MMAM-Cap: A Multimodal Accurate Motion Capture System for Human Pose and Trajectory Estimation

Yuxiang Liu
Beijing University of Posts and Telecommunications
yuxiangliu@bupt.edu.cn

Anlong Ming*
Beijing University of Posts and Telecommunications
mal@bupt.edu.cn

Ruizhe Kang
Beijing University of Posts and Telecommunications
byrkrz@bupt.edu.cn

Yonglong Wang
Beijing University of Posts and Telecommunications
dnqf@bupt.edu.cn

Weihong Yao
Shanghai Vision Era Co., Ltd
yaoweihong@xvgate.com

Huadong Ma
Beijing University of Posts and Telecommunications
mhd@bupt.edu.cn

ABSTRACT

This paper introduces a Multimodal Accurate Mocap (MMAM-Cap) system, designed to integrate monocular video and Inertial Measurement Unit (IMU) sensor data for high-precision human pose and motion trajectory prediction. The MMAM-Cap system is composed of three main modules: 1) the **Two-Stream Motion Estimation Module**, which extracts human motion features from both the camera and IMU coordinate systems; 2) the **Co-Evolution Decoder**, which fuses multimodal features to reconstruct human poses in the camera coordinate system; and 3) the **Global Trajectory Refinement Module**, which leverages foot-ground contact information to refine the trajectory in the IMU coordinate system, improving global consistency. We propose an Adaptive Fusion Module (AFM) based on cross-attention to dynamically fuse the multimodal data, effectively addressing the complementary strengths and weaknesses of each modality under different scenarios. Extensive experiments demonstrate the superiority of the MMAM-Cap system across multiple datasets, where it significantly improves both pose estimation accuracy and motion trajectory smoothness. Our system not only mitigates cumulative errors from using a single data source but also reduces drift from inertial data through global refinement. This work provides a robust solution for multimodal motion capture and establishes a solid foundation for capturing human motion in complex real-world environments.

ACM Reference Format:
Yuxiang Liu, Anlong Ming, Ruizhe Kang, Yonglong Wang, Weihong Yao, and Huadong Ma. 2024. MMAM-Cap: A Multimodal Accurate Motion Capture System for Human Pose and Trajectory Estimation. In *Proceedings of The 5th International Workshop on Human-centric Multimedia Analysis (MM'24)Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author

MM'24, October 28 - November 1, 2024, Melbourne, Australia.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the rapid development of technologies such as virtual reality (VR), augmented reality (AR), and human-computer interaction, accurate and real-time human motion capture has become a critical requirement in many applications. Traditional motion capture systems typically rely on expensive and complex multi-camera setups or marker-based devices, which are often impractical for everyday environments due to their high cost and stringent operational requirements. In recent years, motion capture techniques that combine monocular video with IMU sensor data have gained attention as a more affordable and flexible alternative. These methods explore the strengths of both visual information and inertial sensor data, offering a more accessible way to achieve reasonably accurate human pose estimation.

However, fusing monocular video and IMU sensor data presents significant challenges. Monocular video lacks depth information, and its performance can degrade in the presence of occlusions, rapid movements, or poor lighting conditions. On the other hand, while IMU sensors can provide localized motion data, they are susceptible to drift over time, which leads to accumulating errors. Addressing these challenges requires an efficient multimodal fusion approach that maximizes the strengths of both data sources while compensating for their individual shortcomings.

In this paper, we introduce the **Multimodal Accurate Mocap (MMAM-Cap)** system, which aims to tackle these challenges by dynamically fusing visual and inertial data to achieve accurate and stable human pose and motion trajectory estimation. The MMAM-Cap system consists of three key components: 1) the **Two-Stream Motion Estimation Module**, which extracts motion features from both the visual and inertial data streams; 2) the **Co-Evolution Decoder**, which dynamically fuses the multimodal information to produce robust pose predictions; and 3) the **Global Trajectory Refinement Module**, which further refines the global motion trajectory using foot-ground contact information to mitigate issues such as foot sliding and pose drift.

The primary contributions of this paper are as follows:

(1) We propose a dual-stream motion estimation framework that separately extracts human motion features from visual and inertial signals, followed by a dynamic fusion process for accurate pose estimation.

