困難であるため，複数の短い動画を結合するケースが多い．しかし，人物が映り込んだ動画を直接結合すると，人物の姿勢や位置が急激に変化したり，時系列的な不整合が生じてしまうという課題がある．本研究では，動画の結合部分における人物姿勢情報の補間と，その情報に基づく人物領域生成により，人物動画をスムーズに結合することを目指す．

2　はじめに

近年のスマートフォンや小型カメラの性能向上により，専用の撮影機材がなくとも高品質な動画を誰でも撮影できるようになってきた．特に中高生の間では，自身や友人，家族が映っている日常生活動画を撮影し，その動画を SNS に投稿したり友人同士で共有し合うことは日常茶飯事である．その際，撮影スキルの乏しい一般ユーザにとっては，一度の撮影で所望の動画を撮影することは困難であるため，複数の短い動画を撮影することが多い．そのような複数の動画は，最終的に他者に共有することや SNS 等にアップロードすることが前提であるため，単一の動画に結合できることが望ましい．

複数動画を結合するための最も単純な方法は，動画編集ソフトウェア等を用いて動画を繋げるという方法である．しかし，人物が映り込んだ動画を直接結合すると，人物の姿勢や位置が急激に変化したり，時系列的な不整合が生じてしまうという課題がある．従来，動画結合や動画中の時系列フレームを補間する際に，時系列的な一貫性を保つことで違和感のない動画を生成する手法がいくつか提案されている．Dong らによって提案された映像の各画素の輝度値を時間的連続性を考慮することで動画全体の時系列的な一貫性を保つ動画処理手法 [2] をはじめとして，Yan らは，時系列的な依存関係をモデル化するための Transformer モデルを導入することで長時間にわたる時系列的な一貫性を保ちつつ動画生成を行う生成モデルを提案した [4]．しかしこれらの手法は，動画中に映り込んだ動作中の人物のような，比較的急な動きを有する領域が存在するシーンを想定していない．

そこで本研究は，ユーザが短い人物動画を入力するだけで，動画にある人物姿勢情報などを用いて，動画間の人物姿勢を推定・生成し，最終的に一つの長い動画を出力することを目指す．具体的には，動画に映り込んだ人物の姿勢情報を抽出し，動画間の姿勢情報をスムーズに結合・補完し，その姿勢情報に基づいて人物の外観を生成することで動画を結合する手法を提案する．

1）　大阪府立佐野高校
2）　慶應義塾大学

まとめると，本研究の技術貢献は以下である．(1) 人物が映り込んだ複数の動画を，人物動作が時系列的に一貫するように結合するという新規タスクに取り組む．(2) そのために，複数動画中の人物姿勢を表す関節位置を時系列的に滑らかに結合するための補間手法を提案する．

3　提案手法

図 1 に提案フレームワークを示す．本研究の目的は，2 つの RGB 動画像 $\mathbf{m^1} = [m_1^1, m_2^1, ..., m_T^1]$ および $\mathbf{m^2} = [m_1^2, m_2^2, ..., m_T^2]$ を入力として，その中に映り込んだ人物のアピアランスを時系列的に一貫させた結合動画 M を生成することである．なお，T はシーケンス長を指す．提案フレームワークは，$\mathbf{m^1}$ および $\mathbf{m^2}$ を構成する各画像フレーム毎に映り込んだ人物の姿勢を推定する姿勢推定モジュール，2 つの動画の結合時刻前後において人物姿勢を滑らかに繋ぐ姿勢補間モジュール，そして滑らかに繋がった人物姿勢に基づいて人物動画のアピアランスを合成する動画生成モジュールから構成される．

具体的な処理の流れとしては，まず $\mathbf{m^1}, \mathbf{m^2}$ を入力として，姿勢推定モジュールにて動画中の人物領域を抽出するとともに，人物の二次元姿勢を推定する．次に，姿勢補間モジュールにおいて，これら二つの動画の人物姿勢を時系列的に滑らかに連続するように結合する．最後に，動画生成モジュールにおいて，結合した姿勢情報に条件付けされた人物のアピアランスを生成することで，結合動画を得る．以降では，これらの各モジュールについて，それぞれ述べる．

