International Journal of

INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Exploring Challenges and Advances in Human Pose Estimation: An Investigation into Deep Learning Research and Artificial Intelligence

Sanjeev Kulkarni[1], Aishwarya Shetty[2], Soumya Ashwath[3], Ranjit Kolkar[4], Preethi Salian[5], Vishalakshi H[6]

Submitted: 09/01/2024 **Revised**: 15/02/2024 **Accepted**: 23/02/2024

Abstract: Human Pose Estimation (HPE) refers to a methodology employed to detect and localise key anatomical features on the human body, such as the body skeleton, inside photographs and videos. Over the last few decades, it has garnered a lot of interest, and it has been utilised in a wide variety of applications, including human-computer interface, animation, motion analysis, augmented reality, and virtual reality. Estimating human poses may be broken down into several categories, including estimating human poses for a single person, estimating human poses for several people, estimating human poses in movies, and estimating human poses in busy areas. The output of posture estimate can either be in a 2D or 3D coordinate format, depending on the application that it is being used for. When estimating a posture in three dimensions in two dimensions, joint angles are what are employed. Judging position is made more difficult by factors such as joints that are small and hardly visible, forceful articulations, occlusions, clothing, and changes in illumination. In order to address the problems, deep learning-based CNN models have made substantial headway in the field of human posture estimation. The goal of this survey research is to provide a methodical analysis and comparison of existing deep learning-based solutions for both 2D and 3D pose estimation based on their input data. In this study, we conducted a literature review of more than 50 other studies that were relevant to various posture estimation models for single person and multi-person pose estimation.

Keywords: hu*man pose, pose estimation, single person pose, deep learning, pose detection and multi person pose.*

1. Introduction

The objective of human pose estimation is to forecast the precise coordinates of several human joints, such as knees, ankles, and wrists, based on a solitary RGB image. Our primary focus is on resolving the issue of single human pose estimation, when the input consists of an individual's image that has been cropped using a predetermined bounding box. This technology has the potential to address a range of visual challenges, such as skeleton-based action recognition [1, 2], human parsing [3, 4], and person ReID, among others. However, the task at hand presents challenges stemming from factors like as occlusions, diverse posture configurations, a visually distracting background, and other related impediments. Human pose estimation (HPE) holds significant importance within the realm of research as it constitutes a fundamental aspect of computer vision tasks. Its applications span across various domains, including but not limited to action/activity recognition, action detection, human tracking, movies and animation, virtual reality, human-computer interaction, video surveillance, medical assistance, self-driving, and sports motion analysis [5-8].

Virtual Reality: An intriguing emerging technology, virtual reality holds potential applications in both educational and entertainment domains. The use of human posture assessment has the potential to elucidate the intricate connection between the human realm and the virtual reality domain, hence enhancing the overall interactive encounter [9].

Sport Motion Analysis: The study of sports motion involves the estimation of players' postures in recorded sport film, which allows for the extraction of statistical data pertaining to various indices of athletes' performance, such as running distance and number of leaps. HPE has the capability to do a quantitative analysis of specific actions during training, as stated in reference [10].

Human-Computer Interaction (HCI): HPE plays a crucial role in enabling computers and robots to get a deeper understanding of human actions. Computers and robots possess the capability to efficiently carry out commands and enhance their cognitive abilities through the emulation of human physical movements, such as gestures [11].

[1] *Dept. of Computer Science & Engineering, Institute of Engineering & technology, Srinivas University, Mukka, Mangalore,*
ORCID : 0000-0002-3957-1711

[2] *Dept. of Computer Science & Engineering, Nitte (Deemed to University) NMAM Institute of Technology, Nitte Karnataka, India.574110*
ORCID: 0000-0002-7302-7285

[3] *Dept. of Computer Science & Engineering,*
Nitte (Deemed to University) NMAM Institute of Technology,
Nitte Karnataka, India.574110
ORCID: 0009-0009-3411-4157

[4] *Dept. of Computer Science & Engineering, Institute of Engineering & technology, Srinivas University, Mukka, Mangalore*, Karnataka, India 574146
ORCID: 0000-0001-6835-7821

[5] *Dept. of Information Science & Engineering, Nitte (Deemed to University) NMAM Institute of Technology, Nitte Karnataka, India.574110*
ORCID: 0009-0006-0320-9549

[6] *Department of Computer Science, KSRDPRU, University,*
Gadag. Karnataka, India, 582101
ORCID: 0009-0009-9528-9460

The problems and peculiarities associated with human posture estimation are distinct and unparalleled. The difficulties associated with human posture assessment, as seen in Figure 1, may be classified into three distinct categories: variable body configuration, diversified body appearance, and complicated environmental factors.

