第 28 回日本バーチャルリアリティ学会大会論文集（2023 年 9 月）

イベントデータのみを用いた三次元人物姿勢および形状推定

3D Human Pose and Shape Estimation Using Only Event Data

堀涼介 [1)]，五十川麻理子 [1,3)]，三上弾 [2)]，斎藤英雄 [1)]
Ryosuke HORI, Mariko ISOGAWA, Dan MIKAMI and Hideo SAITO

1) 慶應義塾大学 理工学研究科（〒 223-8522 神奈川県横浜市港北区日吉 3-14-1）
2) 工学院大学 情報学部（〒 163-8677 東京都新宿区西新宿 1-24-2）
3) JST さきがけ

概要: 我々は，悪照明環境に頑健で，かつ省電力・省メモリな人物メッシュ復元手法の実現を目指し，イベントカメラで取得したイベントデータのみを入力とした人物メッシュ復元という新規タスクに取り組む．本稿では，イベントデータを時空間の三次元点群として扱い，人物メッシュを復元するフレームワークである Event Point Mesh を提案する．既存のデータセットを用いてベースライン手法との定量および定性評価実験を行い，提案手法の有効性を確認した．

キーワード： 人物姿勢および形状推定, 人物メッシュ復元, イベントカメラ

1. はじめに

シーン中の人物の姿勢を推定する手法には，リハビリテーション支援や人物の見守り，災害時の救助活動やスポーツ解析など，様々な応用が期待できる．中でも，関節位置のみならず，推定対象の人物の体型も含めた姿勢推定を可能にする技術として，人物メッシュ復元に関する手法が近年提案されており，これは人物の状態をより詳細に解析する上で非常に重要な技術である．従来，RGB(D) カメラ等に代表される可視光信号に基づく手法 [1] や，Radio Frequency/WiFi などの無線信号を利用した手法 [2] が提案されている．しかし，可視光ベースの手法は暗室環境や夜道などの照明条件の悪い環境では良好に機能させることが難しい．また，計測機器に多くの電力や計測メモリ量を要するため，エッジデバイスなどへの搭載に課題がある．一方，無線信号ベースの手法は，航空機内や病室等の電子精密機器を扱う現場では利用を制限されることが多い．

これらの課題を解決し得る計測方法として，イベントカメラを活用する方法が知られている．イベントカメラは撮影対象の輝度の変化を検出し，座標，極性（輝度変化方向），時間のみを含むイベントデータを出力する．従来のカメラが一定のフレームレートで撮像記録を行うのに対し，画素毎に独立して輝度変化を検出し，明暗の変化が一定閾値以上となった画素のみの記録を非同期に行うため，その高時間分解能や低消費電力性が期待できる．また，輝度変化のみを感知することから，撮像可能なダイナミックレンジが広いことでも知られている．つまりイベントカメラを活用することで，無線信号ベースの手法のように活用シーンを制限されることがなく，かつ RGB(D) カメラよりも暗所に強く，また，省電力・省メモリな手法を実現可能なのである．

そこで我々は，イベントカメラを活用した人物メッシュ復元手法に着目する．従来，イベント情報を活用した人物メッシュ復元手法がいくつか提案されている [3, 4]．しかし，これらの手法では，イベント情報と同時に撮影されたグレースケールのフレーム画像を必要とする．これでは，本来イベントベース手法に期待していた暗所への耐性や，省電力・省メモリ性が大きく制限されてしまう．

そこで本稿では，**イベント情報のみを用いた人物メッシュ復元**，という新規タスクに取り組む．これにより，照明環境の変化や悪照明環境に頑健で，かつ省電力・省メモリな人物メッシュ復元の実現を目指す．その目的のために，イベントデータを時空間の三次元点群として扱うフレームワークを提案し，イベントカメラの特徴である時空間のスパース性を活かした人物メッシュ復元手法を実現する．

