2CM-GPT による関節の空間情報を考慮した 2 次元モーション生成に関する研究

TP23003 井上涼太　　指導教授：山下隆義

1. はじめに

テキストからのモーション生成は，人の動作を記述したテキストをプロンプトとして入力し，その動作を表す関節の座標，速度及び回転などのモーション情報を指定したフレーム数分生成するタスクである．大規模言語モデルや拡散モデルの進歩に伴い，3 次元モーション生成手法は数多く提案されている [1]．3 次元モーションは奥行きを含む空間的な情報を持つため，人の動作を詳細に表現できる一方，データセット作成には専門的な設備や技術が必要であり，高コストである．また，人のモーションを用いた動画生成では，3 次元モーションよりも必要な計算量が少なく，データセットの作成が容易な 2 次元モーションを入力として必要とするケースが多い．そこで本研究では，テキストから 2 次元モーションを生成する 2-dimensional Convolutional Motion Generative Pre-trained Transformer (2CM-GPT) を提案する．2 次元モーションは関節間の相対的な配置や空間的な構造が動きの意味を大きく左右するため，2CM-GPT では空間的な関係性を考慮するために 2 次元畳み込みを採用する．また，既存のテキストと 3 次元モーションのペアデータセットである HumanML3D [2] を基に，2 次元モーションのデータセットを新たに作成し，提案手法の学習を行う．

2. 関連研究

本章では，従来のテキストからの 3 次元モーション生成手法である MotoinGPT，テキストと 3 次元モーションのペアで構成された HumanML3D データセットについて述べる．

2.1. MotionGPT

人のモーションには言語と同様の意味的構造を持つと仮定し，MotionGPT [1] はモーションを言語として統一的に扱う手法である．MotionGPT はテキストトークナイザ，モーショントークナイザ及び言語モデルから構成される．モーショントークナイザには Vector Quantised Variational AutoEncoder (VQ-VAE) [3] を採用し，モーションを離散的なモーショントークンへ変換する．この変換により，モーション語彙が構築され，言語処理と同様の方法でモーションを扱うことができる．言語モデルには Text-To-Text Transfer Transformer (T5) [4] を採用し，モーショントークンとテキストトークンを組み合わせた語彙を用いて事前学習を行い，各タスクに対応した質問形式のテキストテンプレートを用いてインストラクションチューニングを行う．

2.2. HumanML3D

HumanML3D は，HumanAct12 と Archive of Motion Capture As Surface Shapes のモーションを統合したデータセットであり，14,616 のモーションと，5,371 の異なる単語を含む 44,970 のテキストから構成される．このデータセットは，日常的な活動（歩く，跳ぶなど），スポーツ（泳ぐ，空手など），アクロバット（回し蹴りなど）及び芸術（踊るなど）の様々な動作の 3 次元モーションを含む．またテキストは，少なくとも 5 つの単語からなる 3 つの文章がペアとして，そのモーションを説明する．

3. データセットの作成

テキストと 3 次元モーションのペアで構成されたデータセットは存在するが，テキストと 2 次元モーションがペアのデータセットは存在しない．また，2 次元モーションを用いた動画生成では，Common Objects in Context (COCO) データセットに基づく 18 関節のモーションが広く利用されている．そこで本研究では，HumanML3D から COCO 形式の 2 次元モーションを作成する．具体的には，まず HumanML3D の 3 次元モーションに対して，Skinned Multi-Person Linear model (SMPL) [5] モデルを用いて SMPL メッシュを推定する．次に，SMPL メッシュに回帰行列を乗算し，2 次元平面上で 18 個の関節座標を取得することで，2 次元モーションを作成する．ここで，作成した 2 次元モーションデータセットは，HumanML3D のテキスト，モーションの種類及びアクションの長さに基づいて構築している．

4. 提案手法

提案手法である 2-dimensional Convolutional Motion Generative Pre-trained Transformer (2CM-GPT) は，テキストを入力として直感的かつ効率的に 2 次元モーションを生成することが可能であり，生成したモーションは 2 次元モーションを用いた動画生成へ直接適用できる．2CM-GPT の構造を図 1 に示す．2CM-GPT は，モーショントークナイザと言語モデルから構成される．