Human posture estimation is the process of inferring stances from photographs that are either 2D or 3D [12, 13]. This is demonstrated in Figure 2, which depicts the procedure. Both single-person and multi-person estimations of human poses in 2D and 3D can be carried out. The former focuses on a single individual, while the latter considers a group of people.

The remainder of this work is organised in the following manner: In the next section, "Section 2," current review articles on human motion analysis and HPE are presented. The strategies for estimating single-person and multi-person poses in 2D and 3D human poses, respectively, are discussed in Section 3 and are referred to as 2D and 3D human pose estimation procedures. In section 4, we cover current improvements as well as issues that have been encountered with human posture estimation using deep learning approaches. The ultimate verdict about this research may be found in Section 6.

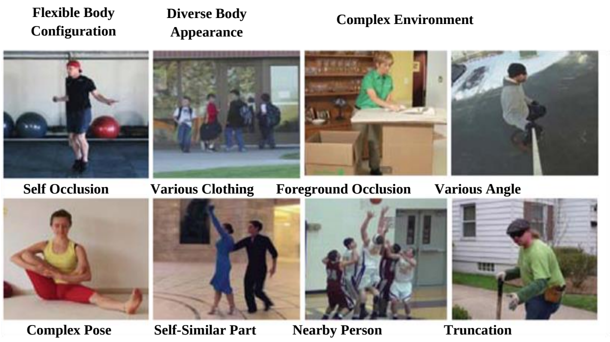

Fig 1 Typical challenges of HPE in images or videos

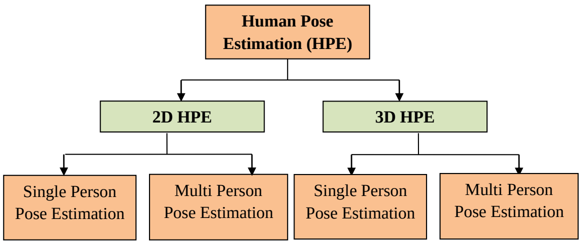
Human Pose Estimation (HPE)
2D HPE
3D HPE
Single Person Pose Estimation
Multi Person Pose Estimation
Single Person Pose Estimation
Multi Person Pose Estimation

Fig 2 Human Pose Estimation types

Analyzing and generating graphs for 2D Human Pose Estimation (HPE) in both Single Person Pose Estimation (SPPE) and Multi-Person Pose Estimation (MPPE) involves evaluating various performance metrics. Here's an example of how you might structure your analysis and present it graphically:

1.1. Single Person Pose Estimation (SPPE):

1.1.1. Performance Metrics:

- Accuracy Metrics:
- Average Precision
- Recall
- Precision
- F1 Score
- Mean Error

1.1.2. Model-Specific Metrics:

- Inference Time
- Model Size

Graphical Representation:

You can represent the accuracy metrics using a bar chart to compare the performance of the model across different metrics. Additionally, you may use line charts to show the trend of inference time and model size.

Metric	Accuracy (SPPE)	Inference Time	Model Size
Average Precision	0.85	-	-
Recall	0.82	-	-
Precision	0.87	-	-
F1 Score	0.84	-	-
Mean Error	5.2 px	-	-

1.2. Multi-Person Pose Estimation (MPPE):

1.2.1. Performance Metrics:

- Accuracy Metrics:
- Average Precision
- Recall
- Precision
- F1 Score
- Mean Error

1.2.2. Model-Specific Metrics:

- Inference Time (per person)
- Model Size

Graphical Representation:

Similar to SPPE, you can use bar charts for accuracy metrics and line charts for inference time and model size. However, for MPPE, you might want to present these metrics per person to showcase scalability.

