Journal of Visual Communication and Image Representation (2021)

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Human Pose Estimation and Its Application to Action Recognition: A Survey

Liangchen Song[a], Gang Yu[b], Junsong Yuan[a], Zicheng Liu[c]

[a] *University at Buffalo*
[b] *Tencent*
[c] *Microsoft Research*

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Communicated by S. Sarkar

2000 MSC: 68T10, 68T45

Keywords: Pose estimation, Action recognition

ABSTRACT

Human pose estimation aims at predicting the poses of human body parts in images or videos. Since pose motions are often driven by some specific human actions, knowing the body pose of a human is critical for action recognition. This survey focuses on recent progress of human pose estimation and its application to action recognition. We attempt to provide a comprehensive review of recent bottom-up and top-down deep human pose estimation models, as well as how pose estimation systems can be used for action recognition. Thanks to the availability of commodity depth sensors like Kinect and its capability for skeletal tracking, there has been a large body of literature on 3D skeleton-based action recognition, and there are already survey papers such as [1] about this topic. In this survey, we focus on 2D skeleton-based action recognition where the human poses are estimated from regular RGB images instead of depth images. We summarize the performance of recent action recognition methods that use pose estimated from color images as input, then show that there is much room for improvements in this direction.

1. Introduction

Localizing human body joint positions in images or videos is the task of human pose estimation. By knowing the motion of a human body in a sequence of images, we can infer what the person is doing, that is, recognizing the action. In this sense, action recognition can be viewed as a natural application of human pose estimation. With the advance of deep neural networks, the last decade has witnessed great progress on human pose estimation techniques. Although there are articles reviewing recent human pose estimation methods [2, 3], the topic of human pose estimation and its application to action recognition has not been well summarized.

Human pose estimation can be used for action recognition, which does not mean that the two tasks are not correlated and action recognition is a down-stream task. As a matter of fact, action recognition could become a driving force for the continued improvement of human pose estimation. For example, recognizing sports activities requires a highly accurate estimation of skeletal joints. If we want to develop an explainable grading system for diving, the precise location of the body joints is needed for the system to provide a reasonable judgment on the diving performance.

As another example, some actions involve challenging poses that are difficult to infer by existing pose estimation models. Using the pose estimator trained on a dataset composed of normal poses may perform poorly for recognizing actions in acrobatics. This is because some of the acrobatics poses are not covered by the normal poses. An acrobat, for example, may walk with his/her arms. Thus even with recent progress of human pose estimation, there is still room to improve when we apply pose estimation into action recognition.

Meanwhile, there are actions involving some specific joints, such as spine joints and hand joints, that are not well covered by the current pose datasets. For example, distinguishing different facial makeup actions requires accurate hand pose as well as head pose. It is not a trivial task to integrate all these poses

into a single system and such a pose estimation system is undoubtedly valuable in practice.

In this survey, we investigate recent progress on pose estimation and will discuss how to use skeletons extracted from RGB videos for action recognition. We begin our survey by introducing common datasets used for evaluating pose estimation and action recognition. Next, we provide a comprehensive survey of human pose estimation methods, with emphasis on deep learning based models. Then, we review some recent works that use the pose estimated from RGB images for action recognition. Also, recent work on learning to simultaneously estimate poses and recognize actions will be discussed. Finally, we conclude by discussing the remaining challenges and future directions. In general, our survey contains three main contributions:

- We provide a comprehensive survey of recent human pose estimation methods, which are grouped into two categories: bottom-up framework based and top-down framework based.
- We review the action recognition algorithms that exploit the human body poses and demonstrate the performance gap of skeleton-based and video-based methods on different testing benchmarks.
- We discuss the reason why we need human pose based action recognition and how human pose estimation can be improved to be better applied to action recognition tasks.

2. Datasets

2.1. Datasets for human pose estimation

Many datasets have been released for evaluating human pose estimation algorithms in recent years. Generally speaking, the datasets are becoming larger and more challenging. In Tab. 1, we list the details of widely used 2D human pose benchmark datasets. Meanwhile, 3D human pose datasets are relatively less diverse due to the constraints of 3D pose capturing sensors. In Tab. 2, we list the details of three commonly used datasets for 3D human pose estimation.

