UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

3D human pose estimation and action recognition using fisheye cameras

A survey and benchmark

Zhang, Yahui; You, Shaodi; Karaoglu, Sezer; Gevers, Theo

DOI
10.1016/j.patcog.2024.111334

Publication date
2025

Document Version
Final published version

Published in
Pattern Recognition

License
CC BY

Link to publication

Citation for published version (APA):
Zhang, Y., You, S., Karaoglu, S., & Gevers, T. (2025). 3D human pose estimation and action recognition using fisheye cameras: A survey and benchmark. *Pattern Recognition*, *162*, Article 111334. https://doi.org/10.1016/j.patcog.2024.111334

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Nov 2026

Pattern Recognition 162 (2025) 111334

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

3D human pose estimation and action recognition using fisheye cameras: A survey and benchmark

Yahui Zhang [a,*], Shaodi You [a], Sezer Karaoglu [a,b], Theo Gevers [a,b]

[a] University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands
[b] 3DUniversum, Science Park 400, 1098XH Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
3D human pose estimation
Human action recognition
Fisheye camera
Benchmark

ABSTRACT

3D human pose estimation based on visual information aims to predict 3D poses of humans in images or videos. The aim of human action recognition is to classify what kind of actions people do. Both topics are widely studied in the field of computer vision.

Existing methods mainly focus on 3D human pose estimation and human action recognition using images/videos recorded by perspective cameras. In contrast to perspective cameras, fisheye cameras use wide-angle lenses capturing wider field-of-views (FOV). Fisheye cameras are used in many applications such as surveillance and autonomous driving.

In this paper, a survey is given on monocular 3D human pose estimation and action recognition. A new benchmark dataset is proposed using a fisheye camera to quantitatively compare and analyze existing methods.

1. Introduction

3D human pose estimation (3D HPE) from a single image is an active research field in computer vision with many applications such as augmented reality [1], virtual reality [2], and human–robot interaction [3]. The goal is to infer 3D human joint locations from a single image. With the emergence of deep learning and large scale datasets, 2D human pose estimation made significant progress recently. However, 3D human pose estimation from a single image is (still) an ill-posed problem due to the inherent depth ambiguity and changing imaging conditions introducing variations in (human) appearance and self-occlusions. Accurate 3D human pose estimation necessitates the annotation of 3D joint locations, a critical component for effective model training. Unfortunately, the process of annotating 3D labels is both labor-intensive and time-consuming, complicating the development of comprehensive 3D pose datasets. Without high-quality annotated datasets, the performance and robustness of 3D pose estimation models can be severely limited. Furthermore, 3D pose datasets are usually collected in constrained environments limited by a motion capture device. Considering the difference between constrained and in-the-wild environments, it is still a challenge for existing methods to generalize well to in-the-wild images and unseen poses (see Fig. 1).

Human action recognition (HAR) is also a very active research field in computer vision with many applications such as medical diagnosis and security. The aim is to recognize actions performed by persons in videos. In general, HAR methods exploit (single) images, videos or skeletons as their input. In this paper, we focus on skeleton-based methods. As it is hard and time-consuming to annotate human poses, HPE is beneficial for HAR.

However, existing research mainly focuses on 3D HPE and HAR utilizing perspective cameras. In contrast, the wider field of view afforded by fisheye cameras has led to their widespread application in areas such as surveillance, photography, and sports. Nonetheless, as illustrated in Tables 1 and 2, there are only a few existing methods that exploit fisheye cameras. Moreover, there is an absence of datasets tailored for fisheye-based 3D HPE and HAR. To this end, we propose a new dataset for multi-person 3D pose estimation (F-M3DHPE), and skeleton-based HAR (F-HAR) captured by a fisheye camera. There are several surveys on HPE and HAR, including monocular HPE [5], multi-person HPE [6], and HAR [7]. However, they all focus on perspective cameras and ignore egocentric 3D HPE. To provide a complete picture of this important research area, our aim is to provide a comprehensive survey on the recent advances of 3D HPE and HAR for both perspective and fisheye cameras.

The remainder of this paper is organized as follows. Camera models are given in Section 2. Sections 3 and 4 provide a survey on deep learning (DL)-based single-/multi-person 3D pose estimation and human action recognition. In Section 5, a comparative study is conducted with existing methods on public datasets (perspective cameras) and on our newly collected dataset (fisheye cameras). Section 6 discusses future perspectives on 3D human pose estimation and skeleton-based human action recognition.

* Correspondence to: University of Amsterdam, Computer Vision Lab, Science Park 904, 1098XH, The Netherlands.
E-mail address: yhzhang2023@gmail.com (Y. Zhang).

https://doi.org/10.1016/j.patcog.2024.111334
Received 10 April 2023; Received in revised form 13 October 2024; Accepted 26 December 2024
Available online 3 January 2025
0031-3203/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

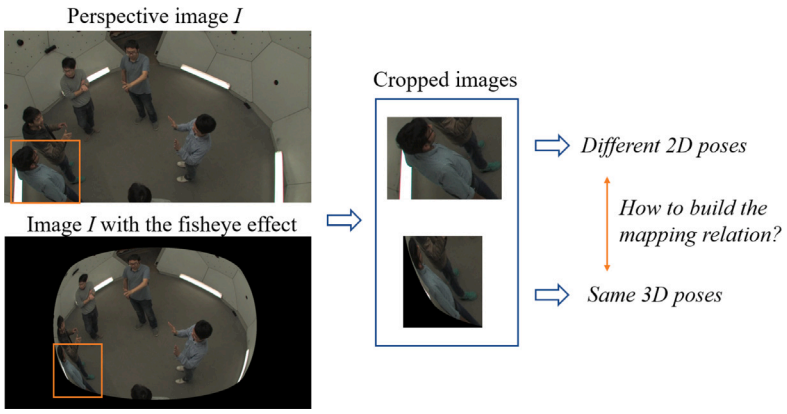

Fig. 1. The challenge of fisheye images for 3D HPE. Although cropped images around the same human exhibit different appearances and 2D poses, they correspond to the same 3D poses.

Table 1
Summary of methods for 3D HPE considering cameras, viewpoints and the number of humans in the images.

Existence		Camera		Viewpoint		The number of persons	
Method	Dataset	Pin-hole	Fisheye	First-person (egocentric) view	Third-person view	Single person	Multi persons
✓	✓	✓			✓	✓	
✓	✓	✓			✓		✓
✓	✓		✓	✓		✓	
✗	✗		✓		✓		✓

Table 2
Summary of skeleton-based methods for HAR considering cameras and viewpoints.

Existence		Camera		Viewpoint	
Method	Dataset	Pin-hole	Fisheye	First-person (egocentric) view	Third-person view
✓	✓	✓		✓	
✓	✓	✓			✓
✗	✗		✓		✓

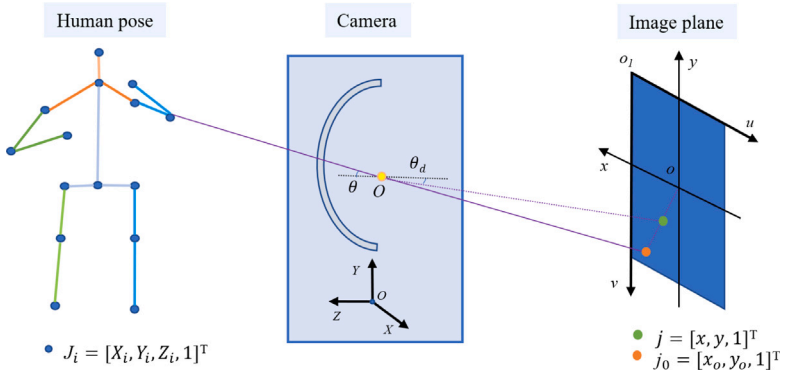

Fig. 2. Projection from 3D joint locations in camera coordinates $OXYZ$ to 2D keypoints in image plane o_1uv [4]. This figure consists of (1) a human pose represented by a series of discrete joints J_i; (2) a camera; (3) an image plane. θ denotes the angle of incidence while θ_d represents the angle of refraction when there is a fisheye camera. j and j_o indicate 2D projections.

2. Camera models

In this section, we outline the details of perspective and fisheye camera models.

2.1. Definition

Fig. 2 shows the projection of a 3D human pose onto a 2D pose. The human pose is represented by a set of discrete joints $\boldsymbol{J}_i$ in the camera coordinate system $OXYZ$. The middle part is a perspective or fisheye camera. Finally, the 2D projection of a joint is denoted as $\boldsymbol{j}_o$ for a perspective camera, otherwise it is represented by $\boldsymbol{j}$ where the angle of refraction is reduced from θ to θ_d.

Let $\mathbf{P}_{3D} = [\boldsymbol{J}_1, \boldsymbol{J}_2, \ldots, \boldsymbol{J}_n]$ denote a 3D human pose in the camera coordinate system, where n is the number of human joints. The human joints $\boldsymbol{J}_i = [X_i, Y_i, Z_i, 1]^T$ are represented by a 4 by n matrix. $\mathbf{p}_{2D}$ and $\mathbf{p}_{o2D}$, denoted by a 3 by n matrix with $\boldsymbol{j} = [x, y, 1]^T$ and $\boldsymbol{j}_o = [x_o, y_o, 1]^T$, are the 2D projections of human poses using a perspective camera and fisheye camera model, respectively.

2.2. Perspective camera model

$\boldsymbol{j}_o$ in Fig. 2 is the 2D projection using a perspective camera model. In this setting, the angle of incidence and refraction is the same, *i.e.,* θ. According to the triangular similarity, the formulation is obtained by:

$$s \cdot \mathbf{p}_{o2D} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3D}, \tag{1}$$

where $\boldsymbol{R}$ and $\boldsymbol{T}$ are extrinsic camera parameters representing rotation and translation parameters, respectively, $\boldsymbol{K}$ denotes the intrinsic camera parameters, and s is a scale factor and equals the depth in the camera coordinate system. As 3D joint locations $\mathbf{P}_{3D}$ are usually defined in the camera coordinate system, $\boldsymbol{R}$ and $\boldsymbol{T}$ are the identical matrices in Eq. (1).