本稿の貢献を以下にまとめる．(1) イベントデータのみを入力とした人物メッシュ復元手法という新規タスクに取り組んだ．(2) イベントデータを時空間の三次元点群として扱い，人物メッシュを対応づけるフレームワークを提案した．(3) 手法の有効性を調査するため，ベースライン手法との定量および定性評価実験を行った．

2. 提案手法

本稿では，イベントカメラの時系列の点群のみから人物メッシュを推定する深層学習モデルである，EventPointMesh を提案する（図 1）．EventPointMesh は，イベント点群から特徴を抽出する Base Module，点群から大域的な特徴ベクトルと人物の二次元関節位置を推定する Keypoint Module，推定された各関節の周辺の点群をグルーピングして特徴を抽出する Anchor Points Module，点群から得られた特徴ベ

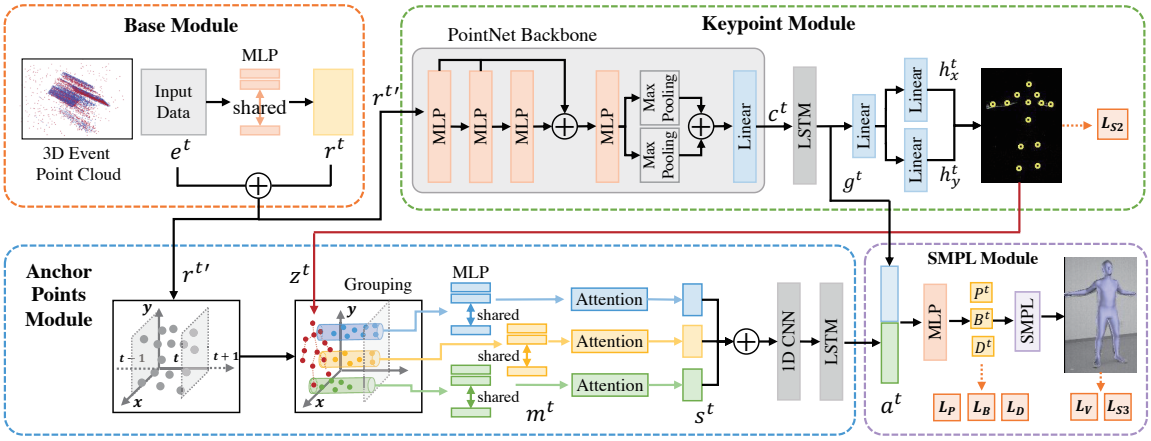

図 1: 提案手法である EventPointMesh のフレームワーク.

クトルから SMPL モデルを推定する SMPL Module から構成されている．各モジュールの詳細を以下に述べる．

2.1 モジュールの概要

Base Module. イベントデータ $\boldsymbol{e}$ を一定時間毎に区切った三次元点群データを入力とし，Multi-Layer Perceptron(MLP) によって高次元の特徴に変換する．具体的には，一定間隔で区切られた時間窓のうち，t 番目の時間窓に記録されたすべてのイベントデータを e^t とし，その時間内に検出された内の i 番目のデータを e_i^t とすると，パラメータ θ_r から成る MLP の変換により高次元の特徴ベクトル $r_i^t = \mathrm{MLP}(e_i^t;\theta_r)$ が得られる．その特徴ベクトルに e_i^t の座標と時間の三次元ベクトルを連結させた，高次元ベクトル $r_i^{t\prime}$ を後続のモジュールの入力とする．