2.2. 2D human pose estimation datasets

LSP. The Leeds Sports Pose (LSP) Dataset [4] contains 2,000 images of mostly sportspeople. The images are collected from Flickr with the following tags: athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball. The most challenging poses lie in the activities of gymnastics, parkour and to a lesser degree athletics. These images are then cropped and resized such that the most prominent person is roughly 150 pixels in height. Later, Johnson *et al.*[5] improve the dataset and collect 10,800 images from Flickr with the challenging activities labels. In total, 10,000 images are sampled from the accepted annotations.

FashionPose. Dantone *et al.*[6] collect images downloaded from a variety of fashion blogs. Each image contains a person where the full body is visible and is annotated by 12 joints and a point for the head, namely the nose. Different from LSP [4], the head is not annotated by the top of the head and the neck, because the top of the head and the neck are difficult to annotate accurately.

J-HMDB. J-HMDB [7] is short for joint-annotated HMDB [18], which is a dataset for action recognition. 928 clips comprising 21 action categories are extracted from HMDB51 and each frame is annotated with scale, pose, segmentation, coarse viewpoint, and dense optical flow for the humans in action.

FLIC. Images from FLIC [8] are collected from popular Hollywood movies. The person detector Poselets [19] is first adopted to detect people in every 10 frames. About 20K detection results of the highest confidence are then manually labeled. Finally, occluded, non-frontal, or just plain mislabeled images are rejected and the rest 5,003 images are used for the FLIC dataset. Note that only upper body keypoints are annotated in the FLIC dataset.

MPII. MPII [9] is the first large scale benchmark dataset for human pose estimation. The image frames are selected from YouTube videos. MPII Human Pose consists of about 25k images and contains around 40k total annotated people, of which three-quarters are available for training. It is worth noting that the MPII can be used for both single-person and multi-person pose estimation evaluation. The multi-person MPII consists of 3844 training and 1758 testing groups of person [20]. The groups are usually a subset of the total people in a particular image. For evaluation, the locations of 14 body parts are considered. Also, action labels are assigned to each video.

MSCOCO. MSCOCO [10] is originally constructed for object detection and instance segmentation in 2014. Later on ECCV 2016, human pose estimation is added to the challenge and OpenPose [21] wins the first place. In Table 1, the statistics shown are from the 2017 competition and kept unchanged for the following 2018 and 2019 competitions. 12 human body parts and 5 facial keypoints are labeled for each person.

AI Challenger. Current largest 2D human pose dataset is AI Challenger [11]. Most of the images in the three large-scale datasets, MPII, MSCOCO, and AI Challenger, contain a small number of persons. That is, the human poses in most of the images are not crowd. To help evaluate the performance under crowd scenes, CrowdPose [13] is constructed by extracting the crowd images from the above three large-scale datasets.

PoseTrack. The PoseTrack dataset is first proposed in [12] and then extended in [22]. PoseTrack 2017 [12] follows the same labeling strategy in MPII [9] and the videos are also from MPII. Typically, PoseTrack 2018 [22] is used for evaluating multi-person pose estimation or tracking algorithms.

Table 1. 2D human pose estimation datasets

Dataset	Year	#Train	#Val	#Test	Single/Multi person	#Keypoints
LSP [4]	2010	1,000 images	-	1,000 images	Single-person	14
LSP extended [5]	2011	10,000 images	-	-	Single-person	14
FashionPose [6]	2013	6,530 images	-	1,000 images	Single-person	13
J-HMDB [7]	2013	31,838 frames	-	-	Single-person	13
FLIC [8]	2013	3,987 images	-	1,016 images	Single-person	10
MPII [9]	2014	28,821 images (40,522 people)	-	11,701 images	Single-person	16
MPII (Multi-person) [9]	2014	3,844 images	-	1,758 images	Multi-person	16
MSCOCO Keypoints [10]	2015	64,115 images (262,465 people)	2,693 images (11,004 people)	40,670 images (test-std) 20,288 images (test-dev)	Multi-person	17
AI Challenger [11]	2017	210,000 images	30,000 images	30,000 images (test A) 30,000 images (test B)	Multi-person	14
PoseTrack 2017 [12]	2017	20 videos	20 videos	20 videos	Multi-person	15
PoseTrack 2018 [12]	2018	292 videos	50 videos	208 videos	Multi-person	15
CrowdPose [13]	2019	10,000 images	2,000 images	8,000 images	Multi-person	14

