複数人物の感情表現を考慮した人物間の動作類似性を抑制する動画生成

Emotion-Aware Multi-Person Video Generation with Inter-Person Motion Similarity Suppression

坂田　真乙†　　渡辺 裕†
Mao Sakata†　and　Hiroshi Watanabe†

†早稲田大学
†Waseda University

Abstract　This study proposes a pipeline that separates multiple characters from a single image, generates individual videos with emotion-specific prompts, and recombines them with the original background. Experiments using pose-derived velocity, acceleration, and Dynamic Time Warping suggest that the method better represents character-specific motion intensity than whole-image generation.

1. まえがき

近年，画像から動画を生成する動画拡散モデルの研究が進展している．FramePack は，過去フレームの情報を圧縮して保持することで，限られた VRAM 環境でも長尺動画を生成できる手法である［1］．

一方，複数人物を含む画像では，人物ごとに異なる感情や動作を指定することが難しい．画像全体に単一のプロンプトを与えると，複数人物が類似した動作を示したり，感情表現が均一化したりする場合がある．複数人物を制御する動画生成手法も提案されている［2］が，静止画像内の各人物に異なる感情に応じた動作を付与する方法は十分に検討されていない．

そこで本研究では，画像内の人物を分離し，人物ごとに異なる感情プロンプトで動画を生成した後，背景上に再合成するパイプラインを提案する．生成動画から姿勢情報を抽出し，平均速度，平均加速度および人物間の動作類似度によって有効性を評価する．

2. 提案手法

提案手法の概要を図 1 に示す．本手法は，人物・背景分離，人物ごとの動画生成，動画合成の 3 段階から構成される．

まず，複数人を含む入力画像に CartoonSegmentation［3］を適用し，背景と各人物領域を分離する．各人物のバウンディングボックスを取得して個別画像として切り出し，人物を除いた領域を背景画像として保持する．

次に，各人物画像を FramePack に入力し，人物ごとに感情と動作を記述したプロンプトを与える．本研究では基本動作を「踊る」に固定し，「怒りながら踊る」「悲しみながら踊る」など，感情のみを変更した．

人物ごとに動画を生成することで，他の人物の影響を抑え，各プロンプトを独立して反映させる．

最後に，生成した人物動画を切り出し時の座標に基づいて元の位置へ配置し，背景画像とフレーム単位で合成する．これにより，同一画面内の人物が異なる感情に基づく動作を示す動画を生成する．

3. 実験条件および評価方法

提案手法の有効性を検証するため，入力画像全体を FramePack に与える比較手法と比較した．

左右に 2 人の人物が描かれた 4 種類の画像を使用し，各人物に anger，sadness，excitement，relaxation の 4 感情を割り当てた．左右の感情の 16 通りの組合せについて 2 手法で生成し，合計 128 本の動画を得た．基本動作はすべて「踊る」とした．

生成動画に ViTPose［4］を適用して関節座標を推定し，位置情報に基づいて左右の人物を追跡した．動作強度の指標として，関節座標のフレーム間変位から平均速度を，その変化から平均加速度を算出した．

