iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52514

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Golf Swing Analysis using Computer Vision

Rahul Patil[1], Yash Nimbalkar[2], Satyajeet Chavan[3], Abhijeet Mahajan[4]
[1, 2, 3, 4]*Department of Information TechnologyPune Institute of Computer Technology* Pune, India

Abstract*: Sports play a significant role in providing entertainment and recreational activities. As technology has advanced, video games featuring various sports and games have been developed using computer vision. In the game of golf, the golf swing is a crucial element that involves the entire body when players strike the ball. Having the correct posture is essential for a strong swing. However, beginners often struggle with identifying the keyframes they should focus on and which areas of their body they need to improve due to uneven timing and lack of expertise. To bridge this gap, this research proposes a neural network-based system for analyzing golf swings. The system utilizes monocular swing footage to offer an autonomous method for estimating the golfer's movement. Since amateur players often lack supervision during self-practice, this method can be particularly useful. The research also includes the design of an architecture that combines parts detection and parts association using image processing. By temporally aligning the swing videos, the system can estimate the golfer's pose, which can be further utilized for analysis.

Index Terms: Computer vision, Neural networks, Golf Swing, Image processing.

I. INTRODUCTION

Sports are currently a popular and highly sought-after form of recreation. Among the various sports, golf is widely enjoyed worldwide, and mastering the golf swing is a challenging task that involves using the entire body. Golfers often dedicate years of consistent practice to achieve a consistent and effective swing. However, many amateur players lack sufficient guidance and support when practicing on their own, which is crucial for analyzing their swing and making necessary adjustments to their mechanics. For amateur golfers, it can be impractical and challenging to set up an environment for practice, as it often requires financial resources and restricts players to a specific location. Therefore, it is essential to develop a simple method that allows amateur players to analyze their swings, aiding coaches in the grading process and providing off-class supervision. This method is in high demand in the field of golf instruction.

Thanks to significant advancements in machine learning technologies, several systems have been developed to recognize various objects, make predictions, and even forecast future events. Researchers have focused on creating self-training systems based on neural networks. One recent study introduced a training system that simulates climbing, where users receive suggested poses and actions based on their current positions, helping them learn and improve in the future. The advantage of machine learning is its ability to enable systems to learn without prior information or explicit programming.

II. LITERATURE SURVEY

A. Human-joint Affinity

A lightweight and scalable attention module called Two-step Squeeze and Excitation (S2E) has been developed. The module aims to achieve a balance between accuracy and speed while consuming minimal processing power. Comparative studies on the COCO dataset and the MPII dataset demonstrate that the S2E module outperforms other state-of-the-art approaches. It significantly improves prediction accuracy, offers a better tradeoff between speed and accuracy, and is more practical. These findings are supported by experiments conducted in [9]. To address the spatial and temporal affinity of human joints in videos, a new approach called Dynamical Spatial/Temporal Graph Convolution (DSG/DTG) has been introduced. This method deviates from conventional graph convolution and utilizes the spatial separation and temporal movement similarity of human joints within the video exemplars. It aims to find the spatial/temporal affinity between human joints. This development is discussed in [10]. Another fast and lightweight method has been proposed for accurate pose estimation. This method includes the FLPN network for pose estimation, a smart bottleneck block to reduce computational costs, and the use of the SSIM method to refine the appropriate ratio of intrinsic feature maps. This refinement helps in reducing the module block size while maintaining high accuracy. Detailed information about this approach can be found in [11]. The UCF sports dataset, which was compiled from broadcast television, offers a more realistic representation of sports actions. The dataset contains 150 action videos with a resolution of 480 x 720, divided into 10 action categories.

However, there is a limited number of action videos in each class, necessitating the use of the leave-one-out cross-validation (LOOCV) approach. On the other hand, the UCF101 dataset consists of 13,320 action videos categorized into 101 different action classes. These classes encompass various types of actions, including human-object interaction, body movement, interaction between people, using musical instruments, and sports. Further information about these datasets can be found in [12].