(2) We design an Adaptive Fusion Module (AFM), which utilizes a cross-attention mechanism to dynamically adjust the weighting of visual and inertial signals, optimizing the fusion based on the reliability of each modality in different scenarios.
(3) We introduce a global trajectory refinement module that incorporates foot-ground contact feedback to refine the global motion trajectory, addressing the challenge of drift and error accumulation in IMU data.

We evaluate our MMAM-Cap on several publicly available datasets, demonstrating its effectiveness and robustness. The experimental results show significant improvements in both pose estimation accuracy and motion trajectory smoothness compared to existing methods. Our system excels in handling complex motion scenarios, making it a valuable tool for multimodal motion capture in real-world applications where occlusions and sensor drift are common challenges.

2 METHOD

2.1 Overview

The overall framework of our Multimodal Accurate Mocap (MMAM-Cap) system is illustrated in Figure 1. MMAM-Cap takes as input monocular video and acceleration and rotation data from six IMU sensors attached to the body. Our objective is to predict a sequence of SMPL model parameters in the IMU coordinate system, which includes local body pose (pose) and global root translation (trans). MMAM-Cap is composed of three main modules: *(1)* **Two-Stream Motion Estimation Module (TSE):** This module extracts human motion features from both the camera and IMU coordinate systems. *(2)* **Co-Evolution Decoder:** It fuses features from both modalities to reconstruct human poses in the camera coordinate system. *(3)* **Global Trajectory Refinement Module:** This module uses foot-ground contact information to refine the trajectory in the IMU coordinate system, improving global consistency.

2.2 Preliminaries

In this study, we utilize the SMPL model [5] to represent the 3D human body. The SMPL is a parametric mesh model denoted as $M(\theta, \beta, r, \pi) \in \mathbb{R}^{6890\times3}$, where $\theta \in \mathbb{R}^{23\times3}$ encodes the relative rotations of 23 body joints, and $\beta \in \mathbb{R}^{10}$ represents the shape parameters that control body shape variation. The parameters $r \in \mathbb{R}^3$ and $\pi \in \mathbb{R}^3$ define the root orientation and translation relative to the camera, respectively. In the world coordinate system adopted in our framework, $T_t = \{r_t, \pi_t\}$ refers to the global position and orientation of the body, while $\Theta_t = \{\theta_t, \beta_t\}$ describes the local body pose and shape at time step t.

2.3 Two-Stream Motion Encoder

The Two-Stream Motion Encoder in our MMAM-Cap system is designed to extract human motion features from two complementary data streams: visual signals and inertial signals. Visual signals, captured by a monocular camera, primarily contain information about the body's pose in the camera coordinate system. In contrast, inertial signals provide localized motion information relative to the root node through IMUs attached to the body. Given that inertial signals may accumulate drift over time and are often sparse, visual signals are more reliable in scenarios where the body is fully visible. However, in cases of occlusion or challenging visual conditions, inertial signals provide a stable reference. To leverage the strengths of both modalities, we propose a novel dual-branch fusion framework that integrates these two streams for accurate pose estimation over time.

2.3.1 IMU Sensor Stream. Follow PIP [8] , the IMU sensor stream is built upon an RNN-based structure, which has shown great success in inertial sensor-based motion capture tasks. This branch adopts a coarse-to-fine strategy, following the Progressive Layered (PL) design to first estimate joint features and then progressively refine the full-body pose.

In the first stage, we input the IMU sensor measurements, including acceleration ($\mathbf{a}$) and rotation ($\mathbf{R}$), into an RNN module to predict the 3D coordinates of leaf nodes (extremities such as hands or feet), denoted as $p_{\text{leaf}} \in \mathbb{R}^{15}$. Each input vector $\mathbf{x_r} \in \mathbb{R}^{(3+9)n}$ contains the acceleration ($\mathbf{a} \in \mathbb{R}^3$) and rotation matrix ($\mathbf{R} \in \mathbb{R}^{3\times3}$) for each of the $n = 6$ IMUs.

In the second stage, the predicted leaf node coordinates p_{leaf} are concatenated with the raw IMU data and passed through the Progressive Aggregation (PA) module. This module uses another RNN to output the full-body pose, estimating the relative 3D coordinates of all 24 joints with respect to the root node. This hierarchical approach effectively captures both low-level joint movements and high-level global pose information.