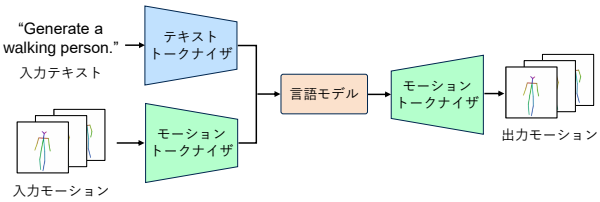


図 1：2CM-GPT のモデル構造

4.1. モーショントークナイザ

従来の 3 次元モーション生成手法は，1 次元畳み込みを用いたモーショントークナイザが主流である．3 次元モーションはフレーム F ごとに 3 次元モーション M を表す関節の位置，速度及び回転などを単一の次元で表現する時系列データ $F \times M$ として構成される．従って，1 次元畳み込みを用いることで，フレーム間の時間的な連続性や変化を効果的に捉えることができる．しかし，2 次元モーションはフレーム F ごとに各関節 J の位置が 2 次元平面上の座標 P として，2 つの異なる次元で表現する時系列データ $F \times J \times P$ として構成される．そのため，モーションの空間的な関係性を捉えるためには，2 次元畳み込みを用いることが効果的である．2CM-GPT では，モーショントークナイザに 2 次元畳み込みを採用することで，各関節の空間的な関係性を効果的に捉える．

4.2. 言語モデル

2CM-GPT の言語モデルには事前学習済みの T5 モデルを使用して，テキストとモーション間の関係性を学習する．入力には，モーションを記述したテキストトークンと，2 次元モーションを表現するモーショントークンを組み合わせたテキストモーション語彙を用いる．テキストモーション語彙は，テキストトークン，モーショントークン，そしてテキストとモーションの両方を表したトークンを表現できるため，テキストとモーションの関係性を効率的に学習できる．

4.3. 学習方法

2CM-GPT は 3 段階の学習戦略を採用する．第 1 段階では 2 次元モーションをモーショントークナイザで学習し，モーショントークンを獲得する．第 2 段階ではテキストのみで事前学習された言語モデルを用いて，テキストとモーションの関係性を学習する．第 3 段階ではテキストからのモーション生成用のテキストテンプレートを用いてファインチューニングする．MotionGPT では，複数タスクのテンプレートを用いているが，2CM-GPT ではより高精度なテキストからの 2 次元モーション生成を行うために，テキストからのモーション生成用のテンプレートのみを用いてファインチューニングする．

5. 評価実験

本実験では，1 次元畳み込みを用いた MotionGPT と 2 次元畳み込みを用いた提案手法の再構成精度及びテキストからの 2 次元モーション生成精度を比較することで，提案手法の有効性を示す．また，生成モーションをポーズ誘導による人物動画生成手法である DisCo へ適用することで，

表 1：再構成精度の比較

Method	MPJPE ↓	FID ↓		Diversity ↑	
		FE_M	FE_2	FE_M	FE_2
Ground Truth	0.00	0.00	0.00	16.98	16.95
MotionGPT	0.17	33.75	34.54	11.38	10.89
2CM-GPT	**0.16**	**25.85**	**25.15**	**12.24**	**12.12**

表 2：テキストからの 2 次元モーション生成精度の比較

Method	FID ↓		Diversity ↑	
	FE_M	FE_2	FE_M	FE_2
Ground Truth	0.00	0.00	16.66	16.71
MotionGPT	27.10	27.13	17.59	18.37
2CM-GPT	**26.35**	**27.07**	**18.36**	**19.15**

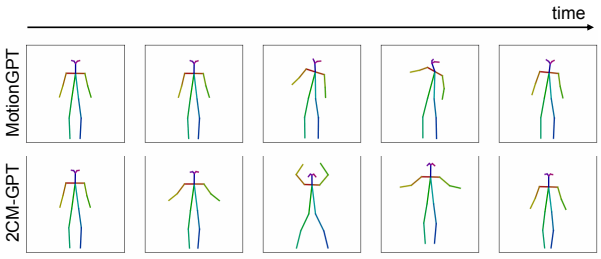


図 2：テキストからの 2 次元モーション生成結果の可視化

提案手法の応用性を示す．評価指標は，Mean Per Joint Position Error (MPJPE) ，Fréchet inception distance (FID) 及び Diversity を用いる．MPJPE は，正解と出力の各関節間の L2 ノルムの平均値であり，0 に近いほど精度が高いことを表す．FID は，正解と出力モーションの特徴ベクトル分布間の Fréchet 距離であり，0 に近いほど精度が高いことを表す．Diversity は，ランダムに選択した出力モーションのペアの特徴ベクトル分布の分散であり，値が高いほど多様性が高いことを表す．