Keypoint Module. イベントデータから人物姿勢を推定するためには，輝度変化の時空間情報のみを持つ点群データから，各点がどの身体部位の動作によって発生したものであるかという情報を得ることが重要である．そこで，Keypoint Module では，後続のモジュールで各体節によって発生した点群の局所的な特徴を得るための基準点（以降，anchor points と記載）として利用するために，入力の点群から主要な関節点の二次元座標を推定する．まず，Base Module によって得られた各点の特徴ベクトル r_i^t から，Point Net [5] をバックボーンとしたネットワークにより点群の大域的な特徴 $c_i^t = \mathrm{PointNet}(r_i^t)$ に変換される．その後，Bidirectional LSTM(BiLSTM) によって時系列の関係性を考慮した特徴ベクトルを c^t から抽出する．これは，動いている身体部位のみを捕捉して，静止している部分を無視するというイベントカメラの特性により，体の一部の情報が少ない疎なデータとなり姿勢推定が難しくなるという問題に対し，前後の時間窓の点群特徴を活かして疎な部分を補うことで対処するためのものである．BiLSTM のパラメータを θ_g, 入力フレーム数を T とすると，特徴ベクトルは $g^t = \mathrm{BiLSTM}(c^{1:T};\theta_g)$ となる．その後，時系列特徴ベクトル g^t は，パラメータ θ_s を持つ分岐した MLP により関節座標を推定するための二つの一次元ヒートマップベクトル $(h_x^t, h_y^t) = \mathrm{MLP}(g^t;\theta_s)$ に変換される．これは二次元のヒートマップによって関節位置推定を行っていた従来手法に対し，低解像度画像での高精度な推定を可能にする新たな座標表現である SimDR/SimCC [6] を点群データに適用したものである．ヒートマップベクトル表現の詳細については，元論文を参照されたい．

Anchor Points Module. 点群の詳細な特徴を捉えて高精度な推定を行うために，二次元関節座標を利用して時空間上に広がるイベント点群を各体節に起因するイベントとしてグルーピングすることで，それらの局所的特徴を抽出する．このネットワーク構成は，Xue ら [2] が提案したミリ波の点群から人物の三次元メッシュを推定する mmMesh から着想を得たもので，Xue らは固定サイズの三次元格子状の anchor points を利用していたのに対し，我々は二次元関節点の anchor points を時空間上でのグルーピングに利用する．図 1 に示すように，anchor points と元のイベント点群との二次元画像座標上でのユークリッド距離に基づき，三次元の時空間内で最近傍点から順に N 個ずつグルーピングすることで，各 anchor point 周辺の体節の動作に起因して計測された可能性の高い点群を分類する．ここで，時間窓 t における k 番目の anchor points の座標を z_k^t とし，グルーピングされた近傍の N 点のインデックスを $NSP(z_k^t)$ とする．また，グルーピング前のイベント点群の三次元座標を $COOR(e_i^t)$ とすると，これらの点群のうちの $i \in NSP(z_k^t)$ の点 e_i^t，すなわちグルーピングされたイベント点は, MLP によって特徴ベクトル $m_i^t = \mathrm{MLP}([z_k^t; COOR(r_i^t) - z_k^t; r_i^t];\theta_m)$ に変換される．つまり，m_i^t は，anchor points の座標，近傍のグルーピングされた点と anchor points の距離，Base Module で得られた各点の特徴ベクトルが連結された特徴ベクトルとなる．この設計により，anchor points とその近傍点の時空間上で位置関係の特徴を捉えた学習が可能となる．特徴ベクトル m_i^t はその後，グルーピングされた点群の中でも特に重要な特徴を抽出するために，アテンションネットワークに入力される．アテンションネットワークを θ_s のパラメータを持つ線型写像関数 L と表現すると，特徴ベクトル m_i^t は次のような特徴ベクトルに変換される：

$$s_k^t = \sum_{i \in NSP(z_k^t)} L(m_i^t;\theta_s) \cdot m_i^t. \tag{1}$$

さらに，anchor point ごとの特徴 s_k^t を集約して，各グループ間，すなわち各体節によって発生したイベント点群間での関連性を考慮した特徴に変換するため，一次元畳み込み層を導入し，特徴量ベクトル $d^t = 1DCNN(s^t; \theta_d)$ を得る．この d^t を，Keypoint Module と同様に BiLSTM を用いて時系列の関係性を考慮した特徴ベクトルに変換する．BiLSTM のパラメータを θ_a とすると，最終的に得られる特徴ベクトルは $a^t = \text{BiLSTM}(a^{t-1}, d^t; \theta_a)$ となる．