2.3.2 Video Feature Stream. For the visual stream, we first extract 2D human keypoints from monocular video frames using an off-the-shelf 2D pose estimator. These keypoints are then normalized to the person's bounding box and further enriched by appending the bounding box center and scale information to the feature vector, following the design of CLIFF [3].

Since 2D keypoints alone are sparse and lack depth information, we utilize a pretrained image encoder, designed for dense human mesh recovery tasks [2], to extract richer semantic features from the images. These features, denoted as ϕ_i, contain detailed visual context relevant to human 3D pose and shape. This additional information compensates for the limitations of the sparse 2D keypoints and significantly improves the precision of 3D pose reconstruction.

To align the video feature stream with the IMU sensor stream, we also employ an RNN as the encoder for the visual branch. The extracted visual features are fused with the IMU features in later stages to capture both motion and appearance information for human pose estimation.

2.3.3 Loss Function Design. To ensure accurate pose estimation, we supervise the model using a combination of 3D joint loss and 2D keypoint loss. The total loss function is defined as:

$$L = \lambda_{3D} L_{3D} + \lambda_{2D} L_{2D} \tag{1}$$

$L_{3D} = ||\hat{J}_{3D} - J_{3D}||^2$ supervises the 3D joint coordinates, where $\hat{J}_{3D}$ are the predicted 3D joints and J_{3D} is the ground truth. $L_{2D} = ||\hat{J}_{2D} - \Pi(J_{3D})||^2$ supervises the 2D keypoints, where $\hat{J}_{2D}$ are the predicted 2D keypoints and $\Pi(J_{3D})$ are the projected 2D keypoints from the predicted 3D joints.

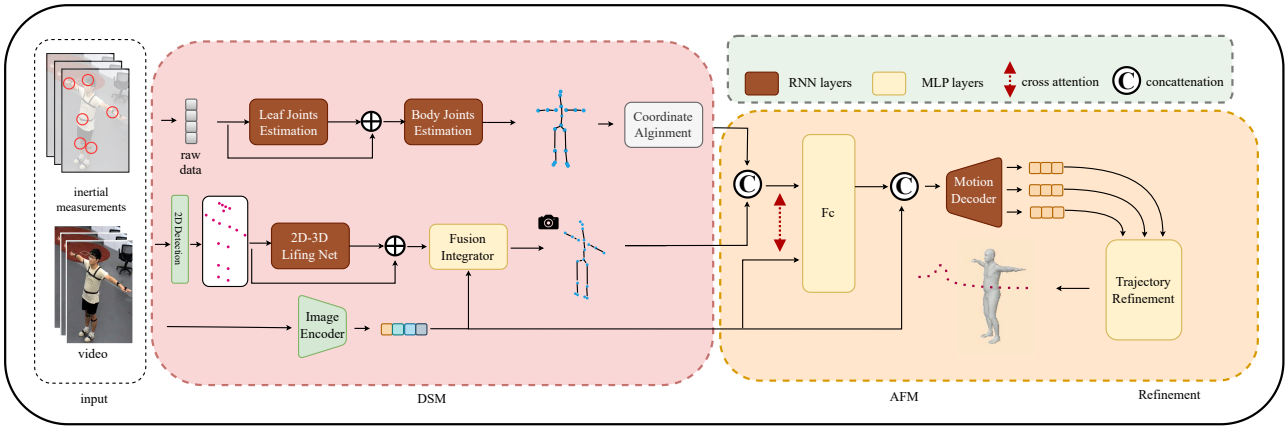

Figure 1: Overview of our Multimodal Accurate Mocap (MMAM-Cap) system. The process starts with monocular video and IMU sensor data from key body regions. Human motion features are extracted through the Two-Stream Motion Encoder (TSE), which processes both visual and inertial data streams. The Co-Evolution Decoder then dynamically fuses these features, leveraging cross-attention to adaptively balance contributions from both modalities. The Motion Decoder predicts SMPL model parameters, including joint rotations, body shape, and global translation. To further improve motion realism, the Global Trajectory Refinement Module adjusts foot-ground contacts and refines the global root trajectory, ensuring accurate and consistent motion capture across different conditions.

This loss design encourages the model to leverage both 2D and 3D information, with the visual stream containing the L_{2D} term exclusively, ensuring that the model learns to align the 3D pose with the 2D observations from video.