Table 2. 3D human pose estimation datasets

Dataset	Year	Size	Characteristics
HumanEva-I&II [14]	2010	56 videos	about 80,000 frames, 4 subjects
Human3.6M [15]	2014	1,376 videos	about 3.6×10^6 poses
ITOP [16]	2016	100K images	20 subjects
MPI-INF-3DHP [17]	2017	1.3M frames	8 subjects, indoor&outdoor

2.3. 3D Human Pose Estimation Datasets

HumanEva-I&II [14], Human3.6M [15] and MPI-INF-3DHP [17] are three commonly used datasets. Compared with the above 2D human pose datasets, 3D pose datasets have less diversity in terms of the background or environment. For example, all samples from Human3.6M are captured at an indoor environment, thus models trained with the RGB images from it are usually not well generalizable to real-life scenes. Also, ITOP [16] is proposed for evaluating depth-based human pose estimation.

2.4. Datasets for action recognition

Action recognition is to assign a label for a video sequence. In [23], the authors utilize the MPII dataset [9], which is more frequently used for human pose estimation, to evaluate the pose based action recognition method. Also, several datasets are proposed for evaluating action recognition datasets, of which some have skeleton data and some not. In Tab. 3, we sum up the details of some common datasets for action recognition.

There are mainly two kinds of datasets. Some are usually used for evaluating the video-based action recognition algorithms, while the others are usually used for skeleton-based action recognition algorithms. For example, UCF101 [24] and Kinetics [25, 26, 27] are the datasets for evaluating the video-based action recognition methods. Since the skeleton is not provided in these datasets, the datasets can have more diversity since only RGB is needed. For the datasets with skeleton provided [28, 29], the datasets with skeleton information are usually small-scale or under a controlled environment due to the limitation of capturing equipment.

Another big difference is that for RGB action recognition datasets, typically the full body of a person is not captured by the camera. For example, a lot of videos from the "Apply Eye Makeup" class in UCF101 only show the head and the hand of the person. However, in the NTU RGB+D dataset, which is a widely used skeleton-based action dataset, the full body of the persons are mostly well captured in the videos.

3. Human pose estimation

Human pose estimation is a long-standing challenge. In the early days, people treat the pose estimation task as a part based inference task and these models can be roughly classified into two categories. One category of commonly used models is called appearance models. Features of body parts are first extracted by feature descriptors like Histogram of Oriented Gradient [36]. Then distinct body parts are combined together, such as the Poselets used in [37, 38]. Another category of models is deformable models or structural models. For these models, articulated constraints are used to parse body parts. Pictorial Structure Model [39] uses pairwise terms for modeling relative distance between two parts. Yang *et al.*[40] proposes a model containing non-oriented parts with co-occurrence constraints, called mixtures-of-parts model, for articulated pose estimation.

After deep convolutional networks being applied to pose estimation, the performance of estimation models improves with great speed. In the beginning, researchers mostly focus on a well cropped single-person pose estimation task, which is a simplified sub-task. Nowadays, the more general multi-person pose estimation has achieved good results and more challenging settings such as pose estimation in the crowd become hot topics. In Tab. 4, we collect the evaluation results of several representative methods introduced above. These methods include single-person pose estimation and multi-person pose estimation.

3.1. Single-person human pose estimation

An early work of deep convolutional neural network-based human pose estimation model is brought by Jain *et al.*in [45]. The authors make a deep learning-based method to meet or outperform the performance of existing traditional architectures. Multiple networks are trained to perform independent

Table 3. Evaluation datasets for action recognition and their key characteristics.

Dataset	Year	Size	Modality			Characteristics
			RGB	Skeleton	Depth	
HMDB51 [18]	2011	6,766	✓	✗	✗	51 classes
UCF101 [24]	2012	13,320	✓	✗	✗	101 classes
ActivityNet [30]	2015	28,000	✓	✗	✗	203 classes
YouTube-8M [31]	2016	800,0000	✓	✗	✗	4716 classes
Kinetics 400 [25]	2017	306,245	✓	✗	✗	400 classes
Kinetics 600 [26]	2018	495,547	✓	✗	✗	600 classes
Kinetics 700 [27]	2019	650,317	✓	✗	✗	700 classes
MSR-Action3D [32]	2012	567	✗	✓	✓	20 classes
UTKinect-Action3D [33]	2012	200	✓	✓	✓	10 classes
MSRDailyActivity3D [34]	2012	320	✓	✓	✓	16 classes
Northwestern-UCLA [35]	2014	1475	✓	✓	✓	10 classes
NTU RGB+D [28]	2016	56,880	✓	✓	✓	60 classes
NTU RGB+D 120 [29]	2019	114,480	✓	✓	✓	120 classes

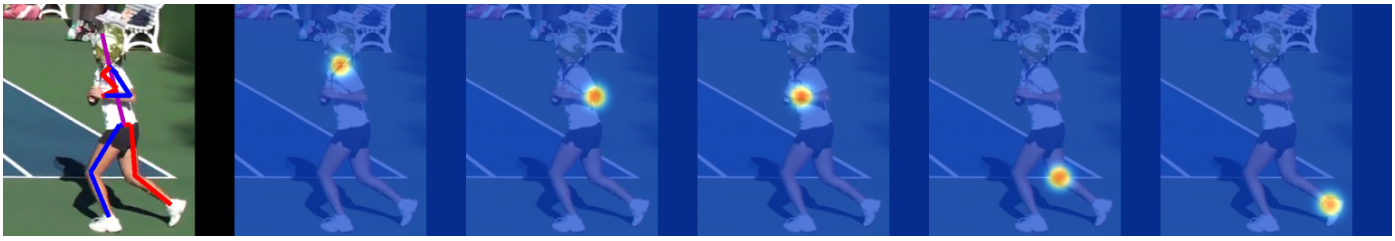

Fig. 1. Heatmap output for human pose estimation. Figure from [41].

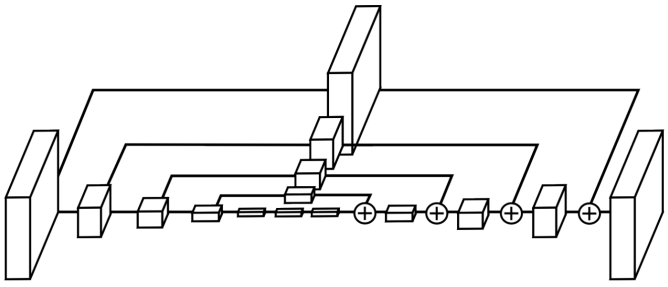

Fig. 2. Network structure of a single hourglass module. Figure from [41].

binary body-part classification. Moreover, in their implementation, the input of these networks is a patch from the original image. In other words, each network is a sliding window on the image and will perform binary classification on these windows. Another deep learning-based method named DeepPose is proposed by Toshev and Szegedy [46]. The deep neural networks work as regressors for the body keypoints. They first generate a rough prediction and then extract related image patches for future refinement on the position of keypoints. Later, some traditional methods are combined with deep networks to achieve better performance. In [47], they combine a convolutional network-based Part-Detector and a Markov Random Field inspired Spatial-Model into a unified learning framework. The output of the network in [47] is a tensor with C channels, where $C = 4$ means the number of joints. Such an output is also known as the heatmap since each joint is encoded in an output channel. An illustration of the heatmap output is shown in Fig. 1. Jain *et al.*[48] incorporates both color and motion features for human pose estimation. The color feature is extracted from RGB images and motion feature is extracted from optical flow maps. A cascaded architecture that combines fine and coarse-scale features is proposed in [49] for single person pose estimation. Later, Wei *et al.*[50] develop a framework named Convolutional Pose Machines, in which a sequence of predictors is trained to make dense predictions at each image location. In the proposed Convolutional Pose Machines, human poses are iteratively refined with different convolutional stages. Another work iteratively refining the pose prediction is [51], in which a network that outputs a correction ϵ for estimation is learned. Following the cascaded structure, Newell *et al.*[41] proposes Stacked Hourglass Networks and outperforms other methods significantly. The Stacked Hourglass Networks include several repeated single Hourglass modules, which are down-sampling and then up-sampling substructures shown in Fig. 2. The biggest difference between Stacked Hourglass Networks and former iterative methods is that the network only requires the image as an input once and no extra image editing such as crop needed, which makes the method can be adapted to other scenarios more easily. The Stacked Hourglass Networks inspired a lot of work. For example, in [52], the authors improve the performance of the Stacked Hourglass Networks with attention modules; In [53], the hourglass structure is improved with the residual connection and a new module named Pyramid