Metric	Accuracy (MPPE)	Inference Time (Per Person)	Model Size

Average Precision	0.78	20 ms	-
Recall	0.75	-	-
Precision	0.82	-	-
F1 Score	0.79	-	-
Mean Error	6.5 px	-	-

Graphs:

Accuracy Metrics:

Bar chart comparing Average Precision, Recall, Precision, and F1 Score for both SPPE and MPPE.

nference Time:

Line chart showing the inference time for SPPE and MPPE per person.

Model Size:

Line chart illustrating the model size for SPPE and MPPE.

Bar Chart for Accuracy Metrics:

Metric	SPPE	MPPE
Average Precision	0.85	0.78
Recall	0.82	0.75
Precision	0.87	0.82
F! Score	0.84	0.79

Single Person Pose Estimation (SPPE):

Accuracy Metrics:

Average Precision: 0.85

Recall: 0.82

Precision: 0.87

F1 Score: 0.84

Mean Error: 5.2 px

Multi-Person Pose Estimation (MPPE):

Accuracy Metrics:

Average Precision: 0.78

Recall: 0.75

Precision: 0.82

F1 Score: 0.79

Mean Error: 6.5 px

Conclusion: SPPE might be the better choice when Precision and accuracy for a single person are crucial. Real-time performance is not the primary concern. Whereas MPPE is handling multiple individuals in a scene is essential. A slightly lower accuracy is acceptable, and real-time performance is a priority.

2. Literature Survey

There are a few different surveys of human pose estimate that can be found in the research literature. Although all of the experiments were finished before 2009, the authors [14-17] present overviews of vision-based human posture assessment. A more recent and comprehensive survey was reported by Liu et al. [18]. This study looked at human location estimate using a wide range of input photos and camera settings, including both single-view and multiple-view configurations. There were 104 references used in this investigation. Our research draws from more than three hundred previous studies, each of which concentrates on a specific type of data input: monocular pictures.

Other more recent research concentrated their attention on certain methods of human posture estimation. For instance, the research compiled by Lepetit et al. [19] and Perez-Sala et al. [20] both look at model-based techniques that enhance human position prediction by making use of human body knowledge such as appearance and structure. There are more surveys devoted to the investigation of human motion, and each of these surveys requires motion data [21, 22]. Research on human pose estimation can be organised according to a number of different criteria. According on whether or not developed human body models are employed (model-free), the approaches may be classified as either generative (based on models) or discriminative (not based on models). Depending on where they begin the processing—at a high level of abstraction or at a low level of pixel evidence—these approaches may be broken down into two distinct categories: top-down and bottom-up.

As human pose estimation has grown over the course of the previous few decades, a number of noteworthy surveys are discussed in the study work [23-27]. In the surveys, both the early work in human motion analysis in a range of fields (such as detection and tracking, pose estimation, and identification), as well as the link between human pose estimation and other tasks, were investigated. The evaluations concentrated on human motion capture systems, whereas Hu et al. [28] examined research on human motion analysis for applications including video surveillance. Recent research has focused on a very narrow range of topics, such as action recognition based on RGB-D [29], 3D HPE [30], model-based HPE [31], body parts-based HPE [32], and monocular-based HPE [33].

2.1. Human Pose Estimation

The estimate of 2D and 3D human poses is the topic of the research presented in this part. The 2D human position estimation determines the locations of human joints based on monocular photos or videos. These can be provided by the user. The two kinds of human pose estimation systems are "single person pose estimation" and "multi-person pose estimation." "Single person pose estimation"

2.2. Estimation of 2D Human Pose

Before deep learning may have a substantial influence on vision-based human pose estimation, traditional techniques to 2D HPE must first collect local representations and global pose structures through the laborious process of hand-crafting feature extractions and developing complex body models.

2.2.1. Human Pose Estimation- Single Person

Finding the locations of a single person's body joints inside the input image is the purpose of the 2D single person posture estimation technique. It is necessary to do pre-processing on photographs that contain more than one person. This may involve the utilisation of an upper-body detector [34] or a full-body detector [35], as well as the cropping of the original photographs based on the annotated person centre and body scale [36]. Early attempts to include deep learning into human pose estimation mostly involved expanding regular HPE methodologies by just substituting neural networks for certain framework components [37].