2.4 Co-Evolution Decoder

The Co-Evolution Decoder is designed to dynamically fuse multimodal information from both the IMU sensor and video streams, enabling the system to adapt to the advantages of each modality under different conditions. Traditional methods, such as HybridCap [4], simply concatenate IMU measurements with 2D keypoints, while RobustCap [6] applies decision-based fusion strategies by linearly weighting the results based on a predefined policy. However, these approaches fail to fully exploit the complementary strengths of each modality. To address this limitation, we propose an Adaptive Fusion Module (AFM) coupled with a motion decoder that dynamically adjusts the contribution of each signal, facilitating more effective and seamless fusion.

In the proposed AFM, cross-attention is used to blend the information from both modalities. First, the 3D joint coordinates from the IMU stream (p_{imu}) are transformed into the camera coordinate system using the camera parameters. These transformed IMU coordinates are then concatenated with the visual stream results (p_{vis}) to form a comprehensive joint representation p. Next, we use the image features (ϕ_i) as keys and values, and p as the query in a cross-attention mechanism. The attention weights from this mechanism determine the significance of each modality's features, allowing the model to dynamically adjust based on the scenario. This adaptive weighting helps the system leverage visual signals when they are reliable while falling back on IMU data during occlusions or other visual disturbances.

After the cross-attention mechanism, the fused features are passed through a fully connected layer (FC), which further integrates both signals into a final feature representation (ϕ_f). This combined feature is then concatenated with the image features and fed into the motion decoder.

The **Motion Decoder**, built on an RNN-based structure, extracts temporal dependencies from the fused feature (ϕ_f) and predicts the final human pose parameters. Specifically, the motion decoder consists of four fully connected layers, each responsible for predicting one of the following outputs: 1. Pose (local body pose): captures the SMPL model's body joint rotations. 2. Shape: estimates the SMPL body shape parameters. 3. Global Translation (trans): outputs the global translation of the body in the camera coordinate system. 4. Contact State (contact): determines whether specific body parts are in contact with the ground.

The loss function for the Co-Evolution Decoder is designed to ensure accurate prediction of 3D joints, body shape, and contact states:

$$L = \lambda_{3D}L_{3D} + \lambda_{SMPL}L_{SMPL} + \lambda_{contact}L_{contact} \tag{2}$$

where each term is defined as: $L_{3D} = ||\hat{J}_{3D} - J_{3D}||^2$, representing the error between predicted and ground truth 3D joints. $L_{SMPL} = ||\hat{\Theta} - \Theta||^2$, which measures the difference between predicted and ground truth SMPL parameters. $L_{contact}$ uses binary cross-entropy (BCE) to model the contact state prediction accuracy.

Unlike the previous terms, $\hat{J}_{3D}$ refers to the 3D joints reconstructed from the SMPL model, ensuring consistency between the predicted parameters and the estimated joint locations.

This adaptive, dynamic fusion strategy is effective in utilizing the strengths of each modality, enhancing the overall precision of pose and shape recovery across different challenging conditions.

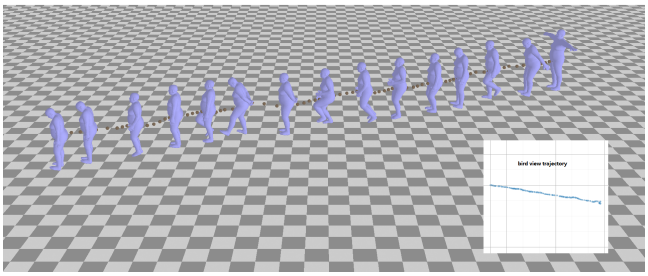

Figure 2: visualization of our results

2.5 Trajectory Correction Module

To enhance the accuracy of global motion prediction,follow WHAM, we introduce an additional decoder (DT) that estimates the coarse global root orientation $\Gamma(t)_0$ and root velocity $v(t)_0$ based on the fused motion features ϕ_m. Since ϕ_m originates from input signals captured in the camera coordinate system, maintaining realism in the predicted human motion necessitates further refinement. Thus, we integrate a trajectory optimization network that utilizes foot-ground contact feedback to adjust the predicted poses and trajectories. In realistic 3D motion within the world coordinate system, proper foot-ground contact is crucial to avoid slippage. Our trajectory correction module addresses this issue by constraining both the foot translation and root trajectory, particularly when the foot-ground contact probability p approaches a high value, ensuring more accurate and physically plausible motion.