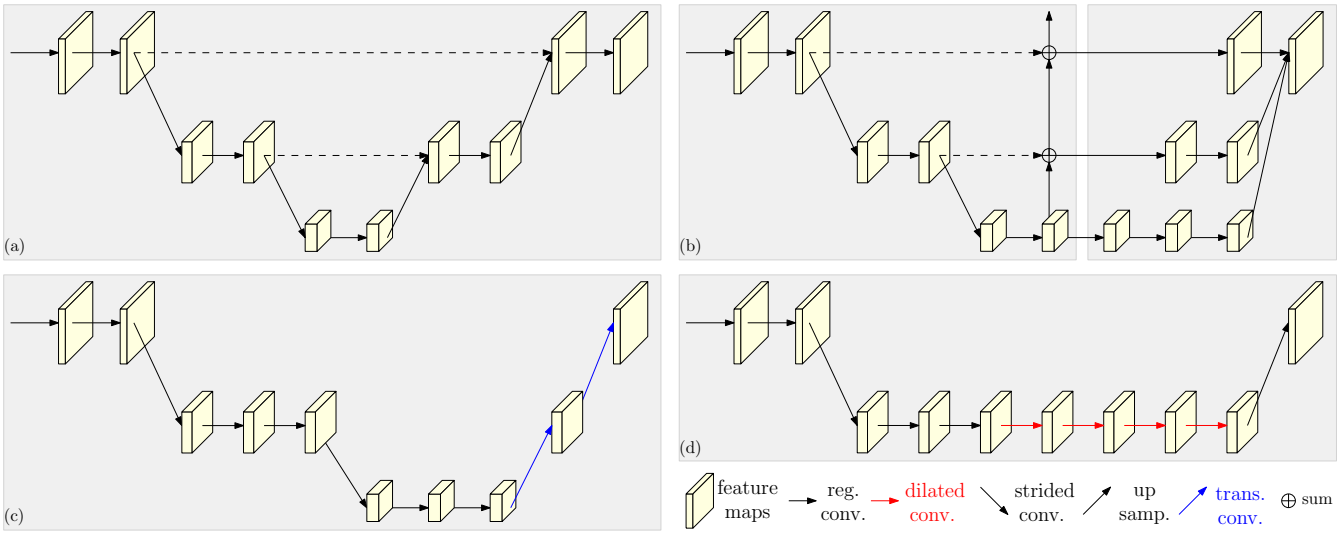

Fig. 3. Several network architectures for pose estimation. (a) Hourglass [41]. (b) Cascaded pyramid networks [42]. (c) SimpleBaseline [43]. Figure from [44].

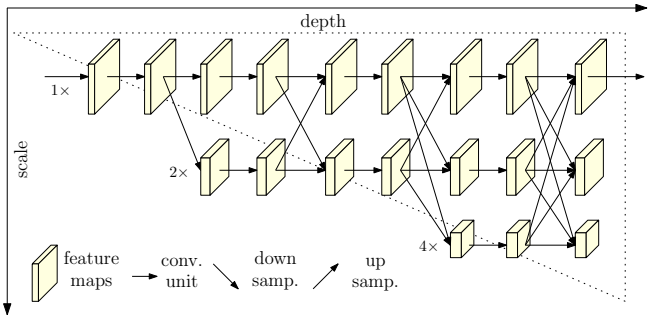

Fig. 4. Network structure of HRNet [44]. Figure from [44].

