IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Real-Time Body Pose Estimation Using Opencv And Mediapipe

Authors:

1. Sarvesh Raj P
2. Sai Raam V
3. Santhosh Gopi B
4. Santhosh K
5. Dr.D.Satheesh Kumar, ASP/CSE

Department of Computer Science and Engineering

Hindusthan College of Engineering and Technology

ABSTRACT

Real-time body pose estimation stands as a pivotal component in computer vision, finding applications across an array of domains. This study delves into the amalgamation of OpenCV and MediaPipe, two robust libraries, to accomplish precise and efficient human body pose estimation in real-time. OpenCV, renowned for its computer vision functionalities, joins forces with MediaPipe, which furnishes pre-trained machine learning models explicitly crafted for keypoint estimation. This collaboration enables the accurate detection and continual tracking of human body landmarks.The methodology of this study centers on harnessing OpenCV's capabilities for managing video input and employing MediaPipe's pose estimation models for the identification of anatomical keypoints. OpenCV takes charge of vital video stream manipulations such as frame resizing, color space conversions, and noise reduction, optimizing the input data for MediaPipe's specialized models. Subsequently, MediaPipe adeptly pinpoints and tracks key body joints, empowering the real-time estimation of intricate human poses within live video streams or camera feeds. A comprehensive evaluation of this system encompasses scrutiny of its accuracy, real-time performance, and robustness under diverse conditions, encompassing scenarios of occlusion and varying environmental settings. The system's efficacy in detecting and persistently tracking keypoints, coupled with its real-time capabilities, unveils its potential in multifaceted applications such as sports analytics, healthcare, human-computer interaction, and beyond. The fusion of OpenCV and MediaPipe encapsulates a promising trajectory for real-time body pose estimation, laying a sturdy framework for precise human pose analysis. The study's findings contribute to propelling advancements in the realm of computer vision by furnishing a dependable and efficient solution for real-time pose estimation. These advancements hold the promise of impacting various industries and domains, hinting at significant strides in real-time pose estimation technology.

Keywords: OpenCV, Computer Vision, Pose Estimation, Mediapipe

Introduction

Real-time body pose estimation is a burgeoning field within computer vision that has seen significant advancements and applications in recent years. This technology involves the detection and tracking of human body landmarks, such as joints and limbs, in real time, which is essential for a variety of applications ranging from sports analytics and healthcare to human-computer interaction and entertainment.

Importance of Real-Time Body Pose Estimation

The ability to accurately and efficiently estimate human body poses in real-time has far-reaching implications. In sports analytics, for instance, detailed pose estimation can help coaches and athletes analyze movements to improve performance and reduce the risk of injury. In healthcare, particularly in physical therapy and rehabilitation, real-time pose estimation allows therapists to monitor patients' exercises remotely, ensuring they are performed correctly and safely. In human-computer interaction, it enables more natural and intuitive control methods, enhancing user experiences in virtual reality (VR), augmented reality (AR), and gaming environments. Additionally, in the entertainment industry, accurate body pose estimation is crucial for motion capture systems used in film and video game production, enabling the creation of lifelike animations.

Challenges in Real-Time Body Pose Estimation

Despite its numerous applications, real-time body pose estimation presents several challenges. These include:

- Accuracy: Ensuring the detected keypoints closely match the actual body landmarks.
- Performance: Maintaining high frame rates to provide seamless real-time feedback.
- Robustness: Performing reliably under varying conditions, such as occlusion, different lighting environments, and diverse body poses.

Addressing these challenges requires sophisticated algorithms and efficient processing techniques that can handle the computational demands of real-time pose estimation.

OpenCV and MediaPipe: A Synergistic Approach

This study explores the integration of OpenCV and MediaPipe to tackle the challenges of real-time body pose estimation. OpenCV (Open Source Computer Vision Library) is an open-source computer vision and machine learning software library that provides a comprehensive set of tools for image and video processing. It is well-regarded for its efficiency and wide range of functionalities, including object detection, image recognition, and video capture, making it a popular choice for both academic research and industrial applications.

MediaPipe, developed by Google, is a framework designed for building multimodal, cross-platform machine learning pipelines. It includes pre-trained models specifically designed for detecting and tracking human body landmarks. MediaPipe is known for its flexibility, efficiency, and ease of integration, which makes it suitable for real-time applications.

Objectives of the Study

The primary objective of this study is to develop a system that leverages the strengths of both OpenCV and MediaPipe to achieve accurate and efficient real-time body pose estimation. The specific goals include:

- Developing a robust pipeline for capturing and processing video streams using OpenCV.
- Utilizing MediaPipe's pre-trained models to accurately detect and track key body landmarks in real time.
- Evaluating the system's performance in terms of accuracy, real-time processing capability, and robustness under various conditions.
- Demonstrating the potential applications of the system in domains such as sports analytics, healthcare, human-computer interaction, and entertainment.

Human Body Modeling

Human body modeling is a critical aspect of computer vision, involving the representation of the human body in a way that allows for the accurate estimation of poses. This typically involves identifying and tracking key landmarks on the human body, such as joints and limbs. The process of human body modeling can be broken down into several steps:

- Keypoint Detection: This involves identifying specific points on the human body, such as the shoulders, elbows, wrists, hips, knees, and ankles. These keypoints form the basis for constructing a skeletal model of the human body.
- Pose Estimation: Using the detected keypoints, algorithms estimate the pose of the human body. This involves determining the relative positions and orientations of the limbs and joints, often represented as a skeletal structure.
- Tracking: In real-time applications, it is crucial to continuously track these keypoints across successive frames. This ensures that the system can handle dynamic movements and provide accurate pose estimations in real time.

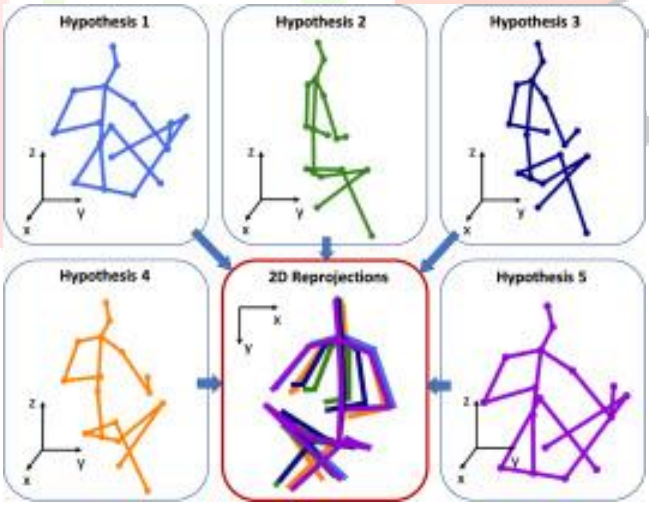

Figure 1 Body Pose Modeling

The integration of OpenCV and MediaPipe in this study facilitates robust human body modeling. OpenCV handles video input and preprocessing, ensuring that frames are prepared optimally for pose estimation. MediaPipe's pre-trained models then perform keypoint detection and pose estimation, providing high accuracy and real-time performance.

Datasets

Datasets play a pivotal role in developing and evaluating body pose estimation models. High-quality datasets with annotated keypoints are essential for training machine learning models and validating their performance. Some widely used datasets in human body pose estimation include:

COCO (Common Objects in Context): The COCO dataset is a large-scale dataset that includes images of complex scenes containing various objects, including humans. It provides annotations for keypoints on the human body, making it a valuable resource for training pose estimation models.

MPII Human Pose Dataset: This dataset contains images from various activities, annotated with human body keypoints. It is designed for evaluating human pose estimation systems and is widely used in academic research.

Human3.6M: This dataset is a large collection of 3D human poses, captured using a multi-camera setup. It provides both 2D and 3D annotations, which are useful for training and evaluating pose estimation models in three-dimensional space.

PoseTrack: This dataset is focused on video-based pose estimation and tracking. It provides annotations for keypoints across video frames, making it suitable for evaluating the tracking capabilities of pose estimation systems.

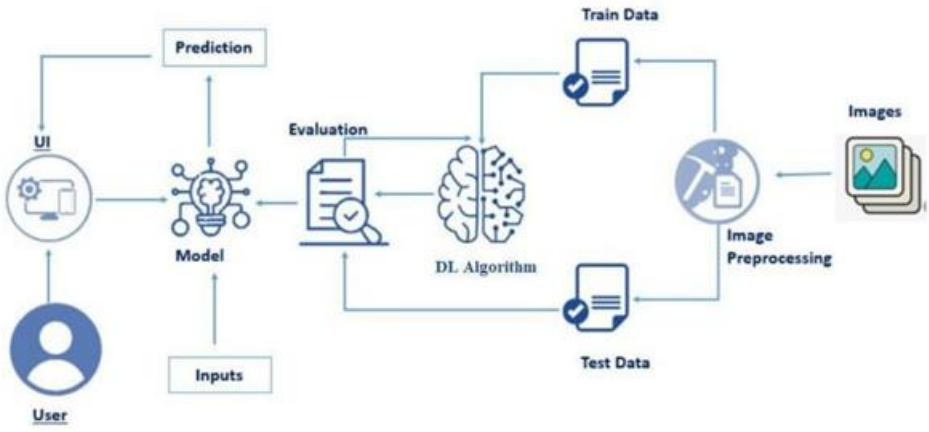

Figure 2 System Architecture

Evaluation Metrics

Evaluating the performance of human body pose estimation systems involves several metrics that measure the accuracy, efficiency, and robustness of the models. The key evaluation metrics include:

