PitcherNet: Powering the Moneyball Evolution in Baseball Video Analytics

Supplementary Material

The supplementary material is organized into the following sections:

1. Section A: Architecture of D2A-HMR 3D human modeling technique.
2. Section B: Comparison of the proposed pitcher identification network with jersey number techniques.
3. Section C: Qualitative comparison of the PitcherNet system and various components including the depth encoders and D2A-HMR 2.0.
4. Section D: Limitations of the proposed system.

A. D2A-HMR Architecture

In this section, we explain the D2A-HMR architecture proposed in [11] in detail. D2A-HMR leverages a transformer-based architecture by incorporating scene-depth information, which is crucial to resolving the ambiguities inherent in single-image data. By jointly learning the distribution of human body shapes and scene-depth, D2A-HMR aims to produce robust 3D human mesh reconstructions, especially for scenarios with unseen data variations.

Algorithm 1 Distribution and Depth Aware Human Mesh Recovery

Input: Image (**I**)
Initialization:
$\boldsymbol{E}(\mathbf{I}) \rightarrow \mathbf{D}$
$\boldsymbol{F}(\mathbf{I}, \mathbf{D})$
Positional Embedding:
$P_e (= \omega_1 P_l + \omega_2 P_s) \rightarrow z_{\text{img}}, z_{\text{depth}}$
Self-Attention (MHSA):
$\text{MHSA}(z_{\text{img}}) \rightarrow z'_{\text{img}}$
$\text{MHSA}(z_{\text{depth}}) \rightarrow z'_{\text{depth}}$
Cross-Attention (MHCA):
$\text{MHCA}(z'_{\text{img}}, z'_{\text{depth}}) \rightarrow z_{\text{c}}$
Learnable Fusion Gates:
$z = \omega_3 z'_{\text{img}} + \omega_4 z'_{\text{depth}} + (1 - \omega_3 - \omega_4) z_{\text{c}}$
Masked Modeling:
$q_{mask} = \text{Mask}(z)$
Distribution Matching:
$\boldsymbol{R}(z) \rightarrow \sigma, \mu$
$\bar{\mu} = (\mu - \mu_{gt})/\sigma \rightarrow NF \rightarrow \mathcal{L}_{RLE}$
Silhouette Decoder:
$\mathbf{I}_{silh} = \boldsymbol{D}(z, k, s, p)$
Output: 3D mesh vertices, $\mathcal{P} = \boldsymbol{R}(z), \mathcal{P} \in \Re^{6890 \times 3}$

Algorithm 1 outlines the core steps of the D2A-HMR model. Initially, a depth encoder $E(I)$ takes an input image (I) and generates a depth map (D). Concurrently, both I and D are fed as input to the feature extractor (F), followed by hybrid positional encoding P_e, yielding tokens z_{img} and z_{depth}. These tokens subsequently undergo processing by self-attention and cross-attention modules, resulting in z'_{img}, z'_{depth} and z_c respectively. Fusion gates then merge these outputs into a singular token, z.

To enhance model performance, three refinement modules are employed: masked modeling, distribution modeling, and a silhouette decoder. A log-likelihood residual approach facilitates distribution modeling, enabling the model to learn deviations in the underlying distribution, and consequently generalize more effectively to unseen data. Additionally, masked modeling and a dedicated silhouette decoder refine the mesh shape and feature representation.

B. Pitcher Identification

The impact of the pitcher identification task is compared with the classical techniques that use jersey number cues are presented in Table 6. The classification labels include pitcher, batter, catcher, and player (which includes the fielders and referee). The inputs for the classification task are all the tracklets obtained from the player detection and tracking algorithm, with outputs representing the class for the tracklet detections.

Table 6. Comparison of our model with state-of-the-art jersey number identification techniques on MLBPitchDB dataset [11].

	Test Accuracy ↑
Gerke *et al.* [21]	64.47
Li *et al.* [30]	88.29
Vats *et al.* [48]	89.46
Balaji *et al.* [2]	93.68
Balaji *et al.* [3]	94.70
Ours	**96.82**

Table 6 shows that methods that only rely on jersey numbers for player identification underperform on this dataset due to the frequent absence of visible jersey numbers in many video frames. This highlights the importance of decoupling player actions within individual tracklets (sequences of detections associated with a single player) for improved identification accuracy. Therefore, our proposed approach, which incorporates a TCN block to decouple the underlying actions in each tracklet, achieves a significant performance increase of 2.12% compared to methods solely dependent on jersey numbers.