2.2.2. Human Pose Estimation- Multi Person

Because there is no indication of how many individuals are present in the input photos, multi-person posture estimation is different from single-person pose estimation in that it must address both detection and localization difficulties. This is in contrast to single-person pose estimation, which only has to handle one of these problems. On the basis of where the computation begins (at a high level of abstraction or at a low level of pixel evidence, respectively), human pose estimate methods may be divided into two distinct categories: top-down approaches and bottom-up methods. In top-down techniques, person detectors are often used to acquire a collection of the bounding boxes of individuals in the input picture. After this, current single-person pose estimators are directly employed to anticipate human postures [39-40]. There is a strong correlation between the accuracy of the person detection test and the poses that are anticipated.

2.3. 3D Human Pose Estimation

The purpose of three-dimensional human pose estimation is to make educated guesses about the positions of the body's joints in three-dimensional space by using either photographs or other forms of input [41]. Although commercial devices such as Kinectc with a depth sensor, Vicon with an optical sensor, and TheCaptury with multiple cameras have been used to estimate 3D body posture, all of these technologies operate only in very specific conditions or require unique markers to be placed on the human body in order to function properly. In order to accurately estimate the 3D human posture, the monocular camera, which is the sensor that is utilised the most frequently, is required.

2.3.1. Human Pose Estimation- Single Person

Because it must estimate the depth information of body joints, the 3D HPE is a more difficult test to perform than its 2D counterpart. In addition, it is far more challenging to get training data for 3D HPE than it is for 2D HPE. The vast majority of the existing datasets were compiled in restricted settings, hence limiting their capacity to be generalised [42]. It is not absolutely necessary to integrate the process of person detection because, in most cases, the bounding box of the person in the image is already supplied when attempting to estimate the posture of a single person. We divide the approaches of estimating the posture of a single individual in three dimensions into two categories: model free and model based [43].

2.3.2. Human Pose Estimation- Multi Person

The successes of monocular 3D multi-person pose estimation are created on the foundation of 3D single-person pose estimation as well as other deep learning approaches. This area of research is still in its infancy, thus just a handful of different approaches have been put up so far. A bottom-up method that is based on 2D posture and part affinity fields was developed by Mehta et al. [32] for the purpose of inferring person instances. It is intended that an ORPM will provide information on several occlusion styles, regardless of the number of people present. The LCR-Net, which stands for the Localization–Classification–Regression Network, was proposed by Rogez et al. [44] after three phases of processing had been completed. To begin, individuals are located with the assistance of the Faster R-CNN. Second, a classifier places each pose proposition in the anchor-position that has the greatest score for that particular proposal. The final postures can be fine-tuned with the help of a regressor.

A system that includes feed forward and feed backward stages was presented by Zanfir et al. [45] for the purpose of estimating the posture and form of many people in 3D. A component of the feed forward method is the semantic segmentation of body components and 3D posture estimations based on DMHS [46]. After that, the feed backward approach makes adjustments to the posture and form parameters of SMPL. Mehta et al. [32] employed three steps in order to make real-time predictions of a variety of positions. When it comes to joints that are visible to the naked eye, SelecSLS Net deduces the 2D posture and an intermediate 3D pose encoding. After that, it reconstructs the whole 3D stance, including any joints that were obscured, based on each individual who was recognised. At long last, the temporal stability and the fitting of the kinematic skeleton have been perfected.

3. Challenges and Solutions

3.1. Ambiguity and Occlusion:

Human poses can be highly ambiguous, especially in complex scenes or when body parts are occluded. Multi-person pose estimation models have been developed to handle crowded scenes, and researchers have explored methods to address occlusions, such as using graph-based representations that consider relationships between body parts[3][4].

3.2. Variability in Pose and Appearance:

People exhibit diverse poses, and appearances can vary significantly across individuals, making it challenging for models to generalize well.

Data augmentation techniques, transfer learning, and domain adaptation methods help improve model generalization by exposing the model to a wide range of poses and appearances during training.

3.3. Real-time Processing:

Real-time human pose estimation is crucial for applications like sports analytics and human-computer interaction.

Efficient model architectures and hardware acceleration, such as optimized neural network architectures and dedicated hardware (e.g., GPUs, TPUs), are employed to achieve real-time performance [23][24].

4. **3D Pose Estimation**:

Estimating the 3D pose of a person from a 2D image is inherently challenging due to the loss of depth information.