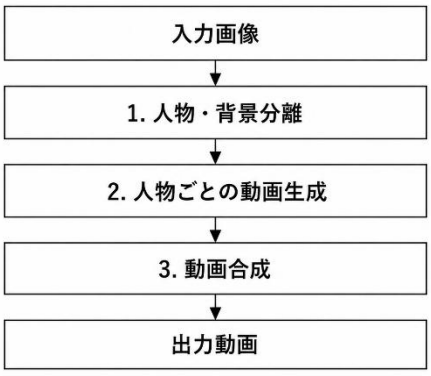

図 1　提案手法の概要

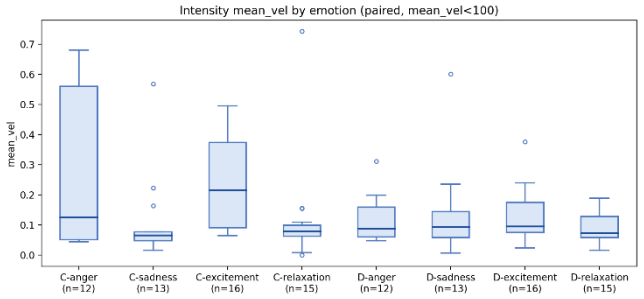

図 2　手法・感情条件別の平均速度分布

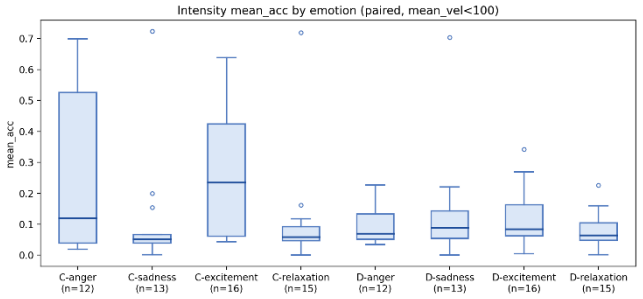

図 3　手法・感情条件別の平均加速度分布

また，左右の人物間の動作類似性の評価には Dynamic Time Warping（DTW）距離を用いた［5］．DTW 距離が小さいほど，両者の動作系列が類似していることを示す．姿勢推定失敗による影響を抑えるため，算出した平均速度指標の値が 100 以上となった動画を評価対象から除外した．

4. 結果および考察

図 2，図 3 に手法・感情条件別の平均速度および平均加速度の分布を示す．図中の横軸ラベルでは，c は提案手法（combine），d は比較手法（default）を表す．提案手法では，anger および excitement が比較的大きく，sadness および relaxation が小さくなる傾向が見られた．人物ごとに与えた感情が，動作の速さや激しさの違いとして反映された可能性がある．

一方，比較手法では感情間の差が提案手法より小さい傾向が見られた．画像全体を入力した場合，複数人物が一体の対象として解釈され，感情や動作が人物ごとに分離されずに反映されたと考えられる．

DTW 距離では，全条件を通じた明確な優劣は確認できなかった．ただし，左右の人物の全身が画像内に収まる条件では，比較手法の DTW 距離が小さく，人物間の動作が類似する傾向が見られた．提案手法では人物ごとに独立して生成するため，一部の条件において動作系列の類似性が低下した可能性がある．

以上より，人物を分離して個別に動画を生成することで，感情に応じた動作強度の差を表現しやすくなることが示唆された．ただし，速度や加速度の差だけでは，動作が意図した感情として認識されるかを直接評価できない．また，DTW 距離には姿勢推定誤差や生成の不安定性も影響するため，感情表現の正確性には追加評価が必要である．

5. まとめ

本研究では，複数人物を含む画像を人物ごとに分離し，異なる感情プロンプトで動画を生成した後，背景上に再合成するパイプラインを提案した．

実験の結果，画像全体を直接入力する場合と比べ，提案手法では感情条件による平均速度および平均加速度の差が明確になる傾向が確認された．また，画像全体を入力した場合，人物間の動作が類似しやすい条件が見られた．以上より，人物単位の動画生成は，複数人物の動作を個別に制御する上で有効である可能性が示された．

今後は，画像数や人物数を増やすとともに，人物の重なりや欠損を含む画像での性能を検証する．また，感情分類モデルや主観評価によって感情表現の正確性を評価し，人物領域の境界に生じる合成の不自然さを改善する．

文　献

[1] L. Zhang, S. Cai, M. Li, G. Wetzstein, and M. Agrawala : “Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models”, arXiv:2504.12626 (Apr. 2025)

[2] Y. Chen, S. Liang, Z. Zhou, Z. Huang, Y. Ma, J. Tang, Q. Lin, Y. Zhou, and Q. Lu : “HunyuanVideo-Avatar: High-Fidelity Audio-Driven Human Animation for Multiple Characters”, arXiv:2505.20156 (May 2025)

[3] J. Lin, C. Li, X. Liu, and Z. Ge : “Instance-Guided Anime Editing with a Curated Large-Scale Dataset”, The Visual Computer, 41, 9, pp.6715-6727 (July 2025)

[4] Y. Xu, J. Zhang, Q. Zhang, and D. Tao : “ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation”, Advances in Neural Information Processing Systems, 35, pp.38571-38584 (Nov.-Dec. 2022)

[5] M. Müller : “Dynamic Time Warping”, Information Retrieval for Music and Motion, Chap. 4, pp.69-84, Springer, Berlin, Heidelberg (2007)

†早稲田大学　基幹理工学研究科
情報理工・情報通信専攻
〒169-0072　東京都新宿区大久保 3-14-9
TEL.070-8471-3380　E-mail: mao.s@moegi.waseda.jp