B. Human Pose Estimation

In order to understand human actions in photos and videos, person detection and pose estimation are essential. A two-step process is proposed to extract whole-body bounding boxes and body keypoint proposals. The coarse-pose proposal sub-net performs this extraction, while the pose refinement sub-net utilizes multi-scale supervision and regression to enhance context feature learning. Additional techniques such as structure-aware loss and keypoint masking are employed to improve posture refinement. Experimental results on COCO and OCHuman datasets demonstrate the effectiveness of this framework [1]. To achieve accurate, fast, and interpretable Human Pose Estimation (HPE), a novel transformer structure called LGPose is suggested. This structure combines large-kernel CNNs, including DW-CNN, dilated CNN, and PW-CNN, to capture regional patterns among neighboring pixels. To bridge the gap between 2D-based DW-CNN and 1D-based ViT, matrix flattening and sequence reshaping techniques are introduced for mutual conversions [2]. An unsupervised learning method is proposed to facilitate future applications and research in various sports and skill-training procedures. This method incorporates neural networks to develop a motion synchronizer, aligning motions with different phases and timings. Additionally, a motion discrepancy detector is employed to identify subtle variations between motions in latent regions learned by the networks [3]. The PCA-CBVR (Content-based video retrieval based on prototypical category approximation) method is utilized for video retrieval based on contextual similarities without additional information like tags. This method involves two main steps: category prediction of the user's query video and fine searching to retrieve the most similar video for each query video [4].

C. *Motion Tracking Methods*

A new approach based on deep reinforcement learning has been developed to evaluate racquet sports, allowing for a more detailed analysis of player movements beyond just considering the outcomes or scores. By inputting player poses and shuttlecock locations, an LSTM model is utilized to learn the underlying function. This method is discussed in [5]. In the context of grading moving poses, monocular swing footage is employed. The initial step involves extracting a set of 3D human poses from a swing video. Dynamic temporal warping (DTW) is then used to align the moving poses between a query and a reference video. The aligned swing videos are subsequently utilized for a distance-based grading technique. This approach is presented in [6].

Various methods for identifying violent behavior and actions involve pose estimation, spatio-temporal models, long short-term memory networks, and 3D convolutional neural networks (3D-CNNs). This particular method focuses on forecasting pedestrian activity in video frames through pedestrian identification, tracking, pose estimation, and neural networks. The joint angles outputted by the posture estimation algorithm are extracted and used as features for behavior classification. This research is discussed in [7]. Comparing to LSTM-based networks, a Transformer network utilized for elderly action identification demonstrates notable improvement. Lastly, this research presents an effective method for fall detection and recognizing senior behavior in surveillance camera settings. It examines the quantitative and qualitative performances of several networks. Further details about this method can be found in [8].

D. Dataset Survey

The state-of-the-art MPII Human Pose dataset serves as a cutting-edge benchmark, comprising more than 25,000 photos featuring over 40,000 individuals with annotated body joints. In addition to body joint annotations, the MPII dataset also provides scale annotations. Although the dataset primarily focuses on upper body images, the performance of the proposed Refinement-Correction approach, which is designed for close positions, was evaluated on the FLIC dataset due to its better results [13]. For pedestrian recognition, a unique PoseEmbedding Network is introduced, which combines visual description and human pose information. This network consists of two parts: a Region Proposal Network and a Pedestrian Recognition Network. The Region Proposal Network functions similarly to faster R-CNN, while the Pedestrian Recognition Network (PRN) consists of a Visual Feature Module, Human Pose Module, and Classification Module [14].

The UCF101 dataset is currently the largest dataset for human actions. It includes over 13,000 videos, covering 101 action categories, with a total of 27 hours of video content. The database incorporates realistic user-uploaded videos with camera movements and complex backgrounds.

The researchers also presented baseline action recognition results on this dataset using the conventional bag of words method, achieving an overall performance of 44.5%. UCF101 is considered one of the most challenging action datasets due to the large number of classes, video clips, and the unrestricted nature of the clips [15].