2.3. Fisheye camera model

In contrast to perspective cameras, fisheye cameras contain wide-angle lenses capturing a wider field-of-view (FOV) i.e. capturing more of a scene, but cause image distortions. As shown in Fig. 2, 2D projections are displayed for a fisheye camera. Particularly, the farther the object is from the center of the image, the stronger the distortion will be.

For the fisheye camera model, the relationship between the angle of incidence and refraction, *i.e.,* θ and θ_d, needs to be considered to calculate the 2D predictions. Therefore, the distortion matrix $\boldsymbol{D}$, to compute 2D projections using Eq. (1), is given by:

$$s \cdot \mathbf{p}_{2D} = \boldsymbol{K}\boldsymbol{D}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3D}. \tag{2}$$

$\boldsymbol{D}$ is defined by:

$$\boldsymbol{D} = \begin{bmatrix} \theta_d/l & 0 & 0 \\ 0 & \theta_d/l & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where $l = \frac{\sqrt{X^2+Y^2}}{Z}$.

Based on [8], the angle of refraction is calculated by:

$$\theta_d = \theta(1 + k_1\theta^2 + k_2\theta^4 + k_3\theta^6 + \cdots), \tag{4}$$

where $\theta = \arctan(l)$.

2.4. Discussion

In this section, the discussion of the difference between perspective and fisheye camera models are provided. The perspective camera model as defined by Eq. (1) depends on extrinsic and intrinsic camera parameters. For simplicity, we only consider the common case that the 3D joint locations are in the camera coordinate system. Hence, the projection only relies on the intrinsic camera parameters, *i.e.,* focal length and principal coordinates. The focal length primarily influences the scale of the human in the image, while the principal coordinates affect the position of the human within the frame. Existing methods [9] leverage cropped images to estimate these parameters for 3D human pose projection, allowing for effective determination of scale and translation.

In contrast, fisheye cameras introduce distortion due to the wide field of view, which is accounted for by additional distortion parameters, as defined in Eq. (3). This distortion affects the entire image, making it necessary to use the full 2D image when estimating both the intrinsic parameters and the distortion parameters. Cropping the image would exclude crucial information about the fisheye projection, resulting in inaccurate estimations. Thus, for accurate 3D human pose estimation from fisheye cameras, it is essential to consider the entire image, particularly when the camera parameters are not provided.

3. Survey on 3D human pose estimation

In this paper, we focus on 3D human pose estimation from a single image. This task is ill-posed due to challenging imaging conditions such as inherent depth ambiguities, appearance changes and occlusions. Most of the existing methods perform 3D pose estimation from perspective cameras. Fisheye cameras (wide-angle lenses) capture a wider field-of-view (FOV) than perspective cameras. Therefore, fisheye cameras are widely used in various applications. In this survey, we focus on the following categories: single-person 3D pose estimation, multi-person 3D pose estimation, and 3D pose estimation from fisheye cameras.

3.1. Single-person 3D pose estimation

Recently, single-person 3D pose estimation shows significant improvement with the use of deep learning and large-scale 3D human pose datasets. 3D single-person pose estimation can be divided into one-stage and two-stage approaches.

(1) One-stage approaches

One stage approaches directly perform single-person 3D pose estimation from a single input image to 3D joint locations without intermediate supervision (*e.g.* 2D keypoints). Li and Chan [10] design a multi-task learning framework for 3D single-person pose estimation, where pose regression and body part detectors are simultaneously trained. Li et al. [11] propose a method taking 3D poses and images as input to compute the score if the two inputs match well. This method can be used as a refinement module for other methods. To alleviate the dependency on 3D poses and input images, Tekin et al. [12] use an auto-encoder to learn structured pose representations in latent space. The latent pose representations are beneficial for 3D human pose estimation. Sun et al. [13] present a regression-based method for 3D pose estimation, where the loss is based on the bone length and vector. Pavlakos et al. [14] propose to use volumetric representation for 3D human poses. They estimate the likelihoods of each voxel in discretized space around the subject. To reduce the computational cost, they adopt a coarse-to-fine scheme. Pavlakos et al. [15] suggest ordinal depth as extra information to infer 3D human joint locations. The method alleviates the restrictions that 3D pose annotations are prerequisite, and utilizes 2D pose datasets with ordinal depth annotations during the training process to improve the performance.

(2) Two-stage approaches

Two-stage approaches first use off-the-self detectors to estimate 2D keypoints, and then infer 3D human joint locations from 2D predictions with/without features extracted from the input images (see Fig. 3).

Inference from 2D poses to 3D poses. Due to the robustness of 2D keypoint detectors, many methods focus on 3D pose estimation directly from the estimated 2D poses. Chen and Ramanan [16] propose a matching scheme to obtain 3D poses from a 3D pose library by estimating 2D poses and depths from a single image. Moreno-Noguer [17] formulates 3D human pose estimation from a single image as a regression problem of a 2D-to −3D distance matrix. Tekin et al. [18] aggregate features extracted from original images and intermediate 2D poses with heatmap representations to infer 3D human joint locations. Martinez et al. [19] design a simple series of fully connected layers but obtain superior performance. To reduce the quantization error caused by extracting keypoints from heatmaps, Sun et al. [20] introduce an integral operation (also referred to as soft-argmax) to reshape the optimization objective. Zhou et al. [21] utilize the representation of three joint heatmaps to learn the local relations between human body parts based on 2D keypoints and relative depth information. Li et al. [22] present a regression-based method by exploring maximum likelihood estimation to deal with the uncertainty of the distribution for 2D and 3D human pose estimation. Chen et al. [23] propose an efficient method to search optimized architectures for 3D human pose estimation. Moreover, Nie et al. [24] aims to let the model adapt the two domains of 2D and 3D

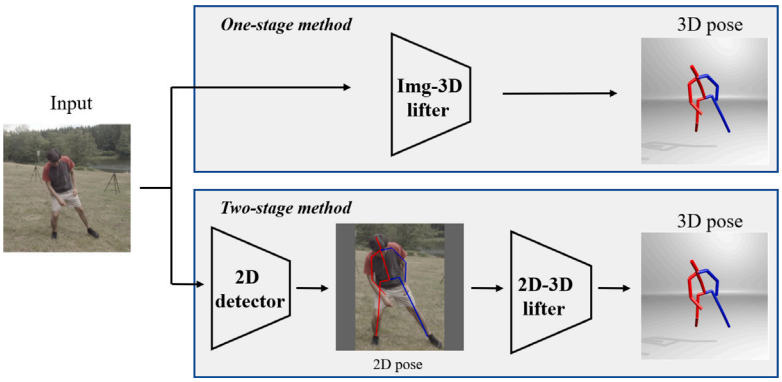

Fig. 3. Overview of single-person 3D pose estimation. Part of the figure is from [9].

poses to generate more plausible 3D predictions. Similarly, Ci et al. [25] generate multiple 3D human poses by utilizing gradient fields derived from learned prior information in existing datasets for guidance.

Graph convolutional network (GCN)-based methods. Recently, researchers focus on GCN-based methods for 3D single-person pose estimation, aiming to obtain pose representations from 2D graphs. Ci et al. [26] introduce a locally connected network combining GCN and a fully connect network for 3D human pose estimation. The aim is to alleviate the limitations of GCN on learning pose representations caused by a weight sharing scheme. To deal with the limitations of GCN for the regression problem, Zhao et al. [27] propose a semantic graph convolutional network (SemGCN) to learn semantic relationships of nodes by combining 2D keypoints with features extracted from input images at 2D keypoints. Liu et al. [28] provide a comprehensive study on the impact of sharing weight schemes with feature transformations. They show that pre-aggregation in GCN, *i.e.,* applying transformations to the 2D input and then aggregating them, is beneficial for 3D human pose estimation. Xu et al. [29] propose a graph stacked hourglass network, aggregating multi-scale and multi-level feature information and 2D keypoints. Li et al. [30] combine the strengths of multi-layer perceptrons for global feature extraction and GCNs for local feature extraction, enabling both local and global spatial interactions for 3D human pose estimation. To improve the generalization and stability, Han et al. [31] introduce an uncertainty-aware scaling factor to convert the predictions for the optimization.

Geometric constraint-based methods. Since public datasets for 3D human pose estimation are usually captured in an constrained environment, generalization is limited to deal with in-the-wild images. To this end, different methods propose geometric constraints to explicitly use information from 2D in-the-wild datasets. Zhou et al. [32] propose a geometric constraint based on the ratio of upper and lower limbs as the prior information. In this way, the designed geometric loss can be used to constrain 3D predictions when trained on 2D in-the-wild pose datasets even without 3D annotations, alleviating the depth ambiguity. Additionally, Tripathi et al. [33] build a connection between the human body and the ground by the introduced pressure map. By ensuring the consistency of predicted joint locations and the pressure map, the predicted 3D poses are further regularized. On the other hand, a perspective camera model serves as an effective alternative way to provide relationships between 2D and 3D poses [9,34,35]. Specifically, Habibie et al. [9] propose a method by explicitly using 2D and 3D features in latent space. They add constraints to ensure consistency between projected 3D predictions and 2D poses. Wandt et al. [34] present a weakly supervised method based on an adversarial re-projection network. Their method can be trained with unpaired 2D-3D poses and generalizes well to in-the-wild images. Chen et al. [35] attempt to regress 3D human joint locations from 2D poses. They use a discriminator to assess the projected 3D projections by minimizing the re-projection error (see Fig. 4).

Although GCN-based methods and geometric constraint-based methods are not entirely mutually exclusive, the classification is intended to reflect their primary focus: the former prioritize learning spatial relations among 2D keypoints of human body through data-driven techniques for 3D pose estimation, while the latter emphasize enforcing human body structure through explicit rules, such as maintaining the ratio of limbs and ensuring the consistency between 3D and 2D poses using camera models.

(3) Comparison

The advantage of one-stage approaches is their efficiency. They directly estimate 3D poses without extra intermediate supervision. However, one-stage approaches heavily rely on fully annotated 3D pose datasets. In contrast, two-stage approaches can make use of in-the-wild images even without 3D annotations increasing their generalization capabilities. In general, two-stage approaches outperform one-stage approaches.