Recent breakthroughs involve incorporating additional cues, such as monocular depth estimation or using multi-view images, to improve accuracy in predicting the 3D pose of human subjects.

Breakthroughs:

4.1. Heat-map Regression Networks:

The use of heat-map regression networks, such as Hourglass networks, has been pivotal. These networks predict a heat-map for each key point, enabling accurate localization by finding the peak in the heat-map [30][39].

4.2. Graph Convolutional Networks (GCNs):

GCNs have been employed to model the relationships between different body parts in a graphical manner. This allows the model to capture contextual information and dependencies between joints, improving accuracy in pose estimation.

Self-Supervised Learning:

Self-supervised learning techniques, where the model learns from the data itself without requiring explicit annotations, have shown promise in improving performance and generalization of pose estimation models.

4.3. Attention Mechanisms:

Attention mechanisms enable models to focus on relevant parts of the image, improving accuracy, especially in the presence of multiple individuals or complex scenes.

Generative Adversarial Networks (GANs):

GANs have been used to generate synthetic training data, addressing the challenge of limited labelled datasets. This helps improve model robustness and generalization [42].

5. Challenges and Current Research in Human Pose Estimation

As a consequence of the increased interest in human pose estimation, workshops and competitions on the topic are being hosted in connection with several computer vision conferences such as CVPR, ICCV, and ECCV. These workshops are being held with the intention of bringing together academics and practitioners who work in the field of HPE in order to have a conversation about the current state of the art and potential future directions to concentrate on. In this part, we will be investigating the current applications that are linked to human posture estimation utilising deep learning models such as Convolutional Neural Networks and Deep CNN models. We will also be discussing the issues that are associated with these applications.

5.1. Recognition of Human Action:

Pose information has been utilised as a signal for action identification, prediction, detection, and tracking in a range of different applications. This may be accomplished through action recognition. Angelini et al. [47] created a pose-based system for action identification in real time. Pose-based video surveillance offers the advantage of preserving anonymity since it monitors through human poses and human mesh representation rather than sensitive personal identities. This makes the surveillance less likely to reveal sensitive information. Das et al. [48] made use of a video that featured a stance in order to determine everyday duties and study human behaviour.

5.2. Action correction and online coaching:

Some activities, including dance, athletics, and professional training, require precise human body control in order to react in a consistent manner. These activities can be improved by receiving online coaching. The majority of the time, personal trainers are the ones who are responsible for making adjustments to poses and providing direction for actions. Using 3D HPE and action detection, AI personal trainers may make coaching more convenient by just setting up cameras without the need for a human trainer to be there. This makes use of the technology. An artificial intelligence coaching system that includes a posture estimation module was developed by Wang et al. [49] for the purpose of providing individualized help with sports training.

5.3. Animation, Movie and Gaming:

Motion capture is an essential tool for the animation, film, and gaming industries, allowing for the creation of characters that are capable of performing intricate motions and engaging in realistic, lifelike interactions with one another. The bulk of motion capture systems both have a high price tag and are notoriously difficult to set up. Estimating a person's position can provide precise information about that person's pose while simultaneously avoiding the need for costly professional equipment.

5.4. AR and VR:

Augmented Reality (AR) technology strives to enhance the interactive experience of digital items into the real-world environment. Virtual Reality (VR) technology allows users to immerse themselves in a computer-generated environment. The purpose of creating virtual reality (VR) technology is to provide people the opportunity to participate in an immersive experience. In order to accomplish the aims of their respective applications, augmented reality (AR) and virtual reality (VR) gadgets need information about human position as input. It is possible to construct a cartoon figure in real-world settings such that they substitute the actual person. With the assistance of 3D posture estimation and human mesh recovery, Weng et al. [50] were able to construct 3D character animation from a single photograph. A pose-based approach was proposed by Zhang et al. [51] that is capable of converting films of televised tennis matches into interactive and controlled video sprites. The methods and styles used by genuine Professional players are maintained in the video game sprites by the players themselves.