SMPL Module では，まず Keypoint Module および Anchor Points Module で得られたベクトル g^t と a^t を連結させた特徴ベクトルを，MLP を用いて $[P^t, B^t, D^t] = \text{MLP}([g^t; a^t]; \theta_p)$ のような SMPL(Skinned Multi-Person Linear Model) [7] パラメータに変換する．SMPL モデルパラメータの詳細については supplemental material を参照されたい．ここで，P^t は姿勢，B^t は形状，D^t は並進運動を表すベクトルで，θ_p は MLP のパラメータである．次に，P^t, B^t, D^t から，学習済みの SMPL モデルによってメッシュの頂点 V と S を $[V^t, S^t] = \text{SMPL}([P^t, B^t]) + D^t$ のように推定する．ただし本稿では，学習済みの SMPL モデルのネットワークの重みを固定して使用した．

2.2 損失関数

本稿でネットワーク学習に用いる損失関数を以下に示す:

$$\mathcal{L} = \sum_{K \in \{P,B,D,V,S_3\}} \alpha_K \left(\frac{1}{T}\sum_{t=1}^{T} \| K^t - \mathcal{GT}(K^t) \|_{L_2}\right) + \alpha_{S_2}\left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{KL}(S_2^t \parallel \mathcal{GT}(S_2^t))\right). \tag{2}$$

P, B, D はそれぞれ SMPL モデルに入力する姿勢，形状，並進を表すテンソル，V, S_3 は SMPL モデルによって推定されたメッシュの頂点の三次元座標と各関節の三次元座標を表すテンソルである．また，S_2 は推定された三次元の関節点をカメラ座標系の画像平面に投影した二次元座標のテンソルを表している．ただし，本手法においては，イベントカメラでの撮影時のカメラパラメータは既知とする．また，$\mathcal{GT}(K)$ は推定されたテンソル K に対応する正解ラベルを，α_K 及び α_{S_2} は各損失のバランスを調整するためのハイパーパラメータを表している．

P, B, D, V, S_3 はそれぞれ正解ラベルとの L_2 ノルムで算出し，S_2 は正解ラベルとのヒートマップの分布差を計算する KL ダイバージェンス $\mathcal{L}_{KL}$ を用いる．ここで，S_2 の正解ラベルとして扱うヒートマップは，正解データの三次元の関節点を縦横 (H, W) の画像に投影した二次元関節点 (x', y') に対し，以下のようなガウシアンフィルタによってヒートマップ h_v に変換する．

$$\begin{cases} \boldsymbol{h_v} = [v_0, v_1, v_2, ..., v_I] \in \mathbb{R}^I, \\ v_i = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-\dfrac{(i - v')^2}{2\sigma^2}\right). \end{cases} \tag{3}$$

上式の記号は $v \in \{x, y\}, I|_x = W, I|_y = H$ を表している．

表 1: 提案手法およびベースライン手法との定量評価結果．

	Method	MPJPE [mm] (↓)	PA-MPJPE [mm] (↓)	PEL-MPJPE [mm] (↓)
実験 1	EventHPE [4]	76.4	41.1	51.4
	Ours (1frame)	165.1	76.2	112.5
	Ours (5frame)	126.3	52.0	83.8
	Ours (15frame)	120.0	49.2	78.9
実験 2	EventHPE [4]	92.2	51.6	66.2
	Ours (1frame)	209.5	86.0	129.4
	Ours (5frame)	182.4	79.1	119.8
	Ours (15frame)	180.0	68.2	103.8