Pitch Hand	| **Pred**: Left	| **GT**: Left
Pitch Position	| **Pred**: Stretch	| **GT**: Stretch
Pitch Velocity	| **Pred**: 90.48 Mph	| **GT**: 87.58 Mph
Release Point	| **Pred**: 90	| **GT**: 90
Extension	| **Pred**: 5.85 feet	| **GT**: 6.13 feet
Pitch Hand	| **Pred**: Left	| **GT**: Left
Pitch Position	| **Pred**: Windup	| **GT**: Windup
Pitch Velocity	| **Pred**: 85.76 Mph	| **GT**: 89.17 Mph
Release Point	| **Pred**: 88	| **GT**: 89
Extension	| **Pred**: 6.01 feet	| **GT**: 6.16 feet
Pitch Hand	| **Pred**: Right	| **GT**: Right
Pitch Position	| **Pred**: Windup	| **GT**: Windup
Pitch Velocity	| **Pred**: 85.46 Mph	| **GT**: 85.65 Mph
Release Point	| **Pred**: 87	| **GT**: 87
Extension	| **Pred**: 6.17 feet	| **GT**: 6.11 feet

Figure 7. **Qualitative results.** Performance of the PitcherNet system in capturing various pitch statistics from the player tracklets. Here, *P red.* denotes the prediction from the 3D pose information and *GT* denotes the ground truth game data.

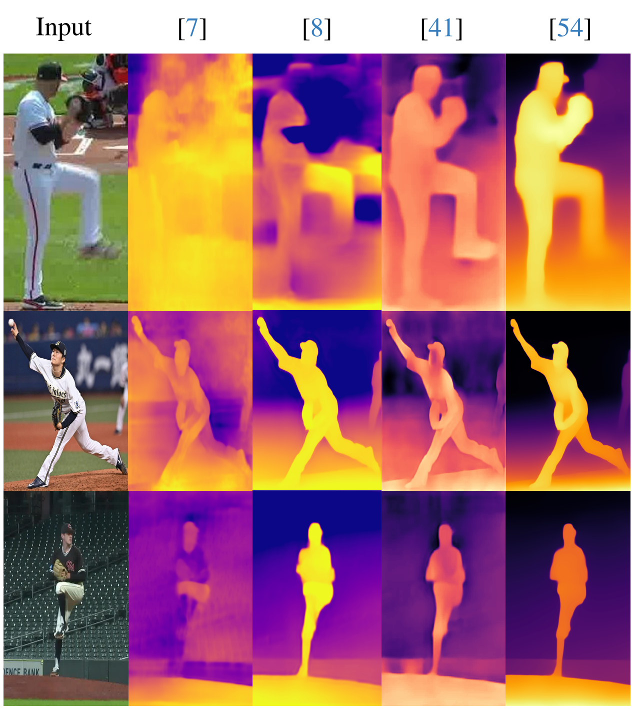

Figure 8. **Qualitative results.** Qualitative comparison of the various depth estimation techniques in MLBPitchDB baseball dataset.

C. Qualitative Results

PitcherNet System. The provided results in Figure 7 highlight the qualitative performance of the PitcherNet system in the MLBPitchDB dataset [11]. These visualizations underscore the effectiveness and robustness of our system in achieving accurate alignment with input pitch tracklets.

Depth Encoder. Our approach utilizes a monocular depth estimation model as the initial step in the 3D human model generation process. Figure 8 qualitatively compares the performance of various techniques, including AdaBin [7], ZoeDepth [8], DINOv2 [41], and Depth Anything [54]. As evident from the figure, Depth Anything [54] exhibits consistently superior depth estimation accuracy compared to the other methods. Consequently, we leverage [54] as the depth encoder within our 3D human model framework.

D2A-HMR 2.0. Figure 9 presents qualitative results obtained by D2A-HMR 2.0 on various outdoor activities. These visualizations demonstrate the model's capability to achieve accurate alignment with input images, even in complex real-world scenarios. This highlights the effectiveness and robustness of D2A-HMR 2.0 for handling diverse outdoor environments.

Figure 9. **Qualitative results.** Qualitative comparison of D2A-HMR 2.0 on COCO and sports datasets with unusual poses.

D. Limitations

PitcherNet, like many video analysis systems, is susceptible to error accumulation due to its reliance on a chain of interconnected components. Each step, from player identification to pitch analysis, introduces a degree of error. These errors can propagate throughout the processing pipeline, potentially leading to inaccuracies in the final extracted statistics. For instance, as shown in Figure 10, motion blur during fast pitching can hinder the ability of the 3D human model (e.g., D2A-HMR 2.0) to accurately estimate joint positions, particularly in the pitching hand. This, in turn, significantly affects the performance of the extracted pitch statistics.

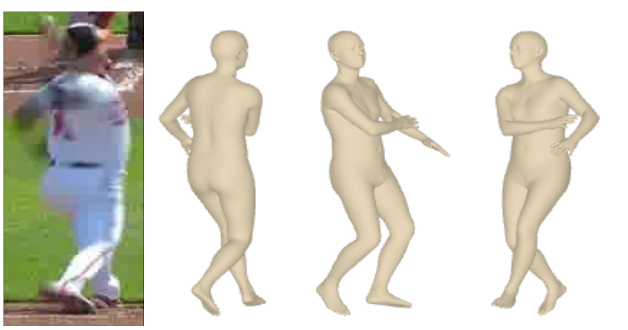

Figure 10. **Limitations of the work.** The 3D human model falters to estimate the mesh vertices with severe motion blur.

To address the issue of severe motion blur and self-occlusion, we propose investigating strategies such as part-based regression (inspired by works such as [27]). This enhancement aims to better equip the model to effectively handle challenging conditions characterized by occlusion and motion blur.