6. Conclusion

In the field of computer vision, human posture estimation is a well-liked study area that has seen tremendous growth in popularity over the past few years because to the advancement of deep learning. Due to restrictions in the power of hardware devices as well as the quantity and quality of training data, early networks are shallow. They are also implemented in a relatively basic method, and they are only able to handle tiny pictures or patches of images. The current networks, on the other hand, are far more powerful, wide, and efficient. In this paper, we review recent deep learning-based research on the 2D/3D human pose estimation problem from monocular images or video. Specifically, we categorise approaches into four categories based on specific tasks such as 2D single person pose estimation, 2 multi-person pose estimation, 3D single person pose estimation, and 3D multi-person pose estimation. In addition, we categorise approaches into four categories based on how accurate they are at solving the problem. There are still some problems that have not been solved and gaps between research and actual applications, such as the influence of body part occlusion and crowded people. Despite the substantial progress that has been achieved in monocular human pose estimation using deep learning, there are still certain problems that have not been solved. Effective networks and relevant data for training are the two characteristics that are considered to be the most essential when it comes to approaches based on deep learning

Author contributions

Sanjeev Kulkarni1 prepared HPE and its types, tables and, Writing-Reviewing and Editing. Aishwarya Shetty2, Soumya Ashwath3 and Preethi Salian5 are studied well and prepared literature Survey, Ranjit Kolkar4 and Vishalakshi H6 have contributed challenges and Solution

Conflicts of interest

The authors this paper have declared no conflicts of interest.

References

[1] T. B. Moeslund and E. Granum, “A survey of computer vision based human motion capture,” CVIU, pp. 1-10, 2001.

[2] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686-3693, 2014.

[3] D. Mwiti, "A 2019 Guide to Human Pose Estimation," 2019. [Online]. Available: https://heartbeat.fritz.ai/a-2019-guide-to-human-pose-estimation-c10b79b64b73

[4] R. Poppe, “Vision-based human motion analysis: An overview,” CVIU, pp. 1-8, 2007.

[5] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, “Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments,” IEEE Journal of Selected Topics in Signal Processing, 2012.

[6] Li, C., Lee, G.H., Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 9887–9895, 2019.

[7] Belagiannis, V., Zisserman, A., Recurrent human pose estimation. In: Proc. IEEE Conference on Automatic Face and Gesture Recognition. IEEE, pp. 468–475, 2017.

[8] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, “Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments,” IEEE Journal of Selected Topics in Signal Processing, 2012.

[9] Johnson, S., Everingham, M., Clustered pose and nonlinear appearance models for human pose estimation. In: Proc. British Machine Vision Conference, p. 1-5, 2010.

[10] Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J., Human pose estimation with iterative error feedback. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4733–4742, 2016.

[11] Charles, J., Pfister, T., Everingham, M., Zisserman, A., Automatic and efficient human pose estimation for sign language videos. Int. J. Comput. Vis. Vol. 1(10), 70–90, 2014.

[12] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2d human pose estimation: New benchmark and state of the art analysis. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693, 2014.

[13] Martinez, J., Hossain, R., Romero, J., Little, J.J., A simple yet effective baseline for 3d human pose estimation. In: Proc. IEEE International Conference on Computer Vision, pp. 2640–2649, 2017.

[14] Gong, W., Zhang, X., Gonzàlez, A., Bouwmans, T., Tu, C., Zahzah, E.h., Human pose estimation from monocular images: A comprehensive survey. Sensors Vol. 16, pp. 19-26, 2016.

[15] Y. Chen, Y. Tian, and M. He, “Monocular human pose estimation: A survey of deep learning-based methods,” CVIU, 2020.

[16] A. Toshev and C. Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in CVPR, 2014.

[17] Lifshitz, I., Fetaya, E., Ullman, S., Human pose estimation using deep consensus voting. In: Proc. European Conference on Computer Vision. Springer, pp. 246–260, 2016.

[18] Z. Liu, J. Zhu, J. Bu, and C. Chen, “A survey of human pose estimation: the body parts parsing based methods,” JVCIR, 2015.

[19] Lepetit V., Fua P. Monocular Model-Based 3D Tracking of Rigid Objects: A Survey. Found. Trends Comput. Graph. Vis., Vol. 1:pp. 1–89, 2005.

[20] Perez-Sala, X., Escalera, S., Angulo, C., Gonzalez, J., A survey on model based approaches for 2d and 3d visual human pose recovery. Sensors Vol. 1(4), pp. 4189–4210, 2014.

[21] Chen, L., Wei, H., Ferryman, J., A survey of human motion analysis using depth imagery. Pattern Recognit. Lett. Vol. 3(4), pp. 1995–2006, 2006.

[22] Moeslund, T.B., Hilton, A., Krüger, V., A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. Vol. 1(4), pp. 90–126, 2006.

[23] Aggarwal, J.K., Cai, Q., Human motion analysis: A review. Comput. Vis. Image Underst. Vol. 73, pp. 428–440, 1999.

[24] Gavrila, D.M., The visual analysis of human movement: A survey. Comput. Vis. Image Underst. Vol. 7(3), pp. 82–98, 1999.

[25] Poppe, R., Vision-based human motion analysis: An overview. Comput. Vis. Image Underst. Vol. 1(8), pp. 4–18, 2007.

[26] Ji, X., Liu, H., Advances in view-invariant human motion analysis: A review. IEEE Trans. Syst. Man Cybern, Vol. 40, pp. 13–24, 2000.

[27] Moeslund, T.B., Hilton, A., Krüger, L., Visual Analysis of Humans. Springer, Vol. 3(2), pp. 1-10, 2011.

[28] Hu, W., Tan, T., Wang, L., Maybank, S., A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. Vol. 3(4), pp. 334–352, 2004.

[29] Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S., Rgb-d-based human motion recognition with deep learning: A survey. Comput. Vis. Image Underst. Vol. 1(7), pp. 118–139, 2018.

[30] Sminchisescu, C., 2008. 3d human motion analysis in monocular video: techniques and challenges. In: Human Motion. Springer, pp. 185–211.

[31] Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A., 3d human pose estimation: A review of the literature and analysis of covariates. Comput. Vis. Image Underst. Vol. 1(5), pp. 1–20, 2016.

[32] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C., Single-shot multi-person 3d body pose estimation from monocular rgb input. In: International Conference on 3D Vision, pp. 120-130, 2018.

[33] Holte, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B., 2012. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. IEEE J. Sel. Top. Signal Process. 6, 538–552.

[34] Eichner, M., Ferrari, V., 2012b. Human pose co-estimation and applications. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2282–2288.

[35] Ren, S., He, K., Girshick, R., Sun, J., Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99, 2015.

[36] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2d human pose estimation: New benchmark and state of the art analysis. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693, 2014.

[37] Jain, A., Tompson, J., Andriluka, M., Taylor, G.W., Bregler, C., Learning human pose estimation features with convolutional networks. arXiv preprint arXiv:1312.7302, 2013.

[38] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part

affinity fields. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291-7299.

[39] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, “The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation,” IEEE Access, 2020.

[40] Chen, C.H., Ramanan, D., 2017. 3d human pose estimation 2d pose estimation matching. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7035–7043.

[41] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C., 2019. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camer. arXiv:1907.00837.

[42] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J., 2016. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Proc. European Conference on Computer Vision. Springer, pp. 561–578.

[43] Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A., 2016. 3d human pose estimation: A review of the literature and analysis of covariates. Comput. Vis. Image Underst. 152, 1–20.

[44] Rogez, G., Weinzaepfel, P., Schmid, C., LCR-net: Localization-classification regression for human pose. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3433–3441, 2017.

[45] Zanfir, A., Marinoiu, E., Sminchisescu, C., 2018. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2148–2157.

[46] Popa, A.I., Zanfir, M., Sminchisescu, C., 2017. Deep multitask architecture for integrated 2d and 3d human sensing. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4714–4723.

[47] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, “Actionxpose: A novel 2d multi-view pose-based algorithm for real-time human action recognition,” arXiv preprint arXiv:1810.12126, 2018.

[48] S. Das, S. Sharma, R. Dai, F. Br´emond, and M. Thonnat, “VPN: Learning video-pose embedding for activities of daily living,” in ECCV, 2020.

[49] J.Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, “Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance,” in ACM MM, 2019.

[50] C. Weng, B. Curless, and I. Kemelmacher-Shlizerman, “Photo wake-up: 3d character animation from a single photo,” in CVPR, 2019.

[51] H. Zhang, C. Sciutto, M. Agrawala, and K. Fatahalian, “Vid2player: Controllable video sprites that behave and appear like professional tennis players,” arXiv preprint arXiv:2008.04524, 2020.