Residual module is proposed; In [54], several coarse detector branches are combined with a fine detector branch.

Apart from the above methods that design different priors or refine the pose iteratively, some works leverage context information when estimating the pose. Hierarchical Contextual Refinement Network (HCRN) [55] is proposed to process human body joints of different complexities at different layers in a context hierarchy. PoseFix [56] is a single network proposed to refine the pose results. The effectiveness of the PoseFix method is built on the observation that the state-of-the-art human pose estimation methods have similar error distributions [57]. The training data for the PoseFix network are synthesized according to the error statistics found in [57].

Then, the idea of combining iterative refining and context information together is proposed. In [42], the authors observe that the eight-stage stacked hourglass module has the roughly same performance as simply stacking two hourglasses networks. Therefore, they propose a cascaded pyramid network (CPN), which is based on the effective ResNet, for refining the pose keypoints. Though these carefully designed structures indeed bring improvements to the overall performance, the algorithm analysis and comparison may be more difficult. In [43], the authors find that simply adding several transposed convolutional layers after the normal ResNet can achieve comparable performance with those carefully designed structures, such as Stacked Hourglass Networks [41] or CPN [42]. Intuitively, after downsampling, the feature map will contain more context information but lose some local details. In both hourglass structure and CPN, the main concept is to construct a path to make the low-resolution and high-resolution feature maps communicate with each other. Several network structures are compared in Fig. 3. In [44], the authors design a network with more multi-scale fusion paths, achieving significant improvements on the performance. The network structure of HRNet is shown in Fig. 4. Su *et al.*[58] proposes to enhance the channel-wise and spacial information contained in the feature maps by channel shuffle operation and attention mechanism.

Besides the network architectures, data is of vital importance to deep learning systems, thus many methods aim to augment the human pose data or transfer the knowledge from other data domains. Peng *et al.*[59] proposes to add the data augmentation into the optimization loop and aim to automatically acquire a more representative training set. The data augmentation step is then revisited in [60], in which the authors claim that a common flipping strategy is unaligned with the original ones in inference. In [61, 62, 63, 64], the structure priors of the human pose are taken into consideration, resulting in a better overall performance. Later works [65, 66] show that the human parsing information can help improve the performance of pose estimation. Also, the authors in [67] exploit the relation among the body parts and introduce a part-based branching network to learn representations specific to each part group.

In real applications, the computation cost is of great importance and some works target at building light-weight pose estimation models. In [68], parameter binarised CNN models for pose estimation is built to accommodate resource-limited platforms. Rafi *et al.*[69] discusses the best practice for de-

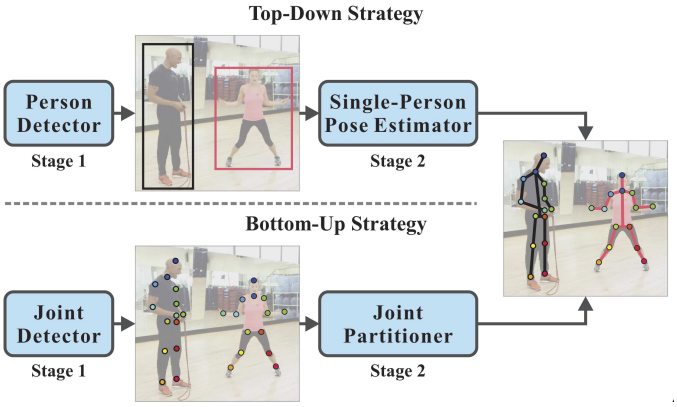

Fig. 5. Top-down approaches and bottom-up approaches for multi-person human pose estimation. Figure from [71].

signing an efficient human pose estimation network. Distilling the pose structure knowledge in a strong teacher network into a lightweight model is shown to be effective in [70].