III. PROPOSED METHODOLOGY

The primary research methodology for this system is a literature survey and Contextual modeling. The very first step in our implementation of the proposed project is through keyframe extraction. This system takes input from the user as aquery video which is then processed for extracting frames. Thekeyframes then undergo key point embedding which is done with heatmaps and PAFs. The model uses graph convolutionalnetwork to form a skeletal frame from the key points. The retrieved co-ordinates can be further utilized for analytics.

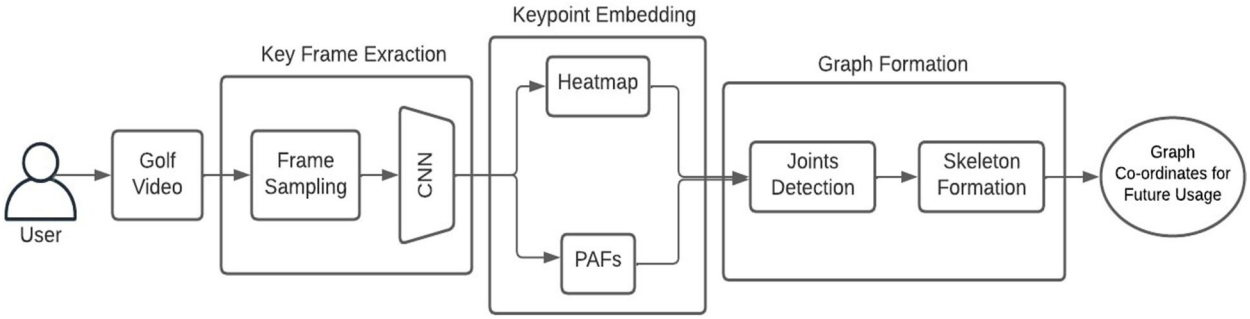


Fig. 1. System Architecture/Workflow

A. Algorithms and Methodologies

1) *VGG-19:* VGG-19 is a convolutional neural network (CNN) composed of 19 layers, including 16 convolutional layers followed by 3 fully connected layers [16]. Each convolutional layer in VGG-19 employs a 3x3 filter size, and the network uses a padding of 1 pixel to maintain the spatial resolution of the feature maps. To reduce the size of the feature maps, a max pooling layer with a 2x2 filter size is applied after every two convolutional layers. The network is trained on the ImageNet dataset, which comprises more than 1 million images spanning 1,000 object categories. In 2014, VGG-19 achieved state-of-the-art performance on the ImageNet dataset, attaining a top-5 error rate of 7.3%. Due to its deep architecture and straightforward design, VGG-19 has been widely utilized as a foundational model for various computer vision tasks, such as object recognition, image segmentation, and transfer learning. Its popularity stems from its ability to handle complex visual tasks effectively while remaining accessible for both researchers and practitioners.
2) *Part Affinity Fields for Part Association:* Part Affinity Fields (PAFs) are a convolutional neural network architecture employed for human pose estimation in images. The concept of PAFs was first introduced in the 2017 paper titled "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields" by Zhe Cao et al. PAFs are designed to predict the likelihood of connections between different body parts, such as the hand and elbow or between two elbows. The network is trained using labeled data, where each image contains annotations of body part locations and their corresponding connections. The PAF network consists of two parallel branches: one branch predicts the likelihood of each pixel belonging to a specific body part, while the other branch predicts the likelihood of connections between pairs of body parts. The output of the network is a set of PAFs, which are vector fields encoding the strength and direction of the connections between body parts. During the inference process, these estimated PAFs are utilized to construct a graphical model, where each body part is represented as a node and each connection as an edge. This graphical model is then utilized to estimate the most probable pose configuration based on the image and the PAFs. PAFs have demonstrated effectiveness in real-time estimation of multiple human poses [2]. They have been applied in various domains, including action recognition, sign language recognition, and virtual reality applications.
3) *Confidence maps for part detection:* Confidence Maps are output generated by convolutional neural networks (CNNs) specifically used for part detection, especially in the field of human pose estimation. In the context of human pose estimation, confidence maps can be visualized as heat maps that provide information about the likelihood or probability of a particular body part being present at each pixel in an image. In a CNN designed for part detection, the final layer typically consists of a collection of 2D maps, where each map corresponds to a specific body part. Each pixel within a map represents the confidence score indicating the probability of the corresponding body part being located at that particular pixel.