In summary, to address the challenge of depth ambiguity in 3D HPE from a single image, the methods can be classified into two categories. The first category is represented by the geometric constraint-based methods. These methods leverage limb characteristics, the projection relationships between 2D and 3D poses, and the estimated pressure distribution of the human body as the prior information to alleviate depth ambiguity in 3D HPE. The second category leverages the strong mapping capabilities of deep learning models, which establish relationships between a single image or 2D pose and the corresponding 3D pose. These methods emphasizes the development of advanced architectural designs specifically tailored for effectively extracting representative features that are critical to human pose analysis.

3.2. Multi-person 3D pose estimation

A more general scenario is to estimate multi-person 3D poses. With development of single-person 3D pose estimation and related deep learning techniques, researchers recently pay more attention to multi-person 3D pose estimation. Generally, there are two categories: bottom-up approaches and top-down approaches. Furthermore, methods can be divided into optimization-based or learning-based methods depending on how the absolute depths are obtained. Optimization-based methods attempt to obtain absolute depths by minimizing the distance between projected 3D poses and 2D ground truths, while learning-based methods exploit the extracted features and geometric constraints to perform absolute depth estimation.

(1) Bottom-up approaches

Bottom-up approaches firstly estimate all human joint locations followed by grouping all joints. Zanfir et al. [37] propose a bottom-up network (MubyNet) to perform 3D multi-person pose estimation as well as 3D shape estimation. MubyNet groups the joints and limbs based on 2D/3D information by limb scoring, transforming the grouping problem into an integer program. Metha et al. [38] introduce occlusion-robust

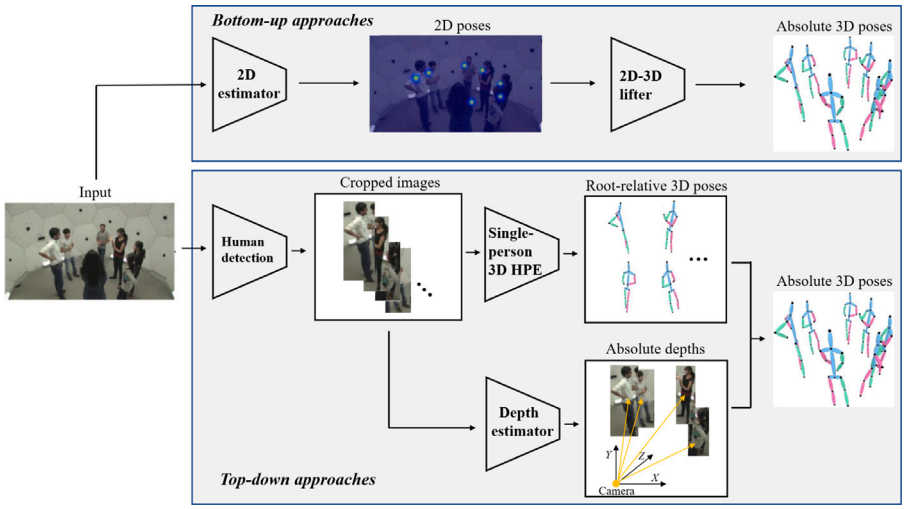

Fig. 4. Overview of multi-person 3D pose estimation. Part of the figure is from [36].

pose-maps to infer 3D poses by estimating 2D poses and Part Affinity Fields [39], even for strong occlusions. Zhen et al. [36] aggregate various information cues from 2.5D representations and 2D keypoints to infer multi-person 3D poses. To reduce the human-defined impact to determine whether the keypoints belong to the same person, Yang et al. [40] propose a instance-aware attention network to exploit the joint grouping.

(2) Top-down approaches

Top-down approaches start by detecting each person and then performing single-person 3D pose estimation to localize human joint positions. Rogez et al. propose LCR-Net [41] and LCR-Net++ [42] to estimate 2D/3D multi-person pose estimation. They consist of localization, classification and regression branches to estimate and refine the predictions. Dabral et al. [43] present a Mask-RCNN based network to detect 2D poses of each person and perform single-person 3D pose estimation from estimated 2D poses. The absolute 3D poses are obtained by minimizing the distance of the projected 3D predictions and 2D keypoints. Recent methods adopt a learning-based manner to compute the absolute depths. Moon et al. [44] introduce a novel depth measure combined with a correction factor to obtain the depth. Guo et al. [45] adopt a data augmentation scheme to deal with the occlusion problem for depth estimation. Lin and Lee [46] consider the depth regression as a classification problem to perform depth estimation and localization of root joints. Wang et al. [47] utilize hierarchical multi-person ordinal relations (HMOR), at joint, body part and human instances levels, to perform 3D multi-person pose estimation. Cucchiara and Fabbri [48] discuss different ways to deal with the occlusion problem and perspective constraints caused by top-down viewpoint in video surveillance. Qiu et al. [49] leverage perspective information to infer the relative depths of human joints for pretraining models, reducing costs by eliminating the need for 3D pose annotations.

(3) Comparison

In terms of computational costs, bottom-up approaches are more efficient than top-down approaches. The former makes a trade-off between computational cost and accuracy. The computational cost of top-down approaches depends on the detection module and the number of humans. Regarding the estimation accuracy, top-down approaches perform better than bottom-up approaches. The reason is that top-down approaches split the task into two subtasks, *i.e.,* human detection and 3D single-person pose estimation. The 3D pose estimator focuses on salient regions to predict 3D human poses. Recently, Cheng et al. [50] combine top-down and bottom-up networks to perform 3D multi-person pose estimation. They introduce an interaction-aware discriminator to integrate two kinds of 3D predictions from top-down and bottom-up network to obtain refined 3D predictions.

3.3. 3D human pose estimation from fisheye cameras

Fisheye cameras are widely used in various applications such as virtual reality, video surveillance and automatic driving. Surprisingly, there are only a few methods focusing on 3D human pose estimation from fisheye cameras. Based on the viewpoint of cameras, the methods are categorized into first-person view (egocentric) 3D HPE and third-person view 3D HPE.

(1) Egocentric 3D HPE

Rogez et al. [51] and Yonemoto et al. [52] propose a method for hand, arm and torso pose inference from RGB-D data. Jiang and Grauman [53] infer full body 3D joint locations from cameras mounted on the chest with the aim to estimate unseen 3D poses. Rhodin et al. [54] first propose a method for full body 3D human pose reconstruction from a pair of fisheye cameras mounted on a helmet. Shiratori et al. [55] present an approach based on structure-from-motion (SFM) for 3D pose estimation from wearable devices, where 16 limb-mounted cameras are used.

To capture the 3D poses of full human body, Xu et al. [56] propose a disentangled method for egocentric 3D pose estimation from a camera mounted on a baseball cap. To improve the estimation performance of the lower body, an original image and an image of the central part are taken as inputs to estimated 3D keypoints based on the known camera parameters. Tome et al. [57] design a novel two-stage method to detect 2D keypoints followed by an auto-encoder with two branches to obtain 3D human joint locations and 2D keypoints. To alleviate the requirement of ground-truth camera parameters, Zhang et al. [58] propose an automatic calibration method for egocentric 3D pose estimation. The influence of image distortions is alleviated by minimizing the error between projected 3D predictions and 2D keypoints. To further regularize the estimated 3D poses, Dhamanaskar et al. [59] introduce the constraint from the third views on egocentric 3D pose estimation.

(2) Third-person view 3D HPE

There is few methods for 3D HPE from third-person view for fisheye cameras. Zhang et al. [4] propose a re-projection module to alleviate the negative effect of images distortion on predicted root-relative poses and absolute depths. For egocentric settings, the distance between the camera and human head joint is fixed. In other words, the position of the head joint in the image is almost the same regardless of the human pose. Therefore, existing methods for egocentric 3D human pose estimation may fall short when applied to single/multi-person 3D pose estimation from a fisheye camera using a third viewpoint.

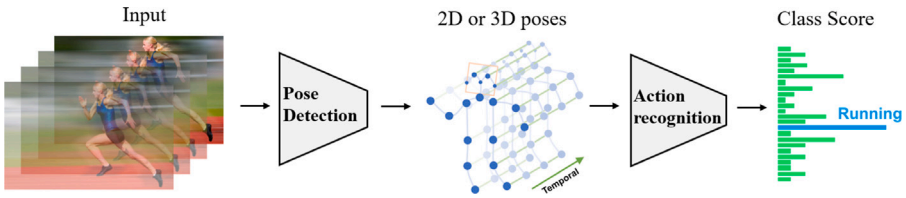

Fig. 5. Overview of human action recognition based on skeletons. Part of the figure is from [60].

4. Survey on human action recognition

Human action recognition generally consists of two elements: action representation and classification. Action representation is provided by a feature extractor. The classifier is used to predict an action label from the extracted features. Regarding feature extraction, methods for human action recognition are divided into traditional and DL-based methods. Due to the progress of deep learning techniques, DL-based methods for human action recognition received more attention recently. In this paper, we mainly focus on DL-based methods. DL-based methods can be grouped into video-based and skeleton-based methods. Video-based methods rely on the objects in the video. Skeleton-based methods focus on the change of 2D/3D skeletons detected in the videos. In this section, we focus on skeleton-based methods (see Fig. 5).

The input of skeleton-based methods is a sequence of 2D/3D human joint locations. Regarding the architecture design, skeleton-based methods can be divided into three categories: CNN-based, RNN-based, and GCN-based.

