Human Pose Estimation for Training Assistance: a Systematic Literature Review

Gisela Miranda Difini
University of Vale do Rio dos Sinos
São Leopoldo, RS, Brazil
gisela.mdi@gmail.com

Marcio Garcia Martins
University of Vale do Rio dos Sinos
São Leopoldo, RS, Brazil
marciog@unisinos.br

Jorge Luis Victória Barbosa
University of Vale do Rio dos Sinos
São Leopoldo, RS, Brazil
jbarbosa@unisinos.br

ABSTRACT

Human pose estimation is an important field of Computer Vision that aims to predict poses of individuals from videos and images. It has been used in many different areas including human-computer interaction, motion analysis, surveillance, action prediction, action correction, augmented reality, virtual reality, and healthcare. This review is focused on the most significant contributions in human pose estimation for training assistance. Executing movements correctly is crucial in all kinds of physical activities, both to increase performance and reduce risk of injury. Human pose estimation is poised to help athletes better analyse the quality of their movements. The systematic review study was conducted in five databases including articles from January 2011 to March 2021. The initial search resulted in 129 articles, of which 8 were selected after applying the filtering criteria. Moreover this study presents the challenges related to pose estimation, which pose estimation methods have been used in recent years, in which specific activities the selected articles have focused on, and a taxonomy of human pose estimation methods.

CCS CONCEPTS

• **Computing methodologies** → *Activity recognition and understanding*; *Motion capture.*

KEYWORDS

human pose estimation, pose tracking, training assistance, sports, systematic review

ACM Reference Format:
Gisela Miranda Difini, Marcio Garcia Martins, and Jorge Luis Victória Barbosa. 2021. Human Pose Estimation for Training Assistance: a Systematic Literature Review. In *Brazilian Symposium on Multimedia and the Web (WebMedia '21), November 5–12, 2021, Belo Horizonte / Minas Gerais, Brazil.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3470482.3479633

1 INTRODUCTION

Human pose estimation (HPE) is a field of Computer Vision that aims to predict the poses of human bodies by extracting joints from images and videos. Pose estimation has a wide range of applications including human-computer interaction, motion analysis, surveillance, action prediction, action correction, augmented reality, virtual reality, and healthcare. The goal of this review is to focus on HPE applied to sports and training assistance. Training assistance is related to online coaching, action correction and could also be used with augmented reality.

HPE methods can be divided into two main categories: two-dimensional (2D) and three-dimensional (3D). The HPE is defined as single-person or multi-person based on the number of people in the image. The single-person pipeline can either be a heatmap-based method or a regression method depending on the way it predicts the human body keypoints. In the multi-person method the position and number of people are unknown and can be classified into top-down or bottom-up approaches. Top-down methods first detect the people and then utilize single-person HPE to estimate the keypoints for each person. Alternatively, bottom-up methods first detect body keypoints without knowing the number of people in an image and then group these keypoints into individual poses.

3D HPE approaches are first classified into two types of input: images and videos or sensors (eg., depth camera). Images and videos can have one or multiple viewpoints, thus it is divided into single-view and multi-view methods respectively. Single-view methods can be classified into single-person and multi-person. The single-person approaches are defined as generative and discriminative [23][19]. The generative method, also referred to as model-based method, uses an a priori model in pose estimation as a reference. The discriminative or model-free method learns a mapping function between image or depth observations and 3D human body poses. Discriminative approaches can be classified as learning-based or example-based. Learning-based methods learn a mapping function from image space to the pose space. Example-based methods store a collection of exemplars in different poses and orientations and estimates the final 3D pose by interpolating the candidates obtained from a similarity search.

In sports, the training quality depends on the movements to be executed precisely to successfully achieve the desired result and decrease risk of injury, thus needing an accurate analysis of the athlete's pose. The use of video to review training and competition performance has been commonplace among athletes and coaches. Now, more than ever due to the COVID-19 pandemic, instructors have started to carry out online coaching [27] [8]. Without any assistive technology, athletes and coaches need to play and replay the video a few times to analyze detailed movements. It's challenging and time consuming for users to analyse massive amounts of videos and visualize training results. HPE can be used to help with sports motion analysis and action correction, reducing the workload of analysing movements from videos.

WebMedia '21, November 5–12, 2021, Belo Horizonte / Minas Gerais, Brazil

ACM ISBN 978-1-4503-8609-8/21/11…$15.00
https://doi.org/10.1145/3470482.3479633

In recent years, new pose estimation techniques have produced good results in detecting keypoints from 2D images using deep neural networks [3], thus helping athletes to analyse their performance from videos or single-view images.

The most recent surveys and reviews focus on deep learning-based human pose estimation. Zheng et al. [28] give an overview of recent advances in HPE and provide a research on deep learning-based 2D and 3D HPE methods. Dang et al. [4] and Liu et al. [17] provide an overview of state-of-the-art deep learning-based 2D HPE methodologies. Khan et al. [13] review different methods of HPE from single image. However, none of them focus on HPE applications related to training assistance (eg., online coaching, action correction and augmented reality).