These confidence maps are generated by applying a softmax activation function to the output of the final layer of the CNN, which ensures that the values are normalized and represent probabilities. During the inference stage, the confidence maps are often subjected to a thresholding process. This involves applying a threshold value to the confidence scores, above which the probability is considered significant. By thresholding the confidence maps, candidate locations of each body part can be identified, providing valuable information for subsequent pose estimation or analysis tasks.

4) *Simultaneous Detection and Association:* Simultaneous detection and association is an approach employed in human pose estimation (HPE) that aims to detect body parts and associate them with the corresponding individuals in an image or video simultaneously. This approach involves a model that predicts both part-to-part associations, represented as affinity fields, and detection confidence maps. The network architecture is divided into two branches, with the bottom branch responsible for predicting affinity fields (shown in blue) and the top branch responsible for predicting confidence maps (shown in beige). The network produces two types of outputs: heatmaps and offset vectors. The heatmaps indicate the likelihood of each pixel belonging to a specific body part, providing information about the spatial distribution of the body parts. On the other hand, the offset vectors encode the relative positions of the body parts in relation to their corresponding joints, allowing for accurate localization. By simultaneously predicting both associations and detections, this approach enables efficient and also the accurate human pose estimation.

B. Datasets

1) *GolfDB Dataset:* A publicly accessible dataset of golf swing videos and the related body posture annotations is called the Golf DB dataset. Over 3,000 golf swing films with 3D body pose annotations for 18 important body joints, including the head, hands, feet, and torso, are included in the dataset. The videos, which show a variety of golf swings made by both amateur and professional players, were taken using a number of cameras. The Golf DB dataset contains annotations that are presented as 3D joint coordinates in a global coordinate system, allowing for the investigation of body mechanics and swing mechanics. Dataset metadata additionally includes player information, shot type, and shot results.

2) *COCO Dataset:* The coarse-pose proposal sub-net extracts the whole-body bounding boxes and body keypoint suggestions in a single step. The coarse-pose filtering stage, which is based on person and keypoint recommendations, may effectively eliminate improbable detections, which enhances further processing. The usefulness of the suggested architecture is demonstrated by experiments using the COCO and OCHuman datasets. The COCO keypoint challenge dataset contains 17 keypoints for each person, including 12 body elements and 5 facial characteristics. The COCO training set consists of approximately 100K human examples and over 1 million tagged keypoints. Each of the test set's "test-challenge" and "test-dev" sections include roughly 20k photographs. The COCO evaluation results, which were positive, employed the mean average accuracy (AP) as the main competition metric.