3 EXPERIMENTS

3.1 Datasets

In the pre-training stage, following the work of WHAM [7], we trained the video stream's 2D lifting model on multiple datasets, including AMASS, 3DPW, Human3.6, and MPIINF-3DHP, using noise injection and interpolation techniques to enhance model performance. For AMASS, we simulated random camera views and weak perspective projections to generate 2D-3D data pairs. For the IMU stream, we followed the pip method and pre-trained on the DIP-IMU dataset. The Minions dataset [1] was used exclusively for fine-tuning and evaluation.

3.2 Implementation Details

During training, we set the temporal window size to 81. We used a batch size of 16 and employed the AdamW optimizer with an initial learning rate of 1×10^{-5}.

Regarding data preprocessing, we utilized several techniques to enhance the input data quality: We used `vit-pose` and `yolo-v8` for extracting 2D keypoints from video frames. These models help ensure accurate detection of human poses from monocular video data. For IMU data, we applied `smoothnet` to smooth both `imu_acc` and `imu_ori` signals. This step was essential to reduce noise and sudden variations in the IMU accelerometer and orientation data, resulting in more stable and reliable motion input.

Other implementation details include:

- **GPU:** RTX 3090
- **Operating System:** Ubuntu 20.04
- **CUDA:** 11.7
- **Language:** Python 3.9
- **Framework:** PyTorch 2.0.1

3.3 Metrics and Results

We evaluated the performance in the IMU coordinate system. To measure the accuracy of the estimated 3D human pose and shape, we computed several metrics, including:

Mean Per Joint Position Error (MPJPE) in millimeters, Procrustes-Aligned MPJPE (PA-MPJPE), Jitter (measured by the second derivative of acceleration in m/s²) to evaluate the frame-to-frame smoothness of the reconstructed motion. In particular, for MPJPE, we only aligned the first frame's trajectory, effectively assessing the accuracy of trajectory estimation in the IMU coordinate system.

The experimental results are shown in Table 1, which presents the changes in each metric using different methods. As can be seen, fusing multimodal information results in more accurate rotation (pose) estimations, and adding the trajectory refinement module helps reduce the cumulative error caused by the inertial data. All metrics, including MPJPE, PA-MPJPE, and jitter, show decreasing values, indicating that the refinement steps progressively enhance the model's performance, achieving a balanced overall effect.

Method	MPJPE ↓	PA-MPJPE ↓	Jitter ↓
Only-cam	182.31	21.17	3.82
Fuse	229.20	11.46	3.84
Fuse+refine	184.81	11.46	3.82
development baseline	693.94	81.55	0.00

Table 1: Comparison of different methods on various metrics. Lower values indicate better performance. Incorporating fusing multimodal data and trajectory refinement leads to reduced MPJPE, PA-MPJPE, and jitter, showing improvement in both accuracy and motion smoothness.

Figure 2 shows the visual comparison of our model's output in Minions dataset. The visual comparison demonstrates our method achieves accurate 3D human pose estimation while also producing smooth, natural motion trajectories.

4 CONCLUSION

In summary, the MMAM system provides an innovative solution to the longstanding challenge of multimodal motion capture. By effectively combining monocular video and IMU sensor data, our approach overcomes the limitations of individual modalities and delivers precise, consistent results, even in complex and dynamic environments. Future work will explore expanding the system's capabilities to more diverse datasets and further optimizing its performance across a broader range of motion scenarios.

REFERENCES

[1] Xiaodong Chen, Wu Liu, Qian Bao, Xinchen Liu, Quanwei Yang, Ruoli Dai, and Tao Mei. 2024. Motion Capture from Inertial and Vision Sensors. *arXiv preprint arXiv:2407.16341* (2024).

[2] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14783–14794.

[3] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. 2022. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*. Springer, 590–606.

[4] Han Liang, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu. 2023. Hybridcap: Inertia-aid monocular capture of challenging human motions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1539–1548.

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

[6] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. 2023. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

[7] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. 2024. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2070–2080.

[8] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13167–13178.