This article presents a systematic review of articles related to sports using pose estimation that aims to understand what are the challenges of using human pose estimation for training assistance, which technologies have been used in recent years, and in which specific sports the selected articles are focused on. The research was conducted in five databases including articles from January 2011 to March 2021. The initial search resulted in 129 articles, of which 8 were selected after applying the filtering criteria.

The organization of this article has the following structure: Section 1 gives an overview of the current topic followed by Section 2 that presents the systematic literature review methodology, the definition of research questions, the search process, the selection and filtering process, and the quality assessment of the articles. Section 3 presents the results and discussions regarding the analysis of the initial search. We answer each research question in Section 4 and present conclusions in Section 5.

2 METHODOLOGY

This article follows the systematic literature review (SLR) process proposed by Kitchenham et al. [14] and we have applied the following methodological steps: (1) Research questions definition; (2) Search process; (3) Criteria for filtering the results; (4) Quality assessment.

2.1 Research questions

The research questions guide the search process of relevant articles that aim to answer how human pose estimation can help with training assistance. The elaboration process of the research questions involved a preliminary research of articles related to the field of interest and analysis of the resulting articles. The following questions were defined:

- **SQ1**: What are the challenges of using human pose estimation for training assistance?
- **SQ2**: Which technologies are being used for the human pose estimation?
- **SQ3**: In which context is human pose estimation being used for?
- **SQ4**: What is the pose estimation accuracy for the applied training context?

The purpose of SQ1 is to understand the technical challenges that the researchers encountered when proposing solutions for training assistance using HPE. SQ2 allows to identify what are the algorithms and frameworks used in HPE scenarios. SQ3 aims to understand in which activities the researchers focused on. And finally, SQ4 helps to understand the accuracy for the scenario where HPE has been applied.

2.2 Search process

The first step of this process is the creation of the search string. Boolean operations are applied into the search string to get more accurate results. The boolean expression was divided into three sets of interest. These sets contemplate terms that are synonymous with the keywords already defined. The following search string was defined:

("human pose estimation" OR "human pose tracking") AND ("exercises" OR "sports") AND ("assistance" OR "correction" OR "guidance")

The selected databases were ACM Digital Library[1], IEEE[2], Science Direct[3], Springer[4], and Wiley Online Library[5].

2.3 Article selection process

After collecting the articles from the selected databases, it is necessary to define the criteria for filtering the results to remove all research that were considered invalid for the purpose of this review. The following exclusion criteria (EC) were applied:

- **EC1**: articles published before 2011.
- **EC2**: duplicated articles.
- **EC3**: articles not directly related to pose estimation.
- **EC4**: articles not directly related to training assistance.
- **EC5**: articles that presented results of surveys or reviews.

As part of the methodology [14], we removed articles published before 2011. Surveys and reviews were used for discussion of related works in Section 1 and removed from the selection process. We also removed duplicated articles that appeared in more than one database. Besides, we removed any work with no scientific character, such as blog posts and magazine articles. After the completion of these steps, named as Impurities Removal on Figure 1, we analysed the remaining articles by its title, abstract, and full text. After that, we applied the EC and then proceeded to the phase of qualitative evaluation presented in the following section.

2.4 Quality assessment

In this final step, the evaluation of the remaining articles by quality assessment is based on the following questions:

- **QA1**: Is there an architecture/framework proposal?
- **QA2**: Does the article present challenges of human pose estimation?
- **QA3**: Does the article present a pose estimation accuracy?
- **QA4**: Does the article present comparisons and results of experiments?

The QA questions are scored based on the following criteria:

[1] https://dl.acm.org/
[2] http://ieeexplore.ieee.org/
[3] https://www.sciencedirect.com/
[4] https://link.springer.com/
[5] https://onlinelibrary.wiley.com/