3.1　姿勢推定モジュール

二つの異なる動画 $\mathbf{m^1}$，$\mathbf{m^2}$ とを連結するために，まずは両動画フレーム中に映り込んだ人物の姿勢を，画像中の二次元関節位置 $\mathbf{p^1} = [p_1^1, p_2^1, ..., p_T^1]$ および $\mathbf{p^2} = [p_1^2, p_2^2, ..., p_T^2]$ として算出する．姿勢推定には，RGB 画像を入力として画像中の二次元人物姿勢を出力する既存モデルである，YOLOv7 [3] を用いる．

3.2　姿勢補完モジュール

滑らかに連結された人物動画 $\mathbf{m^m}$ を生成するために，本手法ではまず，動画 $\mathbf{m^1}$，$\mathbf{m^2}$ 中の人物姿勢 $\mathbf{p^1}$，$\mathbf{p^2}$ を時間軸方向に結合（マージ）させた，連結人物姿勢情報 $\mathbf{p^m} = [p_1^m, p_2^m, ..., p_{2T}^m]$ を生成する．このとき，最も単純な結合方法は，$\mathbf{p^1}$ および $\mathbf{p^2}$ をそのまま時間軸方向に連結するものである．しかし，これでは結合周辺のフレームにおいて，人物の位置や姿勢が急激に変化するなどの不連続性が生じてしまう．そこで提案手法では，以下の方法により，時間的に滑らかな $\mathbf{p^m}$ を生成することを目指す．

まず，$\mathbf{p^1}$，$\mathbf{p^2}$ との間に，N フレーム分の姿勢情報 $\mathbf{p^i} = [p_1^i, p_2^i, ..., p_N^i]$ を挿入する．この挿入した姿勢情報の値を決定するために，挿入フレーム前後の姿勢情報が存在するフレーム，つまり，$\mathbf{p^1}$ の最終フレームにおける姿勢 p_T^1 および $\mathbf{p^2}$ の開始フレームにおける姿勢情報 p_1^2 を活用する．具体的には，p_T^1 と p_1^2 との対応する関

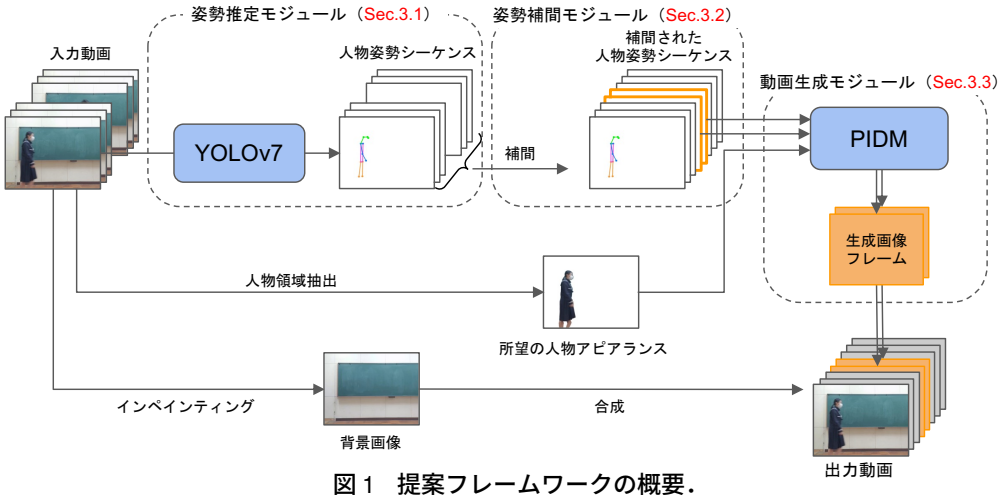

図 1　提案フレームワークの概要．

節座標位置とを繋ぐ線分を $N+1$ 等分する内分点 N 点を取り，その N 点を時系列的に連続するように $\mathbf{p}^{\mathrm{i}}$ に割り当てる，という方法を取る．なお，ここで生成した人物姿勢 $\mathbf{p}^{\mathrm{m}}$ については，この後画像生成モデルに入力するために，入力画像サイズに合わせて空間位置の正規化を行なっている．

3.3　動画生成モジュール

姿勢補間モジュールでは姿勢情報（骨格情報）のレベルで二つの動画を連結したが，最終的に連結した RGB 動画を得るためには，この姿勢情報に基づいて画像フレームを生成する必要がある．そこで本手法では，人物姿勢情報と，合成したい人物のアピアランス情報の手がかりであるソース画像とを入力として，所望の人物のアピアランスを所望の姿勢で合成する手法である PIDM (Person Image Synthesis via Denoising Diffusion Model) [1] を用いて $\mathbf{m}^{\mathrm{m}}$ を生成する．ソース画像としては $\mathbf{p}^{1}$ の開始フレーム $\mathbf{p}_{1}^{1}$ 中に含まれる，YOLO によって抽出された人物領域を用いる．なお，$\mathbf{m}^{\mathrm{m}}$ を生成するために，人物のアピアランスを生成した後に元動画と一貫した背景画像にその人物領域を合成する必要がある．その際の背景画像としては，抽出された人物領域をマスクとして画像修復手法の一種である画像インペインティングを施すことで生成する．

4　実験および結果

4.1　データセット

連結対象となる動画データセットを作成する目的で，1 名の人物が画面内で動作を行う様子を，固定された RGB カメラを用いて撮影した．被験者には，歩行，走行，立ち上がりなどの動作を行なってもらい，それぞれの動画を前半と後半に分割した．各動画は 30fps で撮影され，約 3 秒から 5 秒の動画クリップの撮影を行った．関節点としては，鼻・首・両肩・両肘・両手首・両股関節・両膝・両足首・両目・両耳の計 18 関節を使用した．

4.2　動画生成実験

提案手法は新規タスクに取り組んでいるため，同じタスクに取り組む従来手法が存在しない．そのため，姿勢情報レベルでの補間による提案手法の有効性を調査する目的で，2 つの動画を単純に連結することを想定した人物姿勢フレームおよび，本手法によって補完した人物姿勢フレームを挿入した結果を比較した．その結果を図 2 に示す．図より，今後この姿勢情報に基づいた動画生成

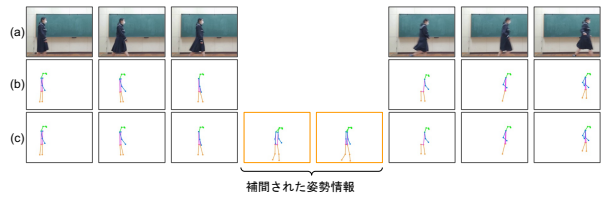

図 2　提案手法により補間した人物姿勢情報．(a) 連結元の動画，(b) 連結元の動画中の人物姿勢情報，(c) 提案手法により補間された人物姿勢情報．

を行う必要はあるものの，少なくとも姿勢情報のレベルにおいては，提案手法を用いることでより連続した動画連結が実現できる可能性があることが示唆された．

5　結論と今後の展望

本稿では，ユーザが撮影した短い人物動画を滑らかに結合することを目的として，動画に映り込んだ人物の姿勢情報を抽出し，動画間の姿勢情報をスムーズに結合・補完し，その姿勢情報に基づいて人物の外観を生成することで動画を結合する手法を提案した．また，人物姿勢補完までの結果を報告した．今後の展望としては，フレームワーク全体を完成させるとともに，更なる生成動画の品質向上に取り組む予定である．また，生成された動画の定量的・定性的な品質評価を行うことで提案手法の有効性を調査する予定である．

参考文献

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5968–5976, 2023.

[2] Xuan Dong, Boyan Bonev, Yu Zhu, and Alan L. Yuille. Region-based temporally consistent video post-processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 714–722, 2015.

[3] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2023.

[4] Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation. In *International Conference on Machine Learning (ICML)*, pages 39062–39098. PMLR, 2023.