FID と Diversity の評価には，通常，基準となる特徴抽出器により抽出した特徴ベクトルを使用する．しかし，作成した 2 次元モーションデータセットに適した特徴抽出器は存在しない．そのため本実験では，従来手法及び提案手法のモーショントークナイザを特徴抽出器として採用することで，それぞれの手法が持つ本来の性能を公平に比較する．

5.1. モーショントークナイザの比較

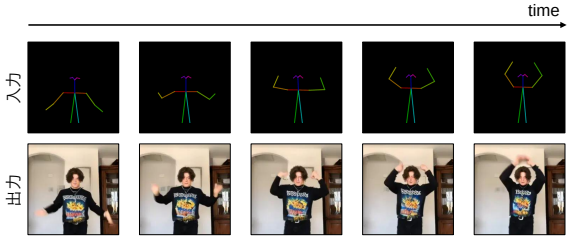
表 1 に MotonGPT と提案手法の再構成精度を示す．ここで，FE_M と FE_2 は基準となる特徴抽出器として用いた MotionGPT と 2CM-GPT のモーショントークナイザを示す．表 1 より，2 次元畳み込みを採用している提案手法の方が高精度であることが確認できる．これにより，2 次元モーションデータを 2 次元畳み込みで処理することは 1 次元畳み込み処理よりも有効であることが分かった．

5.2. テキストからの 2 次元モーション生成

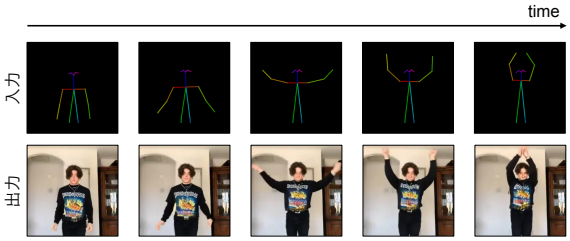
表 2 にテキストからの 2 次元モーション生成精度を，図 2 に生成モーションの可視化を示す．表 2 より，提案手法が MotionGPT に比べて 2 次元モーション生成精度と多様性が向上していることが確認できる．図 2 より，提案手法は入力テキストの動作内容に沿った 2 次元モーションを生成していることが確認できる．特に動作の開始点と終了点が正確に反映されており，動作の一貫性が保たれている．提案手法はテキストの意味的情報を適切に解釈し，それを 2 次元モーションに変換できていることから，現実的で精度の高いモーションを生成できると考えられる．

5.3. ポーズ誘導による人物動画生成への適用

ここでは，テキストから生成したモーションをポーズ誘導による人物動画生成手法である DisCo の入力に使用し，生成動画を可視化することで提案手法の応用性を確認する．



(a) MotionGPT



(b) 2CM-GPT

図 3：ポーズ誘導による人物動画生成結果

DisCo は，人物が含まれる画像と 2 次元モーションを入力として，画像内の人物が 2 次元モーションに従って動作する動画を生成する手法である．図 3 に MotionGPT が生成した 3 次元モーションを 2 次元モーションへ変換し入力した場合と，提案手法が生成した 2 次元モーションを DisCo に入力して，生成した動画を示す．ここで，モーションを生成する際に使用した入力テキストは“A person claps their hands together well above their head.”である．図 3 より，提案手法が自然な動作を生成していることが確認できる．これにより，提案手法では生成モーションの変換を行わずに，従来の 3 次元モーション生成手法と同等の動画生成を行えることが分かる．

6. おわりに

本研究では，テキストから 2 次元モーションを生成する 2CM-GPT を提案した．評価実験より，提案手法のモーショントークナイザは MotionGPT よりも再構成精度が向上することを確認した．また，テキストに従った動作の 2 次元モーションも生成できることを確認した．今後は，個人の体型に合わせたモーション生成を検討する．

参考文献

[1] B. Jiang, *et al.*,“MotionGPT: Human Motion as a Foreign Language”, NeurIPS, 2023.

[2] C. Guo, *et al.*,“Generating Diverse and Natural 3D Human Motions from Text”, CVPR, 2022.

[3] A. Oord, *et al.*,“Neural Discrete Representation Learning”, NeurIPS, 2017.

[4] C. Raffel, *et al.*,“Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, JMLR, 2020.

[5] M. Loper, *et al.*,“SMPL: A Skinned Multi-Person Linear Model”, SIGGRAPH Asia, 2020.

研究業績

[1] R. Inoue, *et al.*,“2D Motion Generation Using Joint Spatial Information with 2CM-GPT”, VISAPP, 2025.
（他 2 件）