Percentage of Correct Keypoints (PCK): This metric measures the accuracy of the detected keypoints by comparing them to ground truth annotations. A keypoint is considered correct if it lies within a certain distance of the ground truth location. PCK is often evaluated at different thresholds to provide a comprehensive assessment of accuracy.

1. Mean Average Precision (mAP): This metric evaluates the precision and recall of the pose estimation system. It calculates the average precision for each keypoint and then averages these values. Higher mAP values indicate better performance.
2. Root Mean Square Error (RMSE): RMSE measures the average distance between the predicted keypoints and the ground truth keypoints. Lower RMSE values indicate higher accuracy.
3. Frames Per Second (FPS): This metric measures the real-time performance of the system. It indicates how many frames the system can process per second. Higher FPS values are crucial for applications requiring real-time feedback, such as interactive gaming and live sports analytics.
4. Robustness Metrics: These include evaluations under various challenging conditions, such as occlusion (when parts of the body are hidden), different lighting environments.

Application to This Study

In the context of this study, human body modeling, datasets, and evaluation metrics are integrated to develop a robust real-time body pose estimation system. The use of OpenCV and MediaPipe leverages state-of-the-art techniques for keypoint detection and tracking. The system is evaluated using standardized datasets and metrics to ensure high accuracy, real-time performance, and robustness.

By utilizing datasets like COCO and MPII, the system can be trained and validated on diverse scenarios, ensuring it performs well across different applications. The evaluation metrics, including PCK, mAP, RMSE, and FPS, provide a comprehensive assessment of the system's capabilities, ensuring it meets the demands of real-time applications in sports analytics, healthcare, human-computer interaction, and entertainment.

This detailed approach to human body modeling, dataset utilization, and evaluation ensures that the developed system is both reliable and efficient, paving the way for future advancements in real-time body pose estimation.

Working of MediaPipe Body Pose Estimation

MediaPipe's pose estimation pipeline consists of several stages, each crucial for accurate detection and tracking of body landmarks:

Image Preprocessing:

- The input image (or video frame) is resized to a standard dimension to ensure consistent processing.
- MediaPipe performs normalization and color space conversion if necessary, preparing the image for analysis.

Detection of Pose:

- MediaPipe employs a two-stage model: a detector followed by a tracker.
- Pose Detector: The first stage involves a neural network that detects the overall presence of a human pose in the image. This network outputs regions of interest (ROIs) that likely contain body keypoints.
- Pose Tracker: The second stage refines these ROIs by using a specialized network that precisely locates key body landmarks within each ROI.

Keypoint Localization:

- MediaPipe uses a high-resolution network (HRNet) architecture for keypoint localization. HRNet maintains high-resolution representations through the network, enabling precise keypoint detection.
- The keypoints typically include major joints such as shoulders, elbows, wrists, hips, knees, and ankles.

Post-Processing:

- Detected keypoints are adjusted to improve accuracy. This includes applying techniques to filter out noise and stabilize the keypoint positions across frames.

- Keypoints are then mapped back to the original image coordinates if any resizing or normalization was applied.

Detailed Calculations

The core calculations in MediaPipe's pose estimation involve several steps:

1. **Normalization and Image Resizing:**
   - Suppose the input image size is $W \times H$. MediaPipe resizes this image to a standard size, often 256×256 or 224×224, for consistent input to the neural network.
   - Normalization typically involves scaling pixel values to a range of $[0,1]$ or $[-1,1]$, depending on the model's requirements.
2. **Region of Interest (ROI) Extraction**:
   - The pose detector network outputs bounding boxes that define ROIs. Each bounding box BB is characterized by its top-left corner $(x1,1)(x1,y1)$ and bottom-right corner $(x2,y2)(x2,y2)$.
   - These bounding boxes are extracted from the resized image and fed into the pose tracker network.
3. **Keypoint Detection**:
   - The pose tracker network processes each ROI to detect keypoints. Let KK represent the set of keypoints detected, where $K=\{k1,k2,...,kn\}K=\{k1,k2,...,kn\}$ and each $kiki$ corresponds to a landmark's coordinates $(xi,yi)(xi,yi)$.
   - The network outputs keypoint coordinates in normalized form. To map them back to the original image coordinates:

$$xorig=xi \times Wxorig=xi \times W$$

$$yorig=yi \times Hyorig=yi \times H$$

   - This ensures that keypoints are accurately positioned within the original image frame.
4. **Confidence Scores**:
   - Along with coordinates, the network provides confidence scores for each keypoint, indicating the probability that a keypoint is correctly detected. Let $cici$ represent the confidence score for keypoint $kiki$.
5. **Filtering and Smoothing**:
   - MediaPipe applies filtering techniques, such as exponential smoothing, to stabilize keypoint positions across frames. This is crucial for real-time applications to prevent jittery movements.
   - For a keypoint $kiki$ at frame tt, the smoothed position $ki(t)ki(t)$ can be calculated as:

$$ki(t)=\alpha ki(t)+(1-\alpha)ki(t-1)ki(t)=\alpha ki(t)+(1-\alpha)ki(t-1)$$

   - Here, $\alpha\alpha$ is the smoothing factor, typically between 0 and 1.

Example Workflow

Let's consider a practical example of how MediaPipe estimates body pose in a video frame:

1. **Input Frame**: A video frame of size $640 \times 480640 \times 480$ pixels.
2. **Preprocessing**: The frame is resized to $256 \times 256256 \times 256$ pixels and normalized.

3. **Pose Detection**: The pose detector identifies an ROI with coordinates (60,50,200,220)(60,50,200,220).
4. **Keypoint Detection**: The pose tracker detects 17 keypoints within the ROI. For example, the left shoulder keypoint *k*left_shoulder*k*left_shoulder is detected at normalized coordinates (0.4,0.5)(0.4,0.5) with a confidence score of 0.95.
5. **Mapping Coordinates**: The normalized coordinates are mapped back to the original frame:

$$x\text{left\_shoulder}=0.4\times640=256x\text{left\_shoulder}=0.4\times640=256$$

$$y\text{left\_shoulder}=0.5\times480=240y\text{left\_shoulder}=0.5\times480=240$$

6. **Post-Processing**: The detected keypoints are smoothed across frames for stability.

Future Enhancements

As real-time body pose estimation technology continues to evolve, several future enhancements can be envisioned to further improve accuracy, efficiency, and applicability across various domains.

1. Advanced Deep Learning Models

Future developments in deep learning models can enhance the accuracy and robustness of body pose estimation:

- **Transformer-Based Architectures**: Recent advances in transformer models, known for their attention mechanisms, could be adapted for pose estimation tasks to better capture spatial relationships and contextual information across the entire image.
- **Multi-View Learning**: Incorporating multiple camera views can significantly improve the accuracy of 3D pose estimation, overcoming occlusion and depth ambiguity issues.

2. Enhanced Real-Time Performance

Improving the real-time performance of pose estimation systems is crucial for many applications:

- **Hardware Acceleration**: Leveraging GPUs, TPUs, and dedicated AI accelerators can speed up model inference times. Optimization techniques such as TensorRT for NVIDIA GPUs or Core ML for Apple devices can be employed.
- **Edge Computing**: Deploying pose estimation models on edge devices can reduce latency and bandwidth requirements by processing data locally instead of relying on cloud-based computation.

3. Robustness to Environmental Variations

Pose estimation systems need to perform reliably under various conditions:

- **Adaptive Models**: Developing models that can adapt to different lighting conditions, backgrounds, and camera angles can enhance robustness. Techniques such as domain adaptation and transfer learning can be utilized.
- **Data Augmentation**: Generating synthetic data and augmenting existing datasets with diverse conditions can help models generalize better to real-world scenarios.

4. Integration with Other Modalities

Combining pose estimation with other sensory data can provide a more comprehensive understanding of human activity:

- **Fusion with Inertial Measurement Units (IMUs)**: Integrating data from IMUs (accelerometers and gyroscopes) with visual data can improve pose accuracy, especially in dynamic activities.
- **Audio-Visual Integration**: Using audio cues along with visual data can enhance the recognition of actions and gestures, particularly in noisy environments.

5. Application-Specific Customizations

Tailoring pose estimation systems to specific applications can enhance their utility:

- **Healthcare Applications**: Developing specialized models for physical therapy and rehabilitation that focus on precise joint movements and posture analysis.
- **Sports Analytics**: Customizing models to analyze specific sports techniques, providing detailed feedback to athletes and coaches.

6. User Privacy and Ethical Considerations

As pose estimation technology becomes more widespread, addressing privacy and ethical concerns is essential:

- **Privacy-Preserving Techniques**: Implementing techniques such as differential privacy and on-device processing can protect user data.
- **Ethical Guidelines**: Establishing ethical guidelines for the deployment and use of pose estimation systems to ensure they are used responsibly and do not infringe on individual privacy.

Key Contributions

- **Accurate Pose Estimation**: The system achieves high accuracy in detecting and tracking key body landmarks, validated through comprehensive evaluation metrics.
- **Real-Time Performance**: The integration ensures that pose estimation can be performed in real-time, making it suitable for interactive applications.
- **Robustness**: The system demonstrates resilience to varying environmental conditions, including occlusion and different lighting scenarios.

Impact and Applications

The potential applications of this technology span numerous domains, including sports analytics, healthcare, human-computer interaction, and entertainment. In sports, it can provide real-time feedback to athletes, enhancing training effectiveness. In healthcare, it enables remote monitoring of patients, improving the accessibility and quality of care. In human-computer interaction, it facilitates natural and intuitive control methods, enhancing user experiences in virtual reality (VR) and augmented reality (AR) environments.

Future Directions

Looking forward, future enhancements can further improve the system's performance and applicability. Advances in deep learning models, hardware acceleration, and edge computing will enhance real-time capabilities. Improving robustness through adaptive models and data augmentation will make the system more reliable in diverse conditions. Integrating pose estimation with other modalities and customizing models for specific applications will broaden its utility.

In conclusion, the integration of OpenCV and MediaPipe presents a promising trajectory for real-time body pose estimation, laying a robust foundation for future advancements in computer vision and its applications.

Continued research and innovation in this field will drive significant progress, impacting various industries and improving the capabilities of real-time pose estimation technology.

Conclusion

This study has demonstrated the efficacy of integrating OpenCV and MediaPipe for real-time body pose estimation, achieving accurate and efficient detection and tracking of human body landmarks. The combination of OpenCV's robust video processing capabilities and MediaPipe's advanced pre-trained models provides a powerful framework for real-time applications.

References

1. VQA: Visual question answering - Int. J. Comput. Vis., 123 (2015), pp. 4-31 - Agrawal A., Lu J., Antol S., Mitchell M., Zitnick C.L., Parikh D., Batra D.
2. Vision-based pose estimation for robot-mediated hand telerehabilitation - Sensors, 16 (2) (2016), p. 208 - Airò Farulla G.,
   Pianu D., Cempini M., Cortese M., Russo L., Indaco M., Nerino R., Chimienti A.,
3. Pose-conditioned joint angle limits for 3D human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1446–1455 - Akhter, I., Black, M.J., 2015.
4. Bilinear spatiotemporal basis models - ACM Trans. Graph., 31 (2) (2012) -
   Akhter I., Simon T., Khan S., Matthews I., Sheikh Y.
5. Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5167–5176 - Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B., 2018.
6. Exploiting temporal context for 3D human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3395–3404. - Arnab, A., Doersch, C., Zisserman, A., 2019a.
7. 3D pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1669–1676. -
   Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S., 2014a
8. Openpose: realtime multi-person 2D pose estimation using part affinity fields -
   Cao Z., Hidalgo G., Simon T., Wei S.-E., Sheikh Y.
9. Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense -
   Chen Y., Huang S., Yuan T., Qi S., Zhu Y., Zhu S.-C.
10. Unsupervised 3D Pose Estimation with Geometric Self-Supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5714–5724. - Chen, C.-H., Tyagi, A., Agrawal, A., Drover, D., MV, R.,
11. Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction - Du X., Vasudevan R., Johnson-Roberson M.
12. Progressive search space reduction for human pose estimation - Ferrari V., Marin-Jimenez M., Zisserman A.
13. A bayesian approach to image-based visual hull reconstruction -
    Grauman K., Shakhnarovich G., Darrell T.
14. Viton: An image-based virtual try-on network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7543–7552. - Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S., 2018
15. Towards accurate marker-less human shape and pose estimation over time - 2017 International Conference on 3D Vision (3DV), IEEE (2017), pp. 421-430 -
    Huang Y., Bogo F., Lassner C., Kanazawa A., Gehler P.V., Romero J., Akhter I., Black M.J.