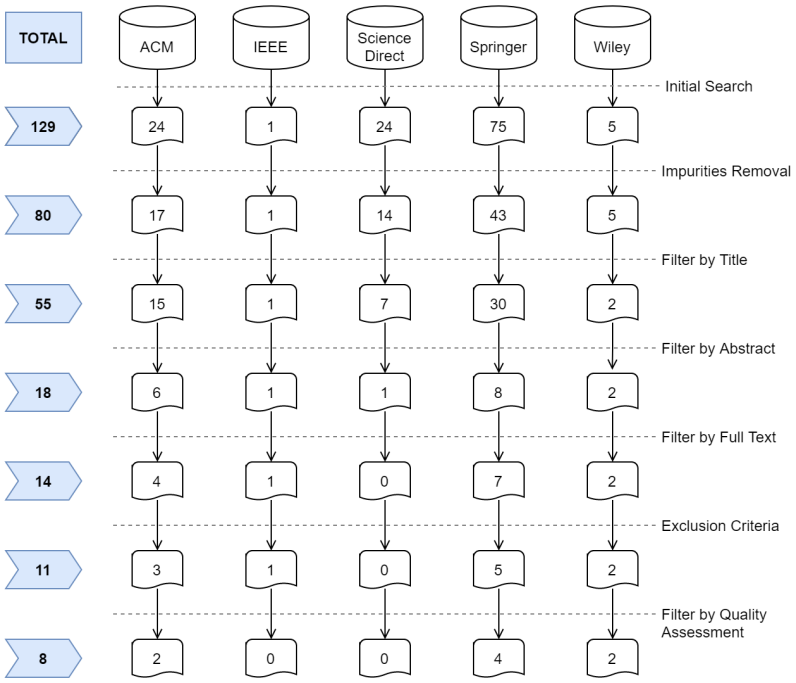

Figure 1: Filtering process of selected articles by database.

- QA1: Y (yes), the architecture/framework is described as well as the human pose estimation algorithm, P (partial), the architecture/framework is not fully explained, N (no) there is no architecture or framework presented.
- QA2: Y, the challenges of the applied pose estimation algorithm are explicitly defined and explained, P, the challenges are implicit or the article mentions challenges of other algorithms for pose estimation, N, if the challenges are not defined.
- QA3: Y, the article explicitly presents the accuracy of human pose estimation, P, the article presents implicit results, N, the article does not present pose estimation results.
- QA4: Y, the article compares the pose estimation result with more than one algorithm/dataset and presents results of experiments/trained network, P, the article presents comparisons with at least one algorithm/dataset or results of experiments/trained network, N, the article does not presents comparisons nor results of experiments/trained network.

Based on the Kitchenham et al. [14] methodology, Table 1 defines three possible answers, each one receiving a grade: Yes = 1, Partial = 0.5, and No = 0. After one of the researchers graded the articles, two researchers discussed the results. Articles that were graded below 2.5 by the researchers were excluded from the corpus since at least 2 of the questions received "no" or "partial" as the answer, indicating a poor reference for this SLR.

After applying the mentioned criteria onto the original set of articles, we read the remaining ones to answer the research question. We discuss the results in Section 3.

Table 1: Answers and grades.

Answer	Description	Grade
Yes	The article explicitly answers the question	1.0
Partial	The article answers part of the question	0.5
No	The article does not mention the topic	0.0

3 SEARCH RESULTS

This section discusses the results of the search process, the selection process and the qualitative analysis of the selected articles. Figure 1 presents the results of the filtering process and the description of each step and the number of remaining articles.

The details of the steps taken based on the SLR methodology have the following order: Section 3.1 discusses relevant articles that are applied to the context of human pose estimation for training assistance but do not meet all the criteria to be part of this SLR. After analyzing the exclusion criteria, details on the quality assessment of the articles are presented in Section 3.2.

3.1 Exclusion of articles from the initial search

Figure 1 shows the number of articles obtained in each database selected in the Initial Search stage with a total of 129 articles. The following stage is called Impurities Removal where duplicated articles, surveys, reviews, and non-scientific work were removed, resulting in a total of 80 articles. In the next step, articles were filtered by title, resulting in 55 articles. Moreover, the remaining articles were filtered by reading by reading their abstracts and removing articles not related to this field of study, resulting in a total of 18 articles.

If the abstract mentioned something related to pose estimation applied to any kind of sport, then the article was moved to the next step for a filter by full text in which resulted in 14 articles. The next step applied the Exclusion Criteria mentioned in Section 2.3, where articles based on the exclusion criteria previously defined were removed. The number of articles considered relevant to this review reduced down to 11 in this phase.

We removed articles that despite focusing on training assistance with video analysis, it did not apply human pose estimation. For example, Kasiri et al. [12] used a fuzzy inference method based on 2D Chamfer distance, depth values, and geodesic distance for detecting boxer body parts. The authors used both a multi-class SVM classifier and Random Forest to classify different punch types.

Jain et al. [10] proposed a framework for analyzing and issuing reports of action segments that were missed or anomalously performed by applying Approximate String Matching technique, the approach included a pose estimation but that was not the main focus of their work.

There were articles that focused mainly on HPE algorithms, but did not have training assistance as the main topic of the study. For example, Mehta et al. [18] presented a real-time method to capture 3D skeletal pose of a human using an RGB camera along with convolutional neural network (CNN) pose regression to obtain joint positions in 2D image space and 3D and a kinematic skeleton fitting against the 2D/3D pose predictions to produce a temporally stable joint angles of a metric global 3D skeleton. We also removed the article from Shamsolmoali et al. [21] that proposes a generative adversarial network (GAN) as the learning model containing two residual Multiple-Instance Learning (MIL) models with identical architecture, one is used as the generator, and the other one is used as the discriminator. It has been validated on two datasets for the human pose estimation task and successfully outperforms the other state-of-the-art models.

3.2 Performing quality assessment to select relevant articles

This section follows the criteria defined in Section 2.4 to guide in the qualitative analysis of the selected articles that have passed in the impurities removal, filter by title, filter by abstract, filter by full text as well as exclusion criteria phase. We answered each question following the Kitchenham's methodology [14]. The possible answers are presented in Table 1. We considered relevant articles, those who have received at least 2.5 as the final grade. The answers to the questions and the resulting scores are presented in Table 2 which is organized in a descending order by the article's score followed by the year of publication and finally the authors' name ordered alphabetically. The articles that scored less than 2.5 are also presented in Table 2.

After performing the quality assessment over 11 articles, 3 articles [7][29][24] that did not reach the desired score were removed.

The research from Tharatipyakul et al.[24] proposed a system that uses human pose estimation to provide visual feedback to users who want to follow tutorial videos. This article was removed based on the quality assessment since it did not explicitly explain the architecture and the pose estimation algorithm being used and its accuracy.

The remaining articles are classified by year of publication and database in Figure 2. The x-axis represents the range of year in an ascending order and on each year we have the articles references.

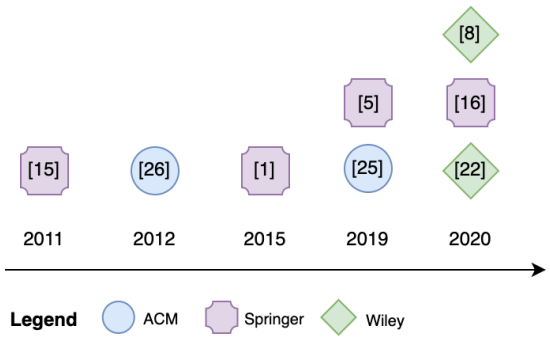

Figure 2: Publication year of the selected articles by database.

Figure 3 shows the articles according to their scores. The x-axis represents the scores in an ascending order and on each score we have the articles references. The results show that 3 articles answered all the quality assessments questions. The articles that scored less than 2.5 were removed from the selected articles.

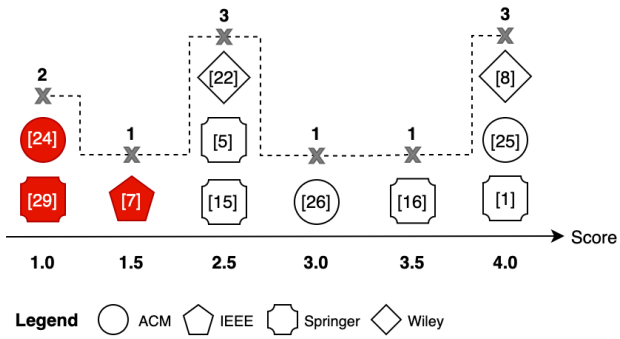

Figure 3: Publication score of the selected articles by database.

Figure 4 shows the total score of each question from the articles presented in Table 2. The most answered question is QA1 and the least answered question is QA3 since a few articles focused on presenting a training assistance application without explicitly explaining which HPE method was applied.

Figure 4: Total score by question.

4 ANSWER TO THE RESEARCH QUESTIONS AND DISCUSSION

This section covers the analysis of the selected articles, the ones that passed all the filtering process and the quality assessment. Each research question defined in Section 4 is answered in one of the subsections.

4.1 SQ1 - What are the challenges of using human pose estimation for training assistance?

Motion blur is one of the challenges in pose estimation due to the fast motion of subjects and cameras typical in sports and physical activities [25]. In dance videos, pose estimation is challenging since the dance movement changes quickly with large magnitude resulting in motion blur in video frames [8]. To overcome this challenge, a post-processing method can be used to correct the pose estimation results using a prior knowledge and timing information [8]. Wang et al. [25] introduced a structural-aware convolution approach that takes into account both spatial and temporal information of human keypoints in video sequences.

For golf swing pose estimation, one of the challenges is to ensure the system provides consistent accuracy and performance because when it comes to 2D video analysis the keypoints of the athlete changes too quickly in a short period of time and evaluating static images doesn't provide enough information for a time sequence analysis. Another challenge is data collection and labelling since golf swing is a very specific scenario [16].

Correctly classifying the joint points of different individuals and respectively matching them can be challenging for bottom-down techniques such as OpenPose and associate embedding which can effectively decouple the number of people and detection speed [8].

In 2D images, the posture estimation can cause ambiguities and singularities between different poses due to occlusion [1] [5] [15]. Multiple cameras can be used to resolve the self occlusion problem. For traditional pose estimation approaches based on image feature extraction, poor contrast between subject and background and loose-fitting clothing can also reduce pose estimation accuracy. Silhouette images cannot resolve self-occlusion problems because this method only extracts information of the body edges and do not provide any information of the image texture [1].

4.2 SQ2 - Which technologies are being used for the human pose estimation?

Leow et al. [15] used sampling techniques by applying an extension of Belief Propagation algorithm [9] to estimate pose samples of each body part. The pose samples are then used to generate posture candidates that have same frontal projection as the corresponding reference posture but different side projections to capture all possible depth ambiguities in a 2D image.

Wei et al. [26] formulated the 3D pose estimation as a per-pixel classification problem and apply trained randomized decision trees to associate each depth pixel with a particular bone segment for automatic labelling. This is a bottom-up approach to estimate 3D joint positions from a single depth image. The proposed framework integrates depth and silhouette information, full-body geometry, and temporal pose priors.

Afrouzian et al. [1] used a multi-view method and for each image the proposed method applies an example-based model known as Shape Context [2] which use a dataset of silhouettes in different poses and orientations to extract features and compare query images against the dataset samples. The nearest 3D poses available in the dataset are selected for each camera and the final 3D posture is selected from 3D pose candidates chosen by all cameras.

Recently, neural networks have been replacing the traditional techniques based on image feature extraction because of its capability to learn high-dimensional features of each joint, thus improving the accuracy of the extraction of the key features of the human body.

Due to the required accuracy in pose estimation for rehabilitation exercises, Escalona et al. [5] applied an end-to-end adversarial learning of human pose and shape approach called Human Mesh Recovery [11] to reconstruct a full 3D mesh of a human body from a monocular RGB image in which a deep learning-based encoder is able to predict the person's shape and pose as well as the camera position for each image and a discriminator is able to validate if the prediction corresponds to a real person or not. The proposed learning-based approach is able to accurately estimate the human pose even under self-occlusion scenarios.

Wang et al. [25] used a heatmap-based model using a CNN to extract features from each frame along with Spacial-Temporal Relation Module (STRM) to extract spatial relation among different keypoints. Spatial relation means the structural information of human body and temporal relation means the smooth movement of a keypoint along time dimension.

Dan Shi and Xin Jiang [22] applied a bottom-up approach using a multi-layer neural network and a confidential map to predict the positions of joint points along with affinity domain to predict the directions and positions of four limbs and obtain the posture of each individual.

Liu et al. [16] used AlphaPose [6], a top-down method, to obtain the key body points for each frame.

Guo et al. [8] used a top-down neural network-based pose estimation method applying confidential map with multiple channels,

Table 2: Quality assessment scores for each article.

Ref.	Year	Authors	HPE Context	QA1	QA2	QA3	QA4	Score
[8]	2020	Guo et al.	Cheer and Dance	Yes	Yes	Yes	Yes	4.0
[25]	2019	Wang et al.	Skiing and sport videos	Yes	Yes	Yes	Yes	4.0
[1]	2015	Afrouzian et al.	Soccer	Yes	Yes	Yes	Yes	4.0
[16]	2020	Liu et al.	Golf Swing	Yes	Yes	Yes	Partial	3.5
[26]	2012	Wei et al.	Wide range of human activities	Yes	Yes	Partial	Partial	3.0
[22]	2020	Dan Shi and Xin Jiang	Sports in general	Yes	Partial	Partial	Partial	2.5
[5]	2019	Escalona et al.	Rehabilitation exercises	Yes	Partial	Partial	Partial	2.5
[15]	2011	Leow et al.	Different type os sports	Yes	Partial	Partial	Partial	2.5
[7]	2019	Gu et al.	Rehabilitation exercises	Partial	Partial	No	Partial	1.5
[24]	2020	Tharatipyakul et al.	Exercise videos	Partial	No	No	Partial	1.0
[29]	2019	Zou et al.	Yoga, Taijiquan, and others	Yes	No	No	No	1.0

Table 3: HPE technologies and approaches.

Ref.	Year	HPE Technology	HPE Approach
[8]	2020	Neural Network	Top-down
[16]	2020	AlphaPose	Top-down
[22]	2020	Multi-layer Neural Network	Bottom-up
[25]	2019	CNN with STRM	Heatmap-based
[5]	2019	Adversarial Learning	Discriminative
[1]	2015	Shape Context	Multi-view
[26]	2012	Classification	Depth Camera
[15]	2011	Belief Propagation	Generative

known as heatmaps, to detect joints of human body and obtain the final skeleton representation. They also applied a post-processing step to increase total accuracy of pose estimation via prior knowledge and timing information.

Figure 5 illustrates a taxonomy of both 2D and 3D HPE methods adding a reference to the article with the respective applied method. This taxonomy combines previous taxonomies elaborated from other HPE surveys and reviews [4] [20] [28].

Table 3 presents pose estimation technologies and methods for each article related to the year of publication in a descending order are presented.

Figure 6 illustrates the HPE methods in relation to their application contexts. The majority of the articles apply pose estimation from videos and images due to the improvement of deep learning-based methods and the affordability of cameras.

4.3 SQ3 - In which context is human pose estimation being used for?

It can be noticed that pose estimation has a wide range of applications when it comes to motion analysis. This article focused on HPE applied to sports and training assistance and inside this field of study there are articles that focused on different contexts.

Guo et al. [8] proposed a visualization-driven approach to analyze dance and cheer leading videos to help with online teaching, reducing the workload of teachers and improving their work efficiency. The user can interactively analyze the quality of dance moves along the time line.

HPE can help with the detailed analysis of a golf swing, which relies on a full-body coordination to be executed precisely. Liu et al. [16] introduced an automatic body motion analysis method to assist the golf player and to reduce the instructor's workload.

HPE has been used to assist with rehabilitation or other exercises by guiding the user on the execution of exercises and by watching a personal trainer from a video or in augmented reality. The user would be able to follow the expert's movements in real time [5][15].

HPE can be applied along with action correction to provide a better analysis of detailed movements and give feedback of what needs to be improved on the exercise [22] [25]. Wang et al. [25] proposed an application that can detect good and bad pose in skiing technique based on keypoints information and is able to show a red and green mark around the individual to indicate a bad and good pose respectively.

Afrouzian et al. [1] applied HPE of soccer players using multi-view method with a dataset of silhouettes with pose configurations and viewpoints with respect to the camera.

HPE can be used with depth cameras to accurately capture full-body motion in a wide range of activities [26].

A summary of the pose estimation context of each article can be found in Table 2.

4.4 SQ4 - What is the pose estimation accuracy for the applied training context?

Guo et al. [8] used the mean of percentage of correct frames (mPCF) to count whether the human skeleton detected in each frame is in line with the real human pose instead of using the average accuracy of each joint. It was noticed that the model trained by the COCO dataset is better in mPCF than the MPII one. A heatmap with higher resolution and deeper backbone network helped to achieve better results. Adding post-processing methods can improve the pose estimation accuracy. The best mPCF achieved without post-processing was 64.30% and with post-processing it achieved 85.97%, both using COCO dataset as the training data.

Afrouzian et al. [1] used a common pose estimation error metric in literature: the Percentage of Correctly estimated body Parts (PCP)

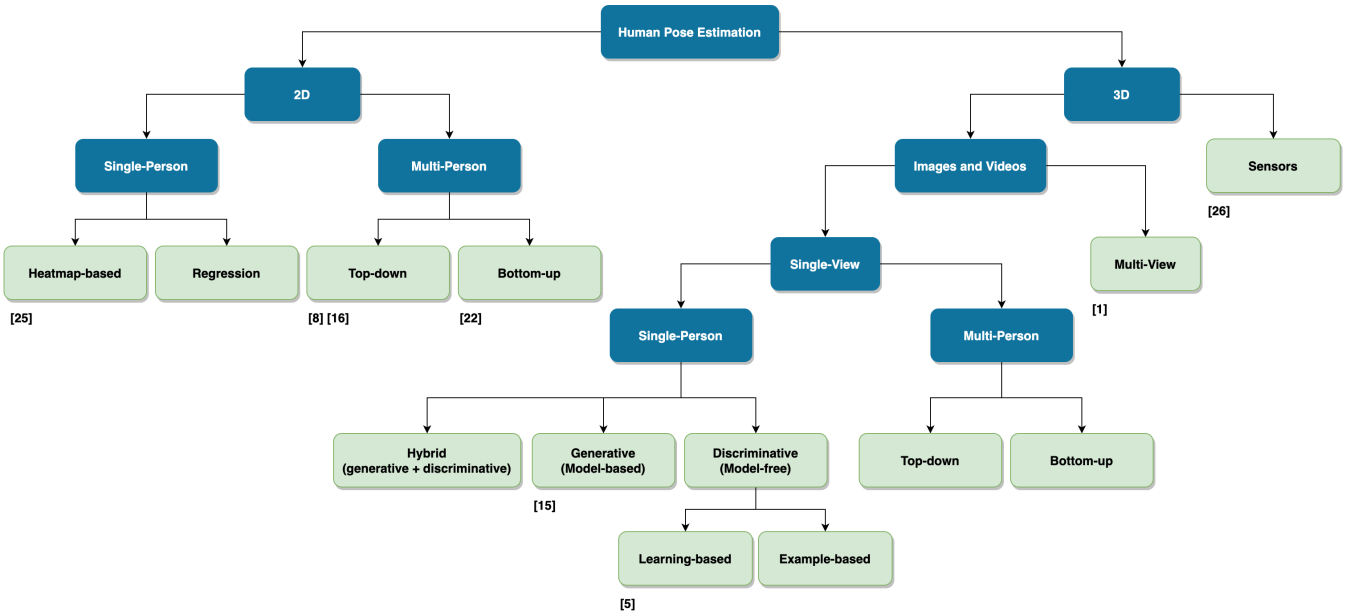

Figure 5: Taxonomy of Human Pose Estimation Methods.

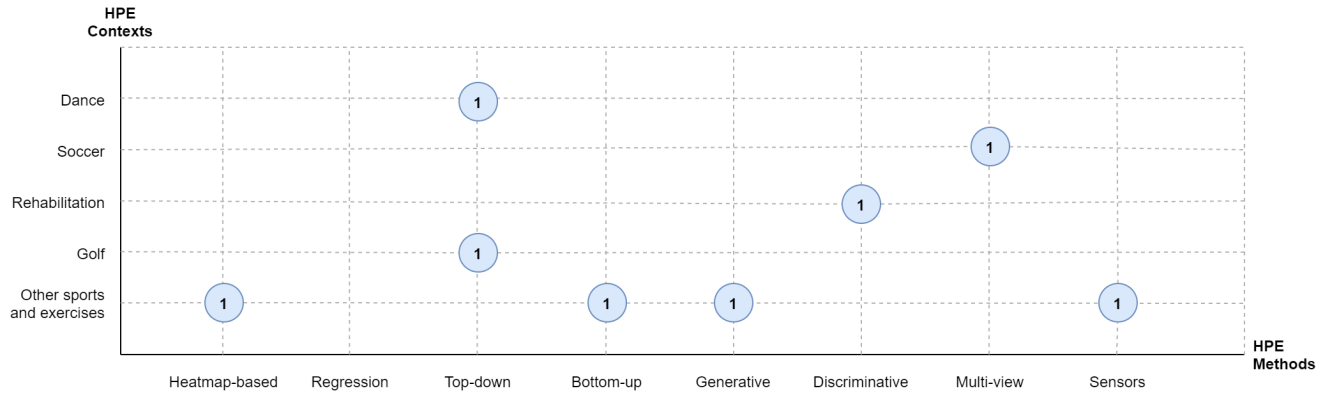

Figure 6: Summary of HPE methods among its application contexts.

to evaluate its human pose estimation system. An estimated body part is correct if:

$$\frac{\|s_n - \hat{s}_n\| + \|e_n - \hat{e}_n\|}{2} \leq \alpha \|s_n - e_n\| \quad (1)$$

Where s_n and e_n are the ground truth 3D coordinates of end points of part n. Also, $\hat{s}_n$ and $\hat{e}_n$ the respective estimations, and α is the parameter that controls the threshold [20], which in this case is between 0.2 and 0.5. Their HPE was able to get a PCP of 80.55% with viewpoint constraint and 72.26% without viewpoint constraint.

Percentage of Correct Keypoints (PCK) can also be used to measure the accuracy of localization of different keypoints within a given threshold [28]. Wang et al. [25] used PCK@0.2 which means that the prediction is considered correct if it lies within $(\alpha = 0.2) \times \max(h, w)$ where h and w represents the height and width of the bounding box respectively. The performance is compared with previous works on Penn dataset as well as sub-JHMDB dataset using a 3-fold cross validation. It was able to achieve a PCK of 94.9% on sub-JHMDB dataset and 99.1% on Penn dataset.

Liu et al. [16] tested their architecture by using a 4-fold validation with 75% of the videos used for training and the remaining 25% for testing. The average accuracy was 87.65%.

In order to test the effectiveness of the system, Dan Shi and Xin Jiang [22] selected 30 students to repeat a yoga movement 10 times. The authors' experimental results show that the accurracy achieved for HPE has reached 98.5%.

The accuracy of the human 3D pose estimation applied by Escalona et al. [5] was evaluated using the KARD dataset. 92% of the joints yieled an error below 20px, and 80% below 12px. The mean error average is 9.58px for an image with 224 X 224 pixels of resolution.

5 CONCLUSION

This research identified the current scenario in articles related to the use of pose estimation to help with sports motion analysis by reviewing which architecture and frameworks were used, what are the challenges of using pose estimation, and what are the main sports that were applied for training assistance applications. Four research questions and four quality assessment questions were defined to guide this systematic literature review.

The study identified an evolution in the architecture used for pose estimation. Recent articles applied deep neural networks to better detect key body points in 2D images, thus improving the pose estimation accuracy. We have noticed that occlusion, motion blur, and fast movements are still challenges for pose estimation, requiring additional processing to improve its accuracy in 2D deep learning-based HPE applications. The authors applied post-processing methods such as prior knowledge and timing information correction [8], spatial and temporal relation of human keypoints [25] and correction filter to eliminate the noise when tracking coordinates of occluded poses in images [16].

Pose estimation methods still require better training dataset to improve its generalization ability [26]. Computed posture error could be mapped to a domain-specific error based on a prior domain knowledge to have more useful and direct feedback to the user. For example, in Taichi a small error in the torso orientation needs to be a major error by the domain-specific criteria [15]. The current pose estimation methods needs improvements to track posture from people with amputated limbs [5]. The use of AR/VR could leverage the users experience in training assistance [5].

Pose estimation can be applied to online coaching, which has increased due to the pandemic, and can reduce the workload of teachers and help improve their work efficiency [8]. It can be used for sports action correction from videos, and for a personalized training experience by showing a virtual personal trainer using augmented reality.

REFERENCES

[1] Reza Afrouzian, Hadi Seyedarabi, and Shohreh Kasaei. 2016. Pose estimation of soccer players using multiple uncalibrated cameras. *Multimedia Tools and Applications* 75, 12 (2016), 6809–6827.

[2] Serge Belongie, Jitendra Malik, and Jan Puzicha. 2002. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence* 24, 4 (2002), 509–522.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[4] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. 2019. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology* 24, 6 (2019), 663–676.

[5] Felix Escalona, Ester Martinez-Martin, Edmanuel Cruz, Miguel Cazorla, and Francisco Gomez-Donoso. 2019. EVA: EVAluating at-home rehabilitation exercises using augmented reality and low-cost sensors. *Virtual Reality* (2019), 1–15.

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2334–2343.

[7] Yiwen Gu, Shreya Pandit, Elham Saraee, Timothy Nordahl, Terry Ellis, and Margrit Betke. 2019. Home-Based Physical Therapy with an Interactive Computer Vision System. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[8] Hong Guo, Shanchen Zou, Chuying Lai, and Hongxin Zhang. 2021. PhyCoVIS: A visual analytic tool of physical coordination for cheer and dance training. *Computer Animation and Virtual Worlds* 32, 1 (2021), e1975.

[9] Gang Hua and Ying Wu. 2004. Multi-scale visual tracking by sequential belief propagation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Vol. 1. IEEE, I–I.

[10] Hiteshi Jain and Gaurav Harit. 2017. Detecting missed and anomalous action segments using approximate string matching algorithm. In *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*. Springer, 101–111.

[11] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7122–7131.

[12] Soudeh Kasiri, Clinton Fookes, Sridha Sridharan, and Stuart Morgan. 2017. Fine-grained action recognition of boxing punches from depth imagery. *Computer Vision and Image Understanding* 159 (2017), 143–153.

[13] Naimat Ullah Khan and Wanggen Wan. 2018. A review of human pose estimation from single image. In *2018 International Conference on Audio, Language and Image Processing (ICALIP)*. IEEE, 230–236.

[14] Barbara Kitchenham, O Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology* 51, 1 (2009), 7–15.

[15] Wee Kheng Leow, Ruixuan Wang, and Hon Wai Leong. 2012. 3-D–2-D spatiotemporal registration for sports motion analysis. *Machine Vision and Applications* 23, 6 (2012), 1177–1194.

[16] Jen Jui Liu, Jacob Newman, and Dah-Jye Lee. 2020. Body Motion Analysis for Golf Swing Evaluation. In *International Symposium on Visual Computing*. Springer, 566–577.

[17] Yi Liu, Ying Xu, and Shao-bin Li. 2018. 2-D human pose estimation from images based on deep learning: a review. In *2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. IEEE, 462–465.

[18] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–14.

[19] Thomas B Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal. 2011. *Visual analysis of humans*. Springer.

[20] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 2016. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding* 152 (2016), 1–20.

[21] Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, and Jie Yang. 2020. AMIL: Adversarial Multi-instance Learning for Human Pose Estimation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 1s (2020), 1–23.

[22] Dan Shi and Xin Jiang. 2021. Sport training action correction by using convolutional neural network. *Internet Technology Letters* 4 (2021), e261.

[23] Cristian Sminchisescu. 2008. *3D Human Motion Analysis in Monocular Video: Techniques and Challenges*. Springer Netherlands, 185–211.

[24] Atima Tharatipyakul, Kenny TW Choo, and Simon T Perrault. 2020. Pose Estimation for Facilitating Movement Learning from Online Videos. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–5.

[25] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. 2019. AI coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *Proceedings of the 27th ACM International Conference on Multimedia*. 374–382.

[26] Xiaolin Wei, Peizhao Zhang, and Jinxiang Chai. 2012. Accurate realtime full-body motion capture using a single depth camera. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 1–12.

[27] Barry D Wilson. 2008. Development in video technology for coaching. *Sports Technology* 1, 1 (2008), 34–40.

[28] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2020. Deep Learning-Based Human Pose Estimation: A Survey. *arXiv preprint arXiv:2012.13392* (2020).

[29] Jiaqi Zou, Bingyi Li, Luyao Wang, Yue Li, Xiangyuan Li, Rongjia Lei, and Songlin Sun. 2018. Intelligent Fitness Trainer System Based on Human Pose Estimation. In *International Conference On Signal And Information Processing, Networking And Computers*. Springer, 593–599.