3. 実験および結果

3.1 実験設定

データセット． 既存の大規模データセットである EventHPE [4] を用いた．被験者は男性 11 名，女性 4 名の計 15 名で，それぞれ速い／中程度／遅い速度の動作が含まれる 3 つのグループの動作（計 21 種類）が 4 回ずつ含まれている．それぞれ 12 個のビデオクリップ（合計 180 本）が収集されており，各ビデオは 15FPS で約 1.5 分間の約 1,300 フレームから構成される．

実装の詳細． ネットワークの学習には Adam Optimizer [8] を使用し，損失関数の各要素の重み α_P, α_B, α_D, α_V, α_{S_3}, α_{S_2} は，それぞれ 10, 10, 10, 100, 10, 1 とした．また，モデルの学習を安定化させてメッシュ推定精度を向上させるために，事前に Base Module と KeypointModule を独立して学習させ，推定された 2 次元関節位置を Anchor Points として利用しネットワーク全体を学習させた．

比較手法． 本手法は，イベントデータの点群のみから人物三次元姿勢と形状を推定する初めての試みであり，同条件で精度比較可能な手法が存在しない．そこで，輝度画像とイベントフレームを組み合わせて姿勢と形状を推定する手法である EventHPE [4] をベースラインとした．また，イベントデータのシーケンスの長さが EventPointMesh の推定精度に与える影響を調べるために，入力画像のシーケンスを 1, 5, 15 フレーム毎に区切った 3 種類の入力方法をとる提案手法である Ours (1frame)，Ours (5frame), Ours (15frame) についても，推定精度比較を行なった．

評価指標． 人物姿勢推定タスクの評価に広く用いられている，Mean-Per Joint Position (MPJPE)，Procrustes-Aligned MPJPE (PA-MPJPE)，Pelvis-Aligned MPJPE (PEL-MPJPE) の 3 つの指標に基づき精度評価を行なった．これらは，推定された姿勢と真値の各関節の位置関係，姿勢の回転・並進状態を一致させた上での各関節の位置関係，また姿勢の並進状態のみを一致させた上での各関節の位置関係の三次元ユークリッド距離を評価する指標である．

3.2 実験 1．異なる撮影データに対する汎化性能の調査

異なるタイミングで撮影されたイベントデータに対するモデルの汎化性能を評価するため，各被験者が動作を 4 回

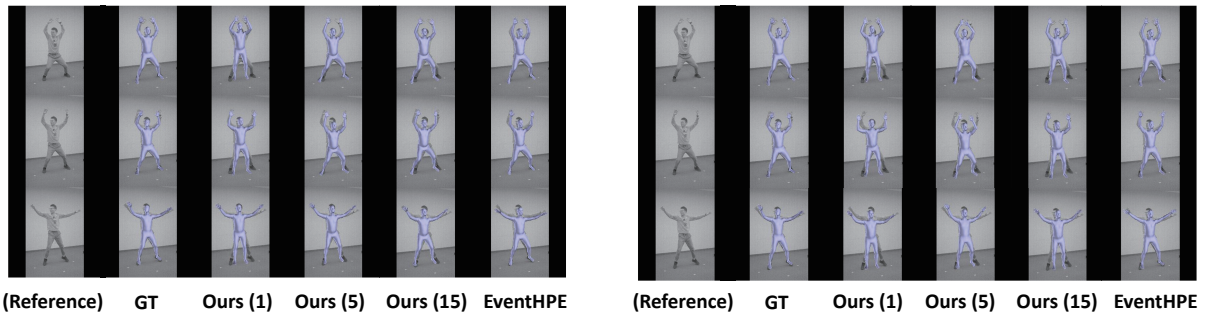

図 2: 実験 1 の定性評価結果 (左) と実験 2 の定性評価結果（右）.