3.2. Multi-person human pose estimation

Unlike the single-person human pose estimation, we have to build the connection between estimated the joints and the persons in the image for the multi-person setting. Generally speaking, recent methods can be roughly grouped into two categories: top-down approaches and bottom-up approaches. For top-down approaches, each person is first detected and then the poses are calculated separately. The other category, i.e., bottom-up approaches, will first estimate all the human parts in the images and then associate the body parts to different persons. The difference between the two categories is shown in Fig. 5.

The two categories do not cover all recent methods and new frameworks are proposed for different scenarios. For embedded computing, one-stage pose estimation methods [71, 73] are proposed to remove the costly association step in bottom-up approaches. Also, estimating poses in crowded scenes is still challenging and new frameworks are needed. In [13], the authors propose a specific dataset named CrowdPose for better evaluating the performance of human pose estimation algorithms. On the more challenging dataset, some algorithms such as [74, 72] that work well on public benchmarks [9, 10] cannot well handle the crowded scenes.

3.2.1. Top-down approaches

Top-down approaches are the straightforward extension of the single-person pose estimation methods, since the first step is to detect and crop each person out and then apply single-person pose estimation algorithms [75, 76, 77, 78]. Therefore, some papers such as [64, 42, 43, 44, 58] that focus on single-person pose estimation are also evaluated on the multi-person setting by running their method on the ground-truth human bounding boxes.

The two steps in two-down approaches are not irrelative. Some methods build an end-to-end workflow to optimize the two steps simultaneously. Mask R-CNN [74], proposed on ICCV'17, add an extra mask branch on two-stage detectors and serves as a strong baseline for top-down multi-person human pose estimation. The mask branch in Mask R-CNN predicts the heatmaps for each keypoints of the pose, making the whole prediction process much faster than before. Another work [79], which is published on the same conference as Mask R-CNN, uses Faster R-CNN [80] to detect the person and then crops the person area out from the original image. One thing worth noting in [79] is that they propose to predict a 2D offset field for each keypoint. The idea of predicting heatmaps and a vector field is later widely used in related pose estimation works including the bottom-up approaches. Also, on ICCV'17, a top-down approach named AlphaPose [81] aims to improve the quality of human bounding boxes generated by Faster R-CNN [80].

3.2.2. Bottom-up approaches

Bottom-up approaches predict all the body parts first and then assemble the parts to infer full body poses. DeepCut [20] first builds a pairwise graph between all the detections and next to the part candidates are clustered belonging to one person. In DeepCut, the part labeling task is turned into an Integer Linear Program formulation. Finally, all the person clusters are combined with labeled body parts and therefore the final multi-person human poses are generated. Later, DeeperCut [83] proposes image-conditioned pairwise terms and incremental optimization strategy to constrain the search space and speed up the assembling step. Also, they use a more powerful backbone ResNet [84] in DeeperCut, resulting in further improvement in the overall performance.

However, solving the Integer Linear Program is time-consuming and thus weaken the practical value of the above DeepCut framework. An important work named OpenPose [21] is then proposed and greatly advance the performance of the bottom-up approaches. In [21], apart from the heatmap for locating the body parts, a vector field named Part Affinity Fields (PAF) is learned for associating body parts with distinct persons. In the post-processing step, a bipartite graph is constructed by connecting predicted adjoint parts and next bipartite matching can be efficiently solved with the PAF output. The pipeline of the whole algorithm is illustrated in Fig. 6. Authors from [21] make their implementation of OpenPose public available and can handle whole-body pose estimation, i.e., including foot, hand, and facial keypoints. In [85], OpenPose is further optimized for dealing with handle scale differences between body/foot and face/hand keypoints.

The idea of predicting a vector field is later developed in PersonLab [86] and PIFPAF [82]. In PersonLab [79], they propose a recurrent scheme to improve the accuracy of long-range predictions. The recurrent scheme is based on three offsets predictions: short-range offsets, middle-range offsets, and long-range offsets. Middle-range offsets are combined with short-range offsets for getting the keypoints, while long-range offsets are further added to enable separating the mask of all persons to distinct persons. PIFPAF [82] predicts two vector fields: Part Intensity Fields (PIF) and Part Association Fields (PAF). PIