TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

BULLETIN OF TUIT:

MANAGEMENT AND COMMUNICATION TECHNOLOGIES

ISSN:2181-1083 Date:2025-06-10 DOI:
Fayzullayeva Z.I., Pirimqulova Z.A., Hojiyev S.N.
№ 1(9)2025

Algorithm for Human Action Recognition Based on Images and Visual Data Using OpenPose

Fayzullayeva Z.I.
Associate Professor of the Department of Basic of Informatics
Tashkent university of information technologies named after Muhammad al-Khwarizmi Tashkent city 100 000, Uzbekistan
zarnigor18z02@gmail.com

Pirimqulova Z.A
PhD student in the Department of Artificial Intelligence
Tashkent university of information technologies named after Muhammad al-Khwarizmi Tashkent city 100 000, Uzbekistan
zilolapirimqulova5@gmail.com

Hojiyev S.N
Assistant of the Department of Basic of Informatics
Tashkent university of information technologies named after Muhammad al-Khwarizmi Tashkent city 100 000, Uzbekistan
hojiyevsunatullo1991@gmail.com

***Abstract*—**Human action recognition is a fundamental task in the field of computer vision and has become increasingly important in applications such as human-computer interaction, intelligent surveillance systems, virtual and augmented reality, and smart transportation. With the rapid advancement of deep learning techniques and cutting-edge algorithms, the effectiveness and accuracy of action recognition systems have significantly improved in recent years. In this study, we propose a real-time human action prediction model based on skeletal keypoints extracted from static RGB images using the OpenPose framework. The model processes spatial configurations of human joints to identify and classify actions with high precision. By leveraging Part Affinity Fields (PAFs) and confidence maps, our system successfully estimates human poses and tracks body movements efficiently. Experimental results demonstrate that the proposed approach outperforms several traditional methods, including DeeperCut (58.1%) and Convolutional Pose Machines (CPM) (55.2%), achieving a Percentage of Correct Keypoints (PCK) of 85.2%. This high accuracy indicates the model's robustness and its potential for real-world applications that require efficient and accurate motion analysis in real-time scenarios. The results highlight the advantages of combining deep learning with skeletal keypoint estimation for advanced and scalable human action recognition systems.

***Keywords* —** OpenPose, YOLO, PAFs, HAR, CNN, Prediction layer, image, skeletal keypoints, RGB.

I. INTRODUCTION

Thanks to advances in convolutional neural networks (CNNs), and object detection algorithms such as YOLOv4 and Single Shot Detector (SSD), computers are now capable of detecting objects in images and videos almost in real time, which has brought significant advantages in various applications [1][2][3][4]. As object detection technologies have matured, a subsequent field known as Human Activity Recognition (HAR) has emerged, which focuses on identifying human actions or activities from visual data. HAR is a growing research

domain aimed at recognizing the behaviors of hu- mans, animals, or other moving entities, and continues to pose numerous challenges and open problems [3][5][6][7][8][9].

In this study, we present a system for human action recognition based on still images by utilizing OpenPose to generate human skeletal keypoints. The RGB-based approach benefits from prior visual knowledge and provides high- accuracy action recognition; however, it demands considerable computational power and storage resources, and is sensitive to background noise and variations in lighting conditions. In contrast, skeleton-based methods are more computationally efficient and less affected by illumination or background changes, but may suffer from limitations due to a lack of contextual information.

II. Related Work

In this paper, we present a robust human activity recogni- tion method that utilizes the open-source OpenPose library to extract anatomical keypoints from RGB images. These keypoints are then analyzed across sequential frames to derive stable motion features. To classify the associated activities, we employ a Recurrent Neural Network (RNN) with Long Short- Term Memory (LSTM) cells, which captures the temporal dependencies of the skeletal features. To enhance generaliza- tion and ensure subject-invariance, the model is trained using data from various subjects and multiple camera viewpoints. The proposed method achieves high performance, reaching an overall accuracy of 92.4%, significantly outperforming traditional approaches such as Support Vector Machines (SVMs), Decision Trees, and Random Forests, whose best accuracy typically does not exceed 78.5% [1].

Despite this progress, action recognition still faces several key challenges, including background interference, illumi- nation variations, privacy concerns, and high computational requirements. Skeleton-based datasets offer a promising al- ternative due to their independence from external environmental factors—such as backgrounds and appearance—and their lower computational overhead [16][17][13]. However, skeleton-based methods suffer from a lack of contextual in- formation and cannot be directly applied to raw video data without first extracting the skeletal representations.

Presti Liliana Lo et al. [13] reviewed 3D skeleton-based action recognition technologies, identifying critical challenges and motivations that remain highly relevant. Similarly, Ren Bin et al. [17] examined the development of deep learning-based methods, specifically the evolution and core technological advancements in RNNs, CNNs, and Graph Convolutional Networks (GCNs).

In their research, these scholars introduced a Kernel-Aware Adaptive

Graph Transformer Network (KA-AGTN), which models high-level spatial correlations between joints using a multi-head self-attention mechanism integrated within a graph transformer operator. Additionally, the Temporal Kernel Attention (TKA) module embedded in KA-AGTN computes attention scores at the channel level by leveraging temporal features, thereby enhancing the temporal correlation of human motion patterns [2].

III. PROPOSED METHOD

The proposed method distinguishes itself from previous approaches by its ability to operate effectively on multi- person images, its integration of Part Affinity Fields (PAFs), and its real-time performance. Earlier algorithms such as DeeperCut [1] and Convolutional Pose Machines (CPM) [2] were primarily designed for single-person pose estimation and often failed to correctly group keypoints in scenes containing multiple individuals. Specifically, joints from one person were frequently misassigned to another, resulting in inaccurate pose estimations.

OpenPose overcomes this limitation through the use of PAFs, which represent the spatial relationships between body parts as vector fields. This allows for more accurate grouping of keypoints, even in crowded scenes with multiple people. Using this approach, the system is capable of correctly detecting and constructing the skeletal structure of up to 10 individuals in a single image.

$$S_j(x,y) = \exp\left(-\frac{\square(x,y) - p_j\ \square^2}{\sigma^2}\right) \quad (1)$$

where:

- $S_j(x,y)$ denotes the confidence score at coordinate (x,y)
- for the j-th keypoint;
- p_j is the ground truth coordinate of the k-th keypoint;
- σ is the variance parameter in the Gaussian distribution

The OpenPose architecture accepts RGB images or video frames as input data and processes them through a deep Convolutional Neural Network (CNN) to detect 2D anatomical keypoints—typically 18 or up to 135 points including body, hand, facial, and foot joints. The spatial locations of these keypoints are determined using confidence maps. To accurately identify key body parts and estimate human poses, OpenPose generates a set of confidence maps and part affinity fields (PAFs), which represent both the likelihood of each keypoint and the spatial relationships between joints. For each person in the scene, the 2D coordinates (e.g., x, y positions of the shoulder, elbow, etc.) are calculated using the confidence map formula described in Equation (1).

During the detection phase, the input image is pre-processed through a sequence of convolutional and pooling layers to extract a comprehensive set of feature maps. These layers help to capture both low-level and

high-level features from the image. Each confidence map indicates the probability of the presence of a specific keypoint at a given location in the image. These feature maps are then used to generate part affinity fields, which encode the spatial relationships between different parts of the body. Each part affinity field represents the likelihood of a pairwise connection between two joints at a particular location in the image.

In the inference process, changes in the spatial position of keypoints across consecutive frames are computed to analyze motion dynamics. This includes evaluating the distances and angular displacements between joints, such as the relative position and direction from the shoulder to the elbow. The confidence maps are generated using a CNN architecture such as VGG-19 or MobileNet, as indicated in Equation (2).

$$F = CNN(I), \quad S = f(F) \qquad (2)$$

where:

• I - the input RGB image,

• F - represents the extracted feature map

• S - denotes the final confidence scores obtained through

• f - a mapping function applied to the features.

These features are robust to variations in viewing angles and individual body differences, making them suitable for person independent and view-invariant activity recognition.

$$L_c(x,y) = \begin{cases} 1, & \text{if } (x,y) \text{ lies on the limb segment} \\ 0, & \text{otherwise} \end{cases}$$

- v — Unit vector between joints (e.g., direction from elbow to wrist).
- L_c — PAFs for the c-th limb pair.

OpenPose adopts a ”bottom-up” approach, in which all keypoints are detected first and then grouped into individual persons using Part Affinity Fields (PAFs). This strategy enhances computational efficiency, as the analysis is performed only once regardless of the number of people present in the image. Traditional algorithms often rely on additional layers or post-processing stages to establish limb connections, which can reduce accuracy and increase computational cost.

PAFs represent the associations between body parts as vector fields, offering a precise and efficient way to group detected keypoints. This method enables accurate estimation of both the direction and distance between joints. The computation of PAFs using a convolutional neural network (CNN) can be formulated as follows:

$$L = g(F)$$

where g — PAFs predicting layers

$$L = \sum_{j=1}^{J}\sum_{p} W(p)\cdot \| S_j(p) - S_j^*(p) \|^2 + \sum_{c=1}^{C}\sum_{p} W(p)\cdot \| L_c(p) - L_c^*(p) \|^2 \qquad (3)$$

S_j^*, L_j^* -Real joint maps and PAFs.

$W(p)$ -Pixel weight.

J - Number of key joints (e.g., 18)

C -Number of limb pair.

$$E = \int_{u=0}^{u=1} L_c(p(u)) \cdot v \, du$$

E - Degree of matching joint accuracy

$p(u)$ - Line segment between joints.

OpenPose pose estimation is used to determine human action recognition, with the

following steps: first step realtime processing using LSTM or 3D CNN architecture:

$$h_t = o_t \cdot \tanh(C_t), \quad C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (4)$$

$$P(y|\ X) = \text{Softmax}(W_h \cdot h_T + b_h)$$

(4) formula is used to generate the output structure, trained using labeled output dataset, and the model is trained in realtime using online data.

Fig. 1. Diagram of the proposed human action recognition model using OpenPose.

A. Experimental Results and Conclusion

The experimental results demonstrated that OpenPose achieved an Average Precision (AP) of 61.8%, outperforming DeeperCut (58.1%) and Convolutional Pose Machines (CPM) (55.2%). Additionally, OpenPose reached a Percentage of Correct Keypoints (PCK) of 85.2%. When integrated with the MobileNet model, OpenPose attained a real-time performance of 10–15 frames per second (FPS) on images of size 432×368.

One of the distinctive advantages of OpenPose lies in its compatibility with lightweight backbone architectures such as MobileNet and MobileNetV2, which enables efficient realtime execution on resource-constrained devices, such as mobile phones and embedded systems.

B. Conclusion

The primary advantages of OpenPose are its bottom-up approach, its use of Part Affinity Fields (PAFs), and its real-time processing capabilities. The system demonstrates high accuracy in multi-person scenarios, high processing speed through lightweight models like MobileNet, and flexible deployment with TensorFlow integration. The algorithm effectively groups keypoints via confidence maps and PAFs, providing superior performance compared to alternative methods such as DeeperCut and Mask R-CNN. OpenPose has found extensive applications in domains such as sports analytics, healthcare, security systems, and gaming. Furthermore, it holds promising potential for future advancements in 3D pose estimation and motion analysis.

REFERENCES

[1] F. M. Noori, B. Wallace, M. Z. Uddin, and J. Torresen, “A Robust Human Activity Recognition Approach Using OpenPose, Motion Features, and Deep Recurrent Neural Network,” in *Proceedings*, 2022.

[2] Y. Liu, H. Zhang, D. Xu, and K. He, “Graph Transformer Network with Temporal Kernel Attention for Skeleton-Based Action Recognition,” Neurocomputing, 2022.

[3] E. Insafutdinov et al., “DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model,” in European Conference on Computer Vision (ECCV), 2016.

[4] S.-E. Wei et al., “Convolutional Pose

Machines," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] K. He et al., "Mask R-CNN," in IEEE International Conference on Computer Vision (ICCV), 2017.

[6] Z. Cao et al., "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[7] Z. Sun, J. Liu, Q. Ke, H. Rahmani, M. Bennamoun et al., "Human action recognition from various data modalities: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 4, pp. 67–77, 2022.

[8] G. Moon et al., "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[9] CMU Perceptual Computing Lab, "OpenPose GitHub Repository," https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[10] B. Xiao, H. Wu, and Y. Wei, "Simple Baselines for Human Pose Estimation and Tracking," in European Conference on Computer Vision (ECCV), 2018.

[11] H. Joo et al., "Panoptic Studio: A Massively Multiview System for Social Motion Capture," in IEEE International Conference on Computer Vision (ICCV), 2015.

[12] J. Liu et al., "NTU RGB+D: A Large-Scale Dataset for 3D Human Activity Analysis," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[13] A. F. Bobick and J. W. Davis, "The Recognition of Human Movement Using Temporal Templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001.

[14] S. Patel et al., "A Review of Wearable Sensors and Systems with Application in Rehabilitation," Journal of NeuroEngineering and Re- habilitation, 2012.

[15] R. Poppe, "A Survey on Vision-Based Human Action Recognition," Image and Vision Computing, 2010.

[16] M. Zyda, "From Visual Simulation to Virtual Reality to Games," IEEE Computer, 2005.

[17] Sadullayeva Sh.A., Fayzullayeva Z.I. Matnlarni semantik tahlil asosida anotatsiyalash// Digital transformation and artificial intelligence, 2(6), Raqamli Transformatsiya va Sun'iy Intellekt ilmiy jurnali, ISSN: 3030-3346, VOLUME 2, ISSUE 6, DECEMBER 2024. 117–122. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v2i624 - 5 %.

[18] Fayzullayeva Z.I., Mengliqulov Sh. Meylikulov Sh. Kobulov M., Education artificial intelligence systems and their use in teaching, AIP Conference Proceedings. – 2024. – Vol. 3244(2024) – P. 030049-1 - 030049-9