ずつ行ったビデオシーケンスのうち，3 回分をネットワークの学習に使い，残りの 1 回分をテストに使用した．

定性評価結果を図 2 左に，定量評価結果を表 1 上部に示す．なお，図 2 の輝度画像は参照用に提示したものである．表 1 の結果から，提案手法は入力フレーム数が増えるほど精度が向上し，輝度画像も利用する EventHPE に対しても数 cm 程度誤差での推定が実現できたことが分かる．図 2 で示す結果に着目すると，Ours (1frame) や Ours (5frame) は，全身では概ね正しい姿勢が推定されているものの，肘から先の姿勢に推定誤差があることがわかる．一方，Ours (15frame) は，輝度画像を入力として使用する EventHPE や正解ラベルと同等の精度でメッシュを推定できている．これは，身体の動作が小さく，発生したイベントが少ない疎な部分があるデータに対し，BiLSTM によって長い画像シーケンスの時系列特徴を抽出したことにより，情報が補間されたためだと考えられる．

3.3 実験 2. 異なる被験者に対する汎化性能の調査

異なる被験者間での汎化性能を調査する目的で，計 15 名の被験者のうち，12 名分のデータをネットワークの学習に，3 名分を推論に用いる設定での推定精度比較実験を行った．実験の定性的結果を図 2 右に，定量的結果を表 1 下部に示す．提案手法の入力フレーム数と精度の関係や EventHPE との誤差の度合いは実験 1 と同様の傾向となったものの，どの手法や設定においても，推定精度は実験 1 の精度と比較して低く，データセットに含まれない人物の推論では推定が悪化することが示唆された．これは，同じ動作であっても被験者による違いが生じることや，服装や体格の差が影響したためであると考えられる．また，表 1 の結果から，いずれの実験においても，提案手法は EventHPE と比較して MPJPE とその他の評価指標との差が大きいことが分かる．これは，提案手法では推定されたメッシュの回転ベクトルおよび並進ベクトルの誤差が大きいことを意味している．今後の研究ではその改善策についても検討する．

4. 結論

本稿では，時系列のイベント点群情報のみから三次元の人物姿勢および人物の形状を推定するという新規タスクに取り組むとともに，それを実現するための手法 EventPointMesh を提案した．従来手法とは異なり，提案手法は輝度画像情報を一切活用しないため，推定精度は従来手法に劣っていたが，いずれの実験設定においても数 cm 程度の推定精度差であった．今回は主に照明条件の良好な条件下で実験を行ったが，暗所ではベースライン手法よりも良好な精度を記録する可能性があると考えている．

謝辞 本研究の一部は，JST さきがけ JPMJPR22C1，慶應義塾学事振興資金，JST 次世代研究者挑戦的研究プログラム JPMJSP2123，および科学研究費基盤研究 (B)23H03422 の助成を受けたものである．

参考文献

[1] G. Moon and K. M. Lee: I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image, ECCV, p. 752–768 (2020).

[2] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang and L. Su: MmMesh: Towards 3D Real-Time Dynamic Human Mesh Construction Using Millimeter-Wave, Mo- biSys, p. 269–282 (2021).

[3] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang and C. Theobalt: EventCap: Monocular 3D Capture of High-Speed Human Motions Using an Event Camera, IEEE/CVF CVPR, pp. 4968–4978 (2020).

[4] S. Zou, C. Guo, X. Zuo, S. Wang, H. Xiaoqin, S. Chen, M. Gong and L. Cheng: EventHPE: Event-based 3D Human Pose and Shape Estimation, IEEE/CVF ICCV, pp. 4710–4720 (2021).

[5] C. R. Qi, H. Su, K. Mo and L. J. Guibas: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, IEEE/CVF CVPR, pp. 652–660 (2017).

[6] Y. Li, S. Yang, S. Zhang, Z. Wang, W. Yang, S. T. Xia and E. Zhou: Is 2D Heatmap Representation Even Necessary for Human Pose Estimation?, ECCV, p. to appear (2022).

[7] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll and M. J. Black: SMPL: A Skinned Multi-Person Lin- ear Model, ACM TOG, Vol. 34, No. 6 (2015).

[8] D. P. Kingma and J. Ba: Adam: A Method for Stochas- tic Optimization, ICLR (2015).