IV. CONCLUSION AND FUTURE SCOPE

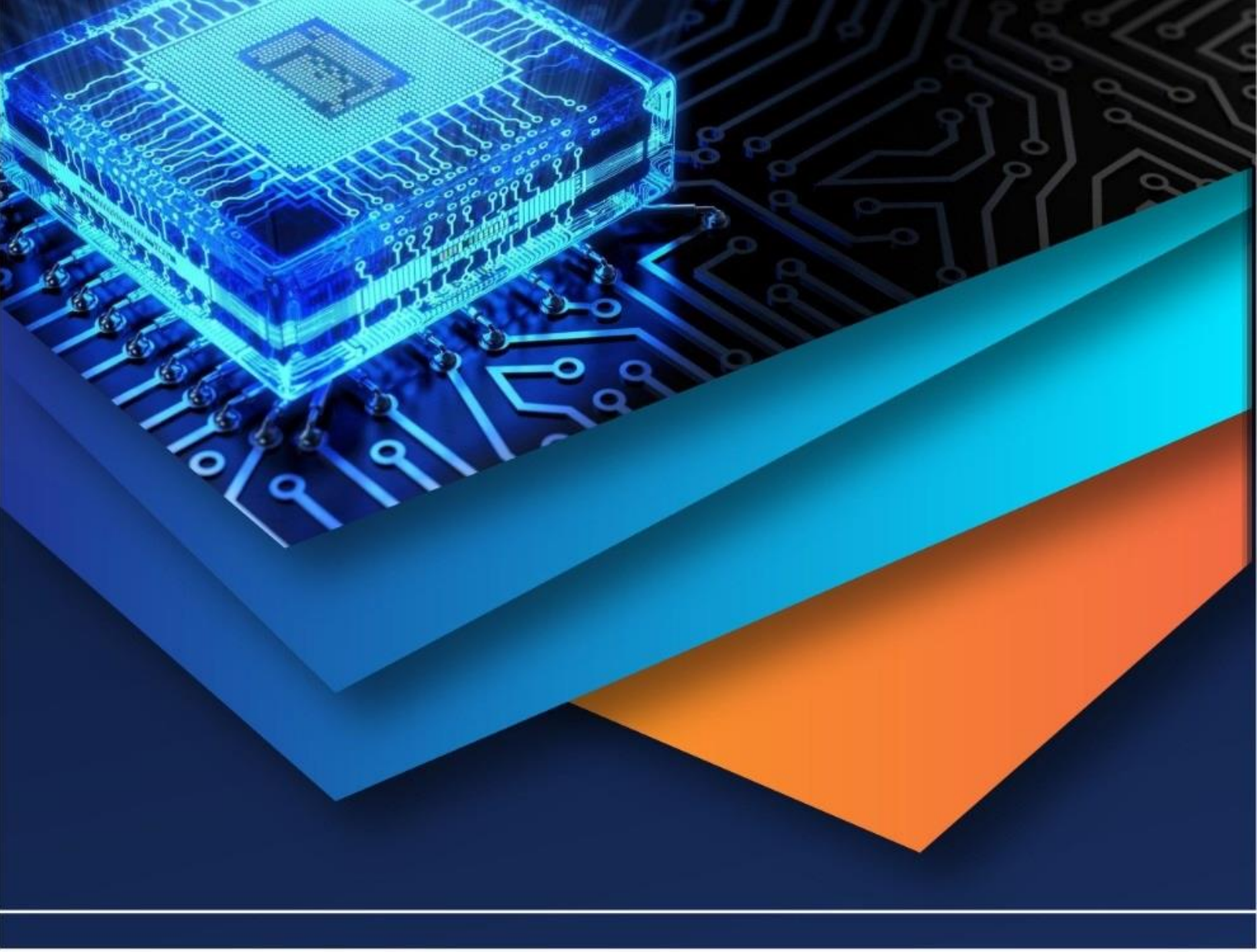
Making a reasonable quality determination based on sports video is difficult. In this document, we offer a starting point strategy for moving golfers around to evaluate their swings. We propose a neural network-based golf swing analysis tool to help consumers grasp intuitively how they differ from professionals. Key frame recognition, orientation alignment, and human position estimate are a few of the steps that make up the process. The tests demonstrate the viability of video-based analysis, but there is still much space for advancement. This will be helpful when the coaching academy is teaching sports.

This study can be used to evaluate grading results in the near future; it can be expanded by increasing the dataset in subsequent work and exploring an end-to-end deep learning-based solution. A programme that compares and visualises the variations between two input golf swing motions can be created. Users may easily distinguish between their swings and those of different specialists for user engagement during self-training. Additionally, we want to let users rapidly learn an ideal form rather than immediately duplicate an ideal motion by understanding the slow changes that take place between the two selected human positions.