CNN-based methods. ConvNets achieve superior performance on feature extraction from images, while skeleton-based methods take human joint locations as their inputs. To bridge this gap, skeleton sequences are converted into image formats by projecting 3D joint locations on three orthogonal planes [61]. Then, 2D ConvNets are applied to extract features from the new representation followed by a classifier for action recognition. Liu et al. [62] propose a view-independent approach for action recognition mitigating the impact of skeleton variations caused by different viewpoints. To obtain additional motion cues, Zolfaghari et al. [63] use the raw image, pose and motion as inputs followed by a Markov chain model to exploit the spatial and temporal information. Hu et al. [64] and Luvizon et al. [65] explore the spatial-channel and visual-pose relation to capture better representations for human action recognition. To exploit richer cues from human poses, Liu et al. [66] adopt heatmap map of human poses to study the evolution of poses. Caetano et al. [67] propose a new representation of skeletons, SkeleMotion, calculated by the magnitude and orientation of each joint. Chen et al. [68] present a color-coding strategy to covert skeleton sequences into pseudo-color images, aiming to learn discriminate features for human action recognition. Li et al. [69] design several learning blocks of elastic units (Else-Net) to explore the relation between pose sequences. Different from the above methods, Asif et al. [70] firstly propose to extract motion and spatial information from facial, hands, and body joints for action recognition. Duan et al. [71] propose a 3D-CNN-based method to effectively learn spatio-temporal information from a 3D heatmap volume.

RNN-based methods. To exploit temporal information, RNNs are used to process sequential data. Du et al. [72] and Shahroudy et al. [73] group the skeleton into five body parts to model temporal dynamics. To increase explicit attention ability, Liu et al. [74] propose an attention-based LSTM for global context information extraction. To mitigate the gradient decay problem, Li et al. [75] present an independent RNN to perform human action recognition. Zhao et al. [76] present a Bayesian-based LSTM method to learn spatial information from joints, temporal information from poses and variation of subjects. To alleviate the variation of sequence skeletons caused by various viewpoints, Zhang et al. [77] introduce a view adaptive module to transform automatically input skeletons into representations under a new viewpoint for action recognition. Friji et al. [78] propose to optimize rigid and non rigid transformations of human poses to exploit the geometric information from skeletons. In addition to raw skeletons, Yadav et al. [79] involve geometrical and kinematic features to better exploit spatio-temporal information.

GCN-based methods. Since human skeletons can be seen as a graph structure, GCN-based methods received most attention for human action recognition. There are several types for GCN-based methods: modifications of GCNs, combination of GCNs and LSTMs, and multi-streams-based GCNs.

Firstly, there are methods to directly use GCNs to learn spatio-temporal representations for human action recognition. Yan et al. [60] propose a pioneer method, ST-GCN, to apply graph convolution networks to human action recognition. Li et al. [80] present an encoder–decoder GCN to learn action-related representations named AS-GCN, where actional and structural links are used by the GCN to extract spatial features. AS-GCN achieves promising performance on both human action recognition and pose prediction. Liu et al. [81] propose a multi-scale aggregation method, MS-G3D, to obtain unbiased information and perform cross-space time information flow effectively. To improve computational efficiency, Zhang et al. [82] propose a semantics-guided GCN to explore spatial and temporal information at joint and frame levels. Dai et al. [83] present a novel network that simultaneously learns spatio-temporal relationships between joints and global temporal dynamics, incorporating subsampling techniques to reduce parameter count while preserving model performance.

Secondly, Graph RNN frameworks are used for human action recognition. Si et al. [84] propose an attention mechanism using a LSTM architecture, named AGC-LSTM, to effectively extract spatio-temporal features for HAR. To obtain high-order information, Wen et al. [85] propose a motif-based GCN and a variable dense block to exploit spatial and temporal relations. Chen et al. [86] present a pooling strategy and a point-wise attention mechanism to learn action-related features from skeleton sequences. To exploit the spatial structure and detailed temporal information, Si et al. [87] propose to extract spatial structural features from each frame while get temporal information from the sequences. Instead of capturing skeleton differences between sequential skeletons, Ding et al. [88] focus on inherent differences in terms of spatio-temporal and context to determine the human actions.

Thirdly, multi-stream GCNs are proposed to fuse spatial and temporal features from different streams for human action recognition. Specifically, several branches are employed to explore the different cues (*e.g.,* joints, bones, motion and relative positions) temporal information extraction. Then, the extracted features are aggregated to perform HAR. The focus includes: flexibility of graph design [89], robustness for feature extraction and fusion [90,91], robustness for incomplete skeletons [92], and structural information preservation [93].

Comparison. From an accuracy and robustness perspective, GCN-based methods show significant promise by effectively exploring spatial and temporal interactions among 2D/3D human joints. RNN-based methods are particularly well-suited for modeling temporal sequences, while CNN-based methods are effective at detecting spatial patterns. Regarding computational cost, CNN-based methods benefit from their inherent parallel processing capabilities, providing faster performance compared to the sequential operations of RNN-based methods, which

Table 3
Public datasets for 3D human pose estimation.

Dataset	Year	Camera type	Environment	Size	The number of persons
HumanEva [94]	2010	Pin-hole	Indoor	6 subjects, 7 actions, 40k frames	Single
Human3.6M [95]	2014	Pin-hole	Indoor	11 subjects, 17 actions, 3.6M frames	Single
MPI-INF-3DHP [96]	2017	Pin-hole	Indoor	8 subjects, 8 actions, 1.3M frames	Single
Occlusion-Person	2020	Pin-hole	Indoor	73k frames	Single
Mo2Cap2 [56]	2018	(Egocentric) Fisheye	Indoor&Outdoor	8 actions, 530k frames	Single
*x*R-EgoPose [57]	2019	(Egocentric) Fisheye	Indoor	46 subjects, 9 actions, 383K frames	Single
CMU Panoptic [97]	2016	Pin-hole	Indoor	8 subjects, 1.5M frames	Multi
MuPoTS-3D [38]	2018	Pin-hole	Indoor&Outdoor	8 subjects, 8k frames	Multi
F-M3DHPE (Our dataset)	2022	Fisheye	Indoor	11 subjects, 10 actions, 2.8k frames	Multi

are generally more time-intensive. RNN-based methods perform better in recognizing intricate, temporally dependent actions, whereas CNN-based methods are more suited for identifying broader, spatially defined movements.

5. Datasets and benchmarks

Many datasets are available for 3D human pose estimation and human action recognition. In this section, we describe the publicly available datasets, our new dataset F-M3DHPE, the evaluation metrics, and the performance of existing methods on the public and proposed datasets.

5.1. Dataset and benchmarks for 3D human pose estimation

In this section, we provide a comparison of existing methods on public datasets. Particularly, we focus on single-/multi-person 3D pose estimation from a single image.

(1) Datasets for 3D human pose estimation

There are different 3D pose datasets for single-person 3D pose estimation for perspective cameras including HumanEva [94], Human3.6M [95], and MPI-INF-3DHP [96]. For 3D multi-person pose estimation, CMU Panoptic [97], the MuCo-3DHP, and MuPoTS-3D [38] datasets are commonly used for quantitative evaluation. In addition, the Mo2Cap2 [56] and *x*R-EgoPose [57] datasets are recently released for egocentric 3D human pose estimation captured by a fisheye camera. The details of the datasets are listed in Table 3.

MPI-INF-3DHP. MPI-INF-3DHP is collected by a marker-less motion capture for single-person pose estimation. The training set consists of more than 1.4M frames captured by 13 cameras. 3D pose annotations and universal skeleton are provided in this dataset. The test set includes both indoor and outdoor images with participants performing various activities such as standing, siting, exercise, sports, etc. Due to the existence of in-the-wild images and complexity, the test set is usually used to verify the generalization of existing methods (normally without fine-tuning on the training set).

MuCo-3DHP and MuPoTS-3D. MuCo-3DHP is a synthesized dataset build on the MPI-INF-3DHP dataset. The number of subjects in the image ranges from 1 to 4. Since 3D poses are known in the MPI-INF-3DHP dataset, MuCo-3DHP provides overlapping scenarios with correct depth ordering.

MuPoTS-3D is used for evaluation of multi-person 3D pose estimation captured by a multi-view marker-less motion capture system. This dataset consists of more than 8000 frames and 8 participants wearing different clothes performing daily activities. The dataset provides 3D pose annotations of up to 3 people collected for 20 scenes, where 5 indoor sequences are taken at 30 fps and 15 outdoor sequences are captured at 60 fps.

F-M3DHPE. The proposed F-M3DHPE dataset, for multi-person 3D pose estimation, is captured in an indoor environment with two Insta360 cameras, and one LiDAR camera. 11 participants with different clothes perform 10 common activities such as posing, discussion, walking, etc. Our dataset includes more than 2000 frames with a resolution of 1920 × 1920 pixels. Example images are shown in Fig. 6. To obtain 3D ground-truth poses, we manually label 2D poses for the images captured by a LiDAR camera. Then, the 3D poses are obtained by labeling 2D poses and their corresponding depth values. Finally, 3D poses for Insta360 cameras are acquired by the transformation matrix of Insta360 camera and LiDAR camera. We provide 3D pose annotations for up to 3 subjects in this dataset.

The advantage of our proposed F-M3DHPE dataset is that our source frames provide different views from one insta360 camera, for the participants, from the panorama cameras, *i.e.,* Insta360 cameras, during collection. Each Insta360 camera contains two fisheye lenses. We provide images taken from one of the fisheye cameras in our dataset. We can project the spherical panorama at any camera angle within a 360-degree field of view to obtain fisheye images that looks like it is taken by a single fisheye camera as shown in Fig. 8. In other words, our dataset can be extended to the multi-view set-ups. In this paper, we only analyze images captured by one fishsye lens from the Insta360 camera.

Table 4
Quantitative results on MuPoTS-3D dataset with the metric of 3DPCK for root-relative 3D poses.

Methods	$3DPCK_{All}$ ↑	$3DPCK_{Matched}$ ↑
Bottom-up approaches		
Mehta et al. [38]	65.0	69.8
Mehta et al. [98]	70.4	75.8
Zhen et al. [36]	80.5	773.5
Top-down approaches		
Rogez et al. [41]	53.8	62.4
Rogez et al. [42]	70.6	74.0
Dabral et al. [43]	71.3	74.2
Moon et al. [44]	81.8	82.5
Wang et al. [47]	82.0	
Chen et al. [99]	87.5	–
Cheng et al. [50]	89.6	–
Lin and Lee [46]	–	83.7

$3DPCK_{all}$ means the results are computed by all annotated poses, while $3DPCK_{Matched}$ are calculated by matched poses (Matched).

Table 5
Quantitative results of existing methods on MuPoTS-3D dataset for absolute multi-person 3D poses.

Methods	Input	$3DPCK_{Matched}$ ↑	$3DPCK_{All}$ ↑
Learning-based methods			
Moon et al. [44]	A single image	31.8	31.5
Zhen et al. [36]	A single image	38.7	35.4
Wang et al. [47]	A single image	–	43.8
Lin and Lee [46]	A single image	35.2	–
Chen et al. [86]	Videos	45.7	–
Cheng et al. [50]	Videos	48.0	–

(2) Evaluation Metrics

Mean Per Joint Position Error (MPJPE). MPJPE is commonly used for the evaluation of 3D HPE. MPJPE is calculated by the Euclidean distance between predicted 3D poses and the ground truth after aligning the root joint, as defined by:

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{J}_i^{rel} - \mathbf{J}_i^{rel*}\right\|_2, \tag{5}$$

Fig. 6. Example images of our F-M3DHPE dataset for multi-person 3D pose estimation. The actions from left to right are Conversation, Taking photo, Phoning, Walking and Stretching. The first row and second row represent two viewpoints for the same pose.

Table 6
The quantitative results on the proposed F-M3DHPE dataset.

Methods	Conversation	Direction	Eating	Greeting	Phoning	Photo	Posing	Purchasing	Stretching	Walking	MPJPE ↓	MRPE ↓
Zhen et al. [36]	574.85	554.99	548.39	657.46	536.91	606.70	690.47	603.03	782.33	594.05	616.23	5662.76
Moon et al. [44]	476.44	446.46	514.00	603.70	509.52	570.30	639.20	553.47	743.41	547.08	560.99	1655.50
Lin and Lee [46]	476.44	446.46	514.00	603.70	509.52	570.30	639.20	553.47	743.41	547.08	560.99	427.02
Zhang et al. [4]	300.39	247.72	307.26	411.14	324.31	393.01	464.37	397.05	573.02	380.21	381.19	384.90

where N denotes the number of test frames, i represents the index of the samples, $\mathbf{J}$ and $\mathbf{J}^*$ indicate the predicted and ground-truth joint locations, *rel* means the joint locations are root-relative. In general, MPJPE is measured in millimeters.

Procrustes Analysis Mean Per Joint Position Error (PA MPJPE). PA MPJPE is the MPJEP after further alignment, *i.e.,* the Procrustes transformation.

Mean of the root position error (MRPE). MRPE is the metric proposed by Moon et al. [44] for absolute root joint estimation. Given predicted and ground-truth root joint locations in the camera coordinate $\mathbf{R}$ and $\mathbf{R}^*$, MRPE is as follows:

$$MRPE = \frac{1}{N}\sum_{i=1}^{N} \left\| \mathbf{R}_i - \mathbf{R}_i^* \right\|_2. \tag{6}$$

3D Percentage of Correct Keypoints (3DPCK). 3DPCK is used in the MPI-INF-3DHP and MuPoTS-3D datasets as the evaluation metric. Specifically, the estimated 3D joint is considered as correct if the distance between predictions and ground truths is within a threshold. Following [96], the threshold is equal to 150 mm.

Area under the Curve (AUC). AUC is an another metric for MPI-INF-3DHP dataset. AUC is defined as the area under the 3DPCK curve with different thresholds.

(3) Evaluation on Public dataset

We summarize the performance of existing methods for multi-person 3D human pose estimation on the MuPoTS-3D dataset. Particularly, the comparison for root-relative multi-person 3D pose estimation is shown in Table 4. Table 5 provides the comparison of absolute depth estimation for multi-person 3D poses.

In general, top-down approaches achieve better performance than bottom-up approaches. The reason is that top-down approaches divide the multi-person 3D pose estimation into single-person 3D person estimation for each detected person making the task simpler. However, absolute depth estimation is more challenging since the input image is cropped and hence ignores the relation between each person. Bottom-up approaches are faster than top-down approaches, because the former directly predicts 2D keypoints and absolute depths of all humans. However, the main weakness of bottom-up approaches is that the accuracy tends to be lower than top-down approaches.

Moon et al. [44] use MRPE on Human3.6M dataset to demonstrate that learning-based methods perform better to estimate the absolute depth than optimization-based methods.

(4) Evaluation on our F-M3DHPE dataset

In contrast to existing datasets, our dataset is captured from Insta360 cameras consisting of two fisheye lenses to capture 3D human poses. Depending on the availability of code, we evaluate four methods for 3D multi-person pose estimation from perspective cameras [36,44, 46] and one method for egocentric 3D pose estimation [58] from a fisheye camera on our dataset. We choose images containing six specific subjects as the training set. The remaining images are used as the test set.

Table 6 lists the experimental results of the existing methods on the F-M3DHPE dataset. It is shown that all methods perform similarly for root-relative 3D human pose estimation. However, the method that considers distortion parameters outperforms other methods. Note that the distortion parameters are not provided in the evaluation. Therefore, taking image distortions into account contributes positively to the performance. On the other hand, bottom-up methods provide worse performance compared to top-down methods. The reason is that 2D-to −3D lifting module follows the same mapping relation to obtain all 3D poses from the detected 2D keypoints for bottom-up methods. However, 2D keypoints located in different regions may suffer from different levels of distortion.

5.2. Dataset and benchmarks for action recognition

In this section, we describe public datasets for human action recognition and our newly dataset F-HAR captured by a fisheye camera (see Table 7). Then, we analyze the performance of existing methods on these
datasets.

(1) Datasets for human action recognition

There are two types of datasets for human action recognition, *i.e.,* with/without skeletons and depths data. In general, datasets without skeletons are usually used for video-based methods, while datasets with skeletons and or depth maps are used for skeleton-based methods. Particularly, existing datasets are captured by perspective cameras. In this paper, we propose a new dataset for human action recognition captured by a fisheye camera.

Kinetics 400/600/700 and Kinetics-Skeleton. Kinetics dataset consists of 400/600/700 human action categories and more than 400 videos per class downloaded from YouTube. This dataset covers different human actions including human–object interactions and human–human interactions. Kinetics-Skeleton is an extension of Kinetics 400

Table 7
Public datasets for human action recognition.

Dataset	Year	Camera type	Modality	Size	Action
HMDB51 [100]	2011	Pin-hole	Image	6766	51
UCF101 [101]	2012	Pin-hole	Image	13,320	101
Sports-1M [102]	2014	Pin-hole	Image	1,133,158	487
ActivityNet [103]	2015	Pin-hole	Image	28,000	203
YouTube-8M [102]	2016	Pin-hole	Image	8,000,000	4716
Kinetics 400 [104]	2017	Pin-hole	Image	306,245	400
Kinetics 600 [105]	2018	Pin-hole	Image	495,547	600
Kinetics 700 [106]	2019	Pin-hole	Image	650,317	700
MSR-Action3D [107]	2012	Pin-hole	Skeleton	567	20
UTKinect-Action3D [108]	2012	Pin-hole	Image + Skeleton	200	10
MSRDailyActivity3D [109]	2012	Pin-hole	Image + Skeleton	320	16
Northwestern-UCLA [110]	2014	Pin-hole	Image + Skeleton	1475	10
SYSU-3D [111]	2015	Pin-hole	Image + Skeleton	480	12
NTU RGB+D [112]	2016	Pin-hole	Image + Skeleton	56,880	60
Kinetics-Skeleton [113]	2017	Pin-hole	Image + Skeleton	266,440	400
NTU RGB+D 120 [114]	2019	Pin-hole	Image + Skeleton	114,480	120
F-HAR(Our dataset)	2022	Fisheye	Image + Skeleton	1000	16

Fig. 7. Example images of the proposed F-HAR dataset for human action recognition. 13 participants perform 10 actions in 5 indoor scenes and 5 outdoor scenes.

dataset. Due to no skeleton information in Kinetics datasets, off-the-shelf 2D detectors are used to obtain 2D poses for skeleton-based methods for human action recognition.

NTU RGB+D and NTU RGB+D 120. NTU RGB+D and NTU RGB+D 120 are large scale datasets with depth information for action recognition. They consists of 56,880 and 114,480 clips captured in an indoor environment, respectively. NTU RGB+D 120 includes 120 action classes, while NTU RGB+D comprises of 60 action classes. There are three types, *i.e.,* daily actions, health-related actions and mutual actions in the datasets. Both datasets include RGB videos with the resolution of 1920×1080 pixels, depth map with the resolution of 512×512, 3D skeletons with 25 body joints.

F-HAR. The proposed F-HAR dataset is captured by an Insta360 camera, consisting two fisheye lenses, capturing 10 scenes with both indoor and outdoor environments as shown in Fig. 7. There are 13 participants (7 males and 6 females), and each participant performs actions in five scenes of an indoor or outdoor environment. Furthermore, 7 participants are present in all indoor scenes while 6 participants are in all outdoor scenes. There are 14 action categories consisting of taking off coat/backpack, wearing coat/backpack, walking, talking, moving, phoning, drinking water, using a computer, taking photos, waving hands, writing, reading, clapping, and eating. Our dataset consists of 1000 sequences. Each sequences includes one action class lasting between 3 to 10 s. The videos are taken at 30fps with the resolution of 1920×1920 pixels. Labels, camera information and skeletons are provided. Skeletons are generated by an off-the-shelf model [115].

Different from previous datasets, F-HAR dataset provides a fisheye view to perform human action recognition. Moreover, F-HAR provides different viewpoints as shown in Fig. 8 because of the panorama cameras as described in Section 5. Therefore, F-HAR is beneficial for (1) human action recognition from a single image captured by a fisheye camera; (2) multi-view human action recognition; and (3) human action recognition from a single panorama image. In this survey paper, we only focus on videos captured by one fisheye lens from the Insta360 camera.

(2) Evaluation Metrics

The mean accuracy is employed to assess methods for HAR at the sequence level. It is specifically defined as the proportion of correctly predicted actions in the sequences to the total number of sequences.

(3) Evaluation on public datasets

We summarize the performance of existing skeleton-based methods on the commonly used NTU, NTU120 and Kinetics-Skeleton datasets. For the NTU and NTU120 datasets, existing methods use 3D information (2D poses with depth maps or 3D joint locations). In terms of Kinetics-Skeleton dataset, the sequences of detected 2D skeletons are used for action classification. The experimental results are shown in Table 8. GCN-based methods tend to perform better than other two kinds of methods. Instead of 3D skeletons, GCN-based methods can also use 2D skeletons for human action recognition. GCN-based methods have the advantage to exploit relationships between nodes presented by the 2D keypoints in the skeleton-based methods for human action recognition.

(4) Evaluation on our F-HAR dataset

Depending on the availability of code, four methods are used to compare the performance on our dataset for human action recognition using a fisheye camera. Based on whether the training set includes both indoor and outdoor environments, there are two split protocols including F-HAR-CS and F-HAR-CP. Specifically, F-HAR-CS represents the training set that only includes videos captured in indoor or outdoor environments. F-HAR-CP indicates the training set consisting of both

Table 8
The quantitative results of existing methods for human action recognition on public datasets.

Methods	NTU CS ↑	NTU CV ↑	NTU120 CS↑	NTU120 CV ↑	KS Top-1 ↑	KS Top-5 ↑
CNN-based methods						
SkeleMotion [67]	76.5	84.7	67.7	66.9	–	–
Chained Net [63]	80.8	–	–	–	–	–
Deep Bilinear [64]	85.4	90.7	–	–	–	–
2D/3D pose [65]	85.5	–	–	–	–	–
Posemaps [66]	91.7	95.3	64.6	66.9	–	–
Else-net [69]	91.6	96.4	–	–	–	–
LSTM-based methods						
GCA-LSTM [74]	76.1	84.0	61.2	63.3	–	–
Ind-RNN [75]	81.8	88.0	–	–	–	–
Geometric [78]	97.0	98.5	90.6	86.7	–	–
GCN-based methods						
ST-GCN [60]	81.5	88.3	–	–	30.7	52.8
SR-TSL [87]	84.8	92.4	–	–	–	–
AS-GCN [80]	86.8	94.2	–	–	34.8	56.5
2S-AGCN [116]	88.5	95.1	–	–	36.1	58.7
SGN [82]	89.0	94.5	79.2	81.5	–	–
AGC-LSTM [84]	89.2	95.0	–	–	–	–
MS-G3D [81]	91.5	96.2	86.9	88.4	38.0	60.9
Channel-wise [117]	92.4	96.8	88.9	90.6	–	–

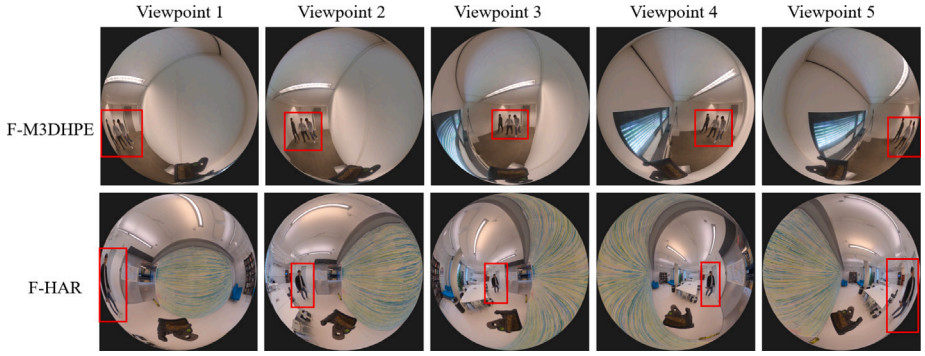

Fig. 8. The frames in F-M3DHPE and F-HAR datasets can be viewed from different angles in one Insta360 camera. It can be seen that the subject at the edge of the image suffers from large distortions, causing 3D HPE and HAR more challenging.

Table 9
Quantitative results of existing methods on our F-HAR dataset for human action recognition.

Methods	F-HAR-CS ↑	F-HAR-CP ↑
ST-GCN [60]	12.21%	23.31%
AS-GCN [80]	10.57%	6.42%
2s-AGCN [116]	19.14%	30.74%
MS-G3D [81]	53.16%	70.27%

indoor and outdoor environments using videos performed by half the number of subjects as the training set. The experimental results are shown in Table 9. The method proposed by Liu et al. [81] outperforms the other three methods on both protocols. The reason is that this method alleviates the biased weighting problem in GCNs and that it combines spatial–temporal connections.

6. Discussion and future perspectives

Although 3D human pose estimation from pin-hole cameras has garnered significant attention, the application of fisheye cameras in 3D HPE and human action recognition has not been fully explored. To address this gap, we present a comprehensive survey of 3D human pose estimation and human action recognition using both pin-hole and fisheye cameras. To facilitate further research in this area, we have introduced a novel dataset specifically collected using fisheye cameras, designed for multi-person 3D pose estimation and human action recognition. Our analysis evaluates existing methods on this proposed dataset, highlighting the advancements made in these fields over recent years while also identifying several persistent challenges. To tackle these challenges, we propose several potential future research directions in this section.

3D human pose estimation. Existing methods for 3D pose estimation mainly focus on perspective cameras, and there are only a few methods for multi-person 3D pose estimation using fisheye cameras. The method [4] exploiting a fisheye camera model to alleviate the negative influence caused by the fisheye lens outperforms other methods. From the experimental results of existing methods on the proposed dataset, the performance of multi-person 3D pose estimation is still far from perfect. The main challenge is that humans, located at different positions, suffer from different levels of distortions. Therefore, combining human positions, distortion parameters and other cues to regularize root-relative 3D poses and absolute depths is a promising direction for multi-person 3D pose estimation.

On the other hand, as there are several large-scale datasets available for 3D human pose estimation (HPE) using pin-hole cameras, effectively utilizing these datasets in conjunction with our proposed dataset captured by fisheye cameras represents a promising direction for advancing 3D HPE from fisheye perspectives. By combining existing datasets for 3D human pose estimation from perspective cameras with our fisheye dataset, we can leverage the extensive information embedded in the perspective camera datasets for pre-training. This pre-training phase can be followed by fine-tuning on the fisheye dataset, potentially enhancing the overall performance of 3D human pose estimation. However, the key challenge in this integration lies in bridging the gap between these two distinct types of datasets. These differences

can significantly impact how poses are represented and interpreted. Therefore, it is essential to develop innovative methodologies for effectively aligning data from both sources, which would help to realize the benefits of combining these datasets and improve pose estimation accuracy in real-world applications.

Another challenge in estimating single- or multi-person 3D poses from perspective or fisheye images is how to account for the inherent uncertainty of each joint. As only a single image is used as input, there exists an inherent ambiguity in the depth dimension for 3D pose estimation. Failure to account for this uncertainty may result in a degraded performance of the method for 3D human pose estimation. For 3D pose estimation from a perspective image, the uncertainty arises mainly from the visibility of each joint. In case of 3D pose estimation from a fisheye image, the uncertainty is due to the visibility of each joint, the image distortion strength, and the human position at the image space. Additionally, there is minor difference in the definition of human joint across different datasets, *e.g.,* Human3.6M [95] and MPI-INF-3DHP [96]. Thus, directly using them to train a model for 3D human pose estimation could be detrimental to its accuracy and performance. As a result, addressing the uncertainty issue in 3D human pose estimation is a propitious research direction.

Human action recognition. Human poses are indeed essential for human action recognition, but obtaining accurate 2D or 3D ground-truth skeletons can be challenging, particularly for in-the-wild images or videos where the subjects may be occluded, in motion, or in varying lighting conditions. To this end, a common strategy is to use off-the-self 2D/3D estimators to obtain 2D/3D skeletons. Although skeleton-based methods for human action recognition have made progress, they often rely on using either 2D or 3D skeletons alone. 2D skeletons are useful at capturing the spatial relationships between joints, while 3D skeletons provide more accurate information about the motion in 3D space. Combining these two types of poses can potentially improve the accuracy of action recognition models. Additionally, there is the possibility of combining pose data with other modalities, such as RGB images or depth maps, to capture complementary information about the action. These are interesting research topics to explore in the future to advance the field of human action recognition.

Unsupervised learning represents a promising avenue for human action recognition using fisheye cameras. Unlike supervised learning, which depends on annotated training data, unsupervised learning methods can automatically learn data representations without requiring explicit labels. By leveraging unsupervised skeleton detection, models can be trained to recognize patterns in 2D and 3D representations of human motion without being restricted to specific tasks or datasets. Furthermore, the use of estimated 2D/3D skeletons can help mitigate the negative effects of image distortions associated with fisheye lenses. This capability may enhance model performance and improve generalization to new tasks and datasets, which is particularly valuable in applications characterized by diverse and challenging input data. The significant challenge remains in accurately identifying representative keypoints of human subjects in an unsupervised manner to effectively indicate human actions.

Limitation. Note that our survey has some limitations. Firstly, our primary focus is on 3D human pose estimation and hence paying less attention to 3D human body reconstruction. Secondly, the focus is more on skeleton-based HAR methods than video-based methods. This survey and recent work [118,119] complement each other, filling in gaps and contributing to a more complete understanding of the more general topic of 3D human pose estimation and action recognition. On the other hand, our collected dataset does not explicitly account for the offset between the skeleton and skin, which is a limitation of our dataset. However, given that the offset from the skeleton to the skin is significantly smaller compared to the distance between each joint and the camera (>1.5 m), we expect that this difference has minimal impact on the overall performance of 3D human pose estimation. In future work, we plan to address this issue by exploring methods to accurately account for the skeleton-to-skin offset, which could further enhance pose estimation accuracy. For instance, incorporating more cameras to capture human poses from multiple angles, combined with optimization techniques, could help refine the labeled 3D poses and better account for this offset.

CRediT authorship contribution statement

Yahui Zhang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Shaodi You:** Methodology, Resources, Supervision, Validation, Writing – review & editing. **Sezer Karaoglu:** Resources, Supervision, Writing – review & editing. **Theo Gevers:** Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] J. Jayanthi, P.U. Maheswari, AI and augmented reality for 3D Indian dance pose reconstruction cultural revival, Sci. Rep. 14 (1) (2024) 7906.
[2] J. Kwon, M. Sa, H. Kim, Y. Seong, C.J. Lee, Egocentric 3D skeleton learning in a deep neural network encodes obese-like motion representations, Exp. Neurobiol. 33 (3) (2024) 119.
[3] J. Celestino, M. Marques, J.C. Nascimento, J.P. Costeira, 2D Image head pose estimation via latent space regression under occlusion settings, Pattern Recognit. 137 (2023) 109288.
[4] Y. Zhang, S. You, S. Karaoglu, T. Gevers, Multi-person 3D pose estimation from a single image captured by a fisheye camera, Comput. Vis. Image Underst. 222 (2022) 103505.
[5] Y. Chen, Y. Tian, M. He, Monocular human pose estimation: A survey of deep learning-based methods, Comput. Vis. Image Underst. (2020) 102897.
[6] E.S. dos Reis, L.A. Seewald, R.S. Antunes, V.F. Rodrigues, R. da Rosa Righi, C.A. da Costa, L.G. da Silveira Jr., B. Eskofier, A. Maier, T. Horz, et al., Monocular multi-person pose estimation: A survey, Pattern Recognit. (2021) 108046.
[7] B. Ren, M. Liu, R. Ding, H. Liu, A survey on 3d skeleton-based action recognition using learning method, Cyborg Bionic Syst. 5 (2024) 0100.
[8] F.A. Van Den Heuvel, R. Verwaal, B. Beers, Calibration of fisheye camera systems and the reduction of chromatic aberration, in: ISPRS Comission V Symposium, Citeseer, 2006, pp. 25–27.
[9] I. Habibie, W. Xu, D. Mehta, G. Pons-Moll, C. Theobalt, In the wild human pose estimation using explicit 2d features and intermediate 3d representations, in: CVPR, 2019, pp. 10905–10914.
[10] S. Li, A.B. Chan, 3D human pose estimation from monocular images with deep convolutional neural network, in: ACCV, Springer, 2014, pp. 332–347.
[11] S. Li, W. Zhang, A.B. Chan, Maximum-margin structured learning with deep networks for 3d human pose estimation, in: ICCV, 2015, pp. 2848–2856.
[12] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, P. Fua, Structured prediction of 3d human pose with deep neural networks, in: BMVC, 2016.
[13] X. Sun, J. Shang, S. Liang, Y. Wei, Compositional human pose regression, in: ICCV, 2017, pp. 2602–2611.
[14] G. Pavlakos, X. Zhou, K.G. Derpanis, K. Daniilidis, Coarse-to-fine volumetric prediction for single-image 3D human pose, in: CVPR, 2017, pp. 7025–7034.
[15] G. Pavlakos, X. Zhou, K. Daniilidis, Ordinal depth supervision for 3d human pose estimation, in: CVPR, 2018, pp. 7307–7316.
[16] C.-H. Chen, D. Ramanan, 3D human pose estimation=2d pose estimation+ matching, in: CVPR, 2017, pp. 7035–7043.
[17] F. Moreno-Noguer, 3D human pose estimation from a single image via distance matrix regression, in: CVPR, 2017, pp. 2823–2832.
[18] B. Tekin, P. Márquez-Neila, M. Salzmann, P. Fua, Learning to fuse 2d and 3d image cues for monocular body pose estimation, in: ICCV, 2017, pp. 3941–3950.
[19] J. Martinez, R. Hossain, J. Romero, J.J. Little, A simple yet effective baseline for 3d human pose estimation, in: ICCV, 2017, pp. 2640–2649.
[20] X. Sun, B. Xiao, F. Wei, S. Liang, Y. Wei, Integral human pose regression, in: ECCV, 2018, pp. 529–545.

[21] K. Zhou, X. Han, N. Jiang, K. Jia, J. Lu, Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation, in: ICCV, 2019, pp. 2344–2353.

[22] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, C. Lu, Human pose regression with residual log-likelihood estimation, in: ICCV, 2021, pp. 11025–11034.

[23] Z. Chen, Y. Huang, H. Yu, L. Wang, Learning a robust part-aware monocular 3D human pose estimator via neural architecture search, Int. J. Comput. Vis. (2022) 56–75.

[24] Q. Nie, Z. Liu, Y. Liu, Lifting 2D human pose to 3D with domain adapted 3D body concept, Int. J. Comput. Vis. (2023) 1–19.

[25] H. Ci, M. Wu, W. Zhu, X. Ma, H. Dong, F. Zhong, Y. Wang, Gfpose: Learning 3d human pose prior with gradient fields, in: CVPR, 2023, pp. 4800–4810.

[26] H. Ci, C. Wang, X. Ma, Y. Wang, Optimizing network structure for 3d human pose estimation, in: ICCV, 2019, pp. 2262–2271.

[27] L. Zhao, X. Peng, Y. Tian, M. Kapadia, D.N. Metaxas, Semantic graph convolutional networks for 3d human pose regression, in: CVPR, 2019, pp. 3425–3435.

[28] K. Liu, R. Ding, Z. Zou, L. Wang, W. Tang, A comprehensive study of weight sharing in graph networks for 3d human pose estimation, in: ECCV, Springer, 2020, pp. 318–334.

[29] T. Xu, W. Takano, Graph stacked hourglass networks for 3D human pose estimation, in: CVPR, 2021, pp. 16105–16114.

[30] W. Li, M. Liu, H. Liu, T. Guo, T. Wang, H. Tang, N. Sebe, GraphMLP: A graph MLP-like architecture for 3D human pose estimation, Pattern Recognit. 158 (2025) 110925.

[31] C. Han, X. Yu, C. Gao, N. Sang, Y. Yang, Single image based 3D human pose estimation via uncertainty learning, Pattern Recognit. 132 (2022) 108934.

[32] X. Zhou, Q. Huang, X. Sun, X. Xue, Y. Wei, Towards 3d human pose estimation in the wild: a weakly-supervised approach, in: ICCV, 2017, pp. 398–407.

[33] S. Tripathi, L. Müller, C.-H.P. Huang, O. Taheri, M.J. Black, D. Tzionas, 3D human pose estimation via intuitive physics, in: CVPR, 2023, pp. 4713–4725.

[34] B. Wandt, B. Rosenhahn, Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation, in: CVPR, 2019, pp. 7782–7791.

[35] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, S. Stojanov, J.M. Rehg, Unsupervised 3d pose estimation with geometric self-supervision, in: CVPR, 2019, pp. 5714–5724.

[36] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao, X. Zhou, Smap: Single-shot multi-person absolute 3d pose estimation, in: ECCV, 2020, pp. 550–566.

[37] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, C. Sminchisescu, Deep network for the integrated 3d sensing of multiple people in natural images, Neural Inf. Process. Syst. (2018) 8410–8419.

[38] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, C. Theobalt, Single-shot multi-person 3d pose estimation from monocular rgb, in: 3DV, IEEE, 2018, pp. 120–130.

[39] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: CVPR, 2017, pp. 7291–7299.

[40] S. Yang, Z. Feng, Z. Wang, Y. Li, S. Zhang, Z. Quan, S.-t. Xia, W. Yang, Detecting and grouping keypoints for multi-person pose estimation using instance-aware attention, Pattern Recognit. 136 (2023) 109232.

[41] G. Rogez, P. Weinzaepfel, C. Schmid, Lcr-net: Localization-classification-regression for human pose, in: CVPR, 2017, pp. 3433–3441.

[42] G. Rogez, P. Weinzaepfel, C. Schmid, Lcr-net++: Multi-person 2d and 3d pose detection in natural images, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1146–1161.

[43] R. Dabral, N.B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, A. Jain, Multi-person 3d human pose estimation from monocular images, in: 3DV, IEEE, 2019, pp. 405–414.

[44] G. Moon, J.Y. Chang, K.M. Lee, Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image, in: ICCV, 2019, pp. 10133–10142.

[45] Y. Guo, L. Ma, Z. Li, X. Wang, F. Wang, Monocular 3D multi-person pose estimation via predicting factorized correction factors, Comput. Vis. Image Underst. (2021) 103278.

[46] J. Lin, G.H. Lee, Hdnet: Human depth estimation for multi-person camera-space localization, in: ECCV, 2020, pp. 633–648.

[47] C. Wang, J. Li, W. Liu, C. Qian, C. Lu, Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation, in: ECCV, 2020, pp. 242–259.

[48] R. Cucchiara, M. Fabbri, Fine-grained human analysis under occlusions and perspective constraints in multimedia surveillance, Trans. Multimed. (2022) 1–23.

[49] Z. Qiu, K. Qiu, J. Fu, D. Fu, Weakly-supervised pre-training for 3D human pose estimation via perspective knowledge, Pattern Recognit. 139 (2023) 109497.

[50] Y. Cheng, B. Wang, B. Yang, R.T. Tan, Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks, in: CVPR, 2021, pp. 7649–7659.

[51] G. Rogez, J.S. Supancic, D. Ramanan, First-person pose recognition using egocentric workspaces, in: CVPR, 2015, pp. 4325–4333.

[52] H. Yonemoto, K. Murasaki, T. Osawa, K. Sudo, J. Shimamura, Y. Taniguchi, Egocentric articulated pose tracking for action recognition, in: MVA, IEEE, 2015, pp. 98–101.

[53] H. Jiang, K. Grauman, Seeing invisible poses: Estimating 3d body pose from egocentric video, in: CVPR, IEEE, 2017, pp. 3501–3509.

[54] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, C. Theobalt, Egocap: egocentric marker-less motion capture with two fisheye cameras, ACM Trans. Graph. (2016) 1–11.

[55] T. Shiratori, H.S. Park, L. Sigal, Y. Sheikh, J.K. Hodgins, Motion capture from body-mounted cameras, in: ACM SIGGRAPH 2011 Papers, 2011, pp. 1–10.

[56] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H.-P. Seidel, C. Theobalt, Mo2Cap2: Real-time mobile 3D motion capture with a cap-mounted fisheye camera, IEEE Trans. Vis. Comput. Graphics (2019) 2093–2101.

[57] D. Tome, P. Peluse, L. Agapito, H. Badino, xR-EgoPose: Egocentric 3D human pose from an HMD camera, in: ICCV, 2019, pp. 7728–7738.

[58] Y. Zhang, S. You, T. Gevers, Automatic calibration of the fisheye camera for egocentric 3D human pose estimation from a single image, in: WACV, 2021, pp. 1772–1781.

[59] A. Dhamanaskar, M. Dimiccoli, E. Corona, A. Pumarola, F. Moreno-Noguer, Enhancing egocentric 3D pose estimation with third person views, Pattern Recognit. 138 (2023) 109358.

[60] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI, 2018.

[61] Y. Du, Y. Fu, L. Wang, Skeleton based action recognition with convolutional neural network, in: ACPR, IEEE, 2015, pp. 579–583.

[62] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognit. (2017) 346–362.

[63] M. Zolfaghari, G.L. Oliveira, N. Sedaghat, T. Brox, Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection, in: ICCV, 2017, pp. 2904–2913.

[64] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, J. Zhang, Deep bilinear learning for rgb-d action recognition, in: ECCV, 2018, pp. 335–351.

[65] D.C. Luvizon, D. Picard, H. Tabia, 2D/3d pose estimation and action recognition using multitask deep learning, in: CVPR, 2018, pp. 5137–5146.

[66] M. Liu, J. Yuan, Recognizing human actions as the evolution of pose estimation maps, in: CVPR, 2018, pp. 1159–1168.

[67] C. Caetano, J. Sena, F. Brémond, J.A. Dos Santos, W.R. Schwartz, Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition, in: AVSS, IEEE, 2019.

[68] Y. Chen, L. Wang, C. Li, Y. Hou, W. Li, ConvNets-based action recognition from skelaeton motion maps, Multimedia Tools Appl. (2020) 1707–1725.

[69] T. Li, Q. Ke, H. Rahmani, R.E. Ho, H. Ding, J. Liu, Else-net: Elastic semantic network for continual action recognition from skeleton data, in: ICCV, 2021, pp. 13434–13443.

[70] U. Asif, D. Mehta, S. Von Cavallar, J. Tang, S. Harrer, DeepActsNet: A deep ensemble framework combining features from face, hands, and body for action recognition, Pattern Recognit. (2023) 109484.

[71] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: CVPR, 2022, pp. 2969–2978.

[72] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: CVPR, 2015, pp. 1110–1118.

[73] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: CVPR, 2016, pp. 1010–1019.

[74] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, A.C. Kot, Skeleton-based human action recognition with global context-aware attention LSTM networks, IEEE Trans. Image Process. (2017) 1586–1599.

[75] S. Li, W. Li, C. Cook, C. Zhu, Y. Gao, Independently recurrent neural network (indrnn): Building a longer and deeper rnn, in: CVPR, 2018, pp. 5457–5466.

[76] R. Zhao, K. Wang, H. Su, Q. Ji, Bayesian graph convolution lstm for skeleton based action recognition, in: ICCV, 2019, pp. 6882–6892.

[77] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive neural networks for high performance skeleton-based human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1963–1978.

[78] R. Friji, H. Drira, F. Chaieb, H. Kchok, S. Kurtek, Geometric deep neural network using rigid and non-rigid transformations for human action recognition, in: ICCV, 2021, pp. 12611–12620.

[79] S.K. Yadav, K. Tiwari, H.M. Pandey, S.A. Akbar, Skeleton-based human activity recognition using ConvLSTM and guided feature learning, Soft Comput. (2022) 1–14.

[80] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in: CVPR, 2019, pp. 3595–3603.

[81] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: CVPR, 2020, pp. 143–152.

[82] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: CVPR, 2020, pp. 1112–1121.

[83] M. Dai, Z. Sun, T. Wang, J. Feng, K. Jia, Global spatio-temporal synergistic topology learning for skeleton-based action recognition, Pattern Recognit. 140 (2023) 109540.

[84] C. Si, W. Chen, W. Wang, L. Wang, T. Tan, An attention enhanced graph convolutional lstm network for skeleton-based action recognition, in: CVPR, 2019, pp. 1227–1236.
[85] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, S. Xia, Graph CNNs with motif and variable temporal block for skeleton-based action recognition, in: AAAI, 2019, pp. 8989–8996.
[86] Y. Chen, G. Ma, C. Yuan, B. Li, H. Zhang, F. Wang, W. Hu, Graph convolutional network with structure pooling and joint-wise channel attention for action recognition, Pattern Recognit. (2020) 107321.
[87] C. Si, Y. Jing, W. Wang, L. Wang, T. Tan, Skeleton-based action recognition with spatial reasoning and temporal stack learning, in: ECCV, 2018, pp. 103–118.
[88] C. Ding, K. Liu, J. Korhonen, E. Belyaev, Spatio-temporal difference descriptor for skeleton-based action recognition, in: AAAI, 2021, pp. 1227–1235.
[89] Y. Liu, H. Zhang, D. Xu, K. He, Graph transformer network with temporal kernel attention for skeleton-based action recognition, Knowl.-Based Syst. 240 (2022) 108146.
[90] J. Kong, S. Wang, M. Jiang, T. Liu, Multi-stream ternary enhanced graph convolutional network for skeleton-based action recognition, Neural Comput. Appl. 35 (25) (2023) 18487–18504.
[91] K. Gedamu, Y. Ji, L. Gao, Y. Yang, H.T. Shen, Relation-mining self-attention network for skeleton-based human action recognition, Pattern Recognit. 139 (2023) 109455.
[92] Y.-F. Song, Z. Zhang, L. Wang, Richly activated graph convolutional network for action recognition with incomplete skeletons, in: ICIP, IEEE, 2019, pp. 1–5.
[93] L. Yu, L. Tian, Q. Du, J.A. Bhutto, Multi-stream adaptive 3d attention graph convolution network for skeleton-based action recognition, Appl. Intell. 53 (12) (2023) 14838–14854.
[94] L. Sigal, A.O. Balan, M.J. Black, Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Int. J. Comput. Vis. (2010) 4.
[95] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. (2013) 1325–1339.
[96] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3d human pose estimation in the wild using improved cnn supervision, in: 3DV, IEEE, 2017, pp. 506–516.
[97] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, Y. Sheikh, Panoptic studio: A massively multiview system for social motion capture, in: ICCV, 2015, pp. 3334–3342.
[98] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, C. Theobalt, XNect: Real-time multi-person 3D motion capture with a single RGB camera, ACM Trans. Graph. (2020) 82–1.
[99] Y. Cheng, B. Wang, B. Yang, R.T. Tan, Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos, in: AAAI, 2021, pp. 1157–1165.
[100] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: ICCV, IEEE, 2011, pp. 2556–2563.
[101] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, 2012, arXiv preprint arXiv:1212.0402.
[102] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: CVPR, 2014, pp. 1725–1732.
[103] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: CVPR, 2015, pp. 961–970.
[104] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, 2017, arXiv preprint arXiv:1705.06950.
[105] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600, 2018, arXiv preprint arXiv:1808.01340.
[106] J. Carreira, E. Noland, C. Hillier, A. Zisserman, A short note on the kinetics-700 human action dataset, 2019, arXiv preprint arXiv:1907.06987.
[107] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: CVPR Workshops, IEEE, 2010, pp. 9–14.
[108] L. Xia, C.-C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: CVPR Workshops, IEEE, 2012, pp. 20–27.
[109] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: CVPR, IEEE, 2012, pp. 1290–1297.
[110] J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning and recognition, in: CVPR, 2014, pp. 2649–2656.
[111] J.-F. Hu, W.-S. Zheng, J. Lai, J. Zhang, Jointly learning heterogeneous features for RGB-D activity recognition, in: CVPR, 2015, pp. 5344–5352.
[112] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: CVPR, 2016, pp. 1010–1019.
[113] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI, 2018.
[114] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 2684–2701.
[115] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, C. Lu, Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, in: CVPR, 2019, pp. 10863–10872.
[116] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: CVPR, 2019, pp. 12026–12035.
[117] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: ICCV, 2021, pp. 13359–13368.
[118] V. Sharma, M. Gupta, A.K. Pandey, D. Mishra, A. Kumar, A review of deep learning-based human activity recognition on benchmark video datasets, Appl. Artif. Intell. 36 (1) (2022) 2093705.
[119] M. Zaka-Ud-Din, Z. Huang, R. Khan, A review of 3D human body pose estimation and mesh recovery, Digit. Signal Process. (2022) 103628.

Yahui Zhang received with the M.Sc. degree from Huazhong University of Science and Technology PR China in 2018, and the Ph.D. degree from University of Amsterdam, Amsterdam, The Netherlands in 2022. He is a postdoc now at Northwestern University, USA. His research focuses on 3D human pose estimation, photometric stereo, and motion transfer.

Shaodi You is an associate professor at University of Amsterdam, Netherlands. He receives his Ph.D. and M.E. degrees from The University of Tokyo, Japan in 2015 and 2012 and his bachelor' s degree from Tsinghua University, P. R. China in 2009. His research interests are physics based vision and deep learning. He serves as area chair of CVPR, ICCV and Neurips regularly. He is an associate editor of IJCV.

Sezer Karaoglu received the double M.Sc. degrees in optics image vision from the Université Jean Monnet, Saint-Etienne, France, and in multimedia technologies from Gjvik University College, Gjvik, Norway, in 2011, and the Ph.D. degree from the Computer Vision Group, University of Amsterdam, Amsterdam, The Netherlands, in 2016. He is the CTO and Co-Founder of 3DUniversum, a spin-off of the Informatics Institute, UvA. His research interests include 2D-3D computer vision and machine learning.

Theo Gevers is a professor of computer vision with the University of Amsterdam. He is the Director of the Computer Vision Laboratory and Co-Director of the Atlas Laboratory in Amsterdam. He is the Co-Founder of AI technology companies 3DUniversum and Sightcorp. His research area focuses on artificial intelligence with the focus on computer vision and deep learning, and in particular image processing, 3D (object) understanding and human-behavior analysis. He has authored or coauthored more than 230 papers and three books.