REFERENCES

[1] DetPoseNet: Improving Multi-Person Pose Estimation via Coarse-Pose Filtering Lipeng Ke; Ming-Ching Chang; Honggang Qi; Siwei LyuMarch 2022, IEEE Transactions on Image Processing, Vol. 31
[2] A Local–Global Estimator Based on Large Kernel CNN and Transformerfor Human Pose Estimation and Running Pose Measurement Qingtian Wu , Member, Yongfei Wu, Yu Zhang, Liming Zhang August 2022, IEEE Transactions on Instrumentation and Measurement, Vol. 71
[3] AI Golf: Golf Swing Analysis Tool for Self-Training Chen-Chieh Liao, Dong-Hyun Hwang, Hideki Koike September 2022, IEEE Access, Vol. 10
[4] Content-Based Video Retrieval With Prototypes of Deep Features HyeokYoon, Ji-Hyeong Han March 2022, IEEE Access, Vol. 10, pp. 30730 - 30742.

[5] Deep Reinforcement Learning in a Racket Sport for Player Evaluation With Technical and Tactical Contexts Ning Ding, Kazuya Takeda, Keisuke Fujii May 2022, IEEE Access, Vol. 10, pp. 54764 - 54772.
[6] Automatic Moving Pose Grading for Golf Swing in Sports Yanting Zhang, Qing'ao Wang, Fuyu Tu, Zijian Wang October 2022, IEEEInternational Conference on Image Processing (ICIP)
[7] Detection of Violent Behavior Using Neural Networks and Pose Esti- mation Kevin B. Kwan-Loo, Jose´ C. Ort´ız-Bayliss, Santiago E. Conant-Pablos, Hugo Terashima-Mar´ın, P. Rad August 2022, IEEE Access, Vol.10
[8] Exploring Human Pose Estimation and the Usage of Synthetic Data for Elderly Fall Detection in Real-World Surveillance Sardor Juraev, Akash Ghimire, Jumabek Alikhanov, Vijay Kakani, Hakil Kim August 2022,IEEE Access, Vol. 10
[9] Optimized S2E Attention Block based Convolutional Network for Hu- man Pose Estimation Yapei Feng; Penghui Liu; Zhe-Ming Lu October 2022, IEEE Access, Vol. 10
[10] Learning Dynamical Human-Joint Affinity for 3D Pose Estimation in Videos Junhao Zhang, Yali Wang, Zhipeng Zhou, Tianyu Luan,Zhe Wang, Yu Qiao September 2021, IEEE Transactions on ImageProcessing, Vol. 30
[11] Fast and Lightweight Human Pose Estimation Haopan Ren, Wenming Wang, Kaixiang Zhang, Dejian Wei, Yanyan Gao, Yue Sun March 2021,IEEE Access, Vol. 9, pp. 49576 - 49589.
[12] HAR-Depth: A Novel Framework for Human Action Recognition UsingSequential Learning and Depth Estimated History Images Suraj PrakashSahoo; Samit Ari; Kamalakanta Mahapatra; Saraju P.Mohanty October 2021, IEEE Transactions on Emerging Topics in Computational Intelli- gence, Vol. 5, Issue 5
[13] Hybrid Refinement-Correction Heatmaps for Human Pose Estimation Aouaidjia Kamel; Bin Sheng; Ping Li; Jinman Kim; David Dagan Feng June 2021, IEEE Transactions on Multimedia, Vol. 23, pp. 1330-1342.
[14] PEN: Pose-Embedding Network for Pedestrian Detection Yifan Jiao, Hantao Yao, Changsheng Xu June 2020, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 31, Issue 3, pp. 1150-1162.
[15] Dynamic Motion Representation for Human Action Recognition Sadjad Asghari-Esfeden, Mario Sznaier, Octavia Camps March 2020, IEEE Winter Conference on Applications of Computer Vision (WACV)
[16] Multi-Person Pose Estimation With Accurate Heatmap Regression and Greedy Association. Jia Li and Meng Wang. IEEE Transactions onCircuits and systems for video technology, Vol. 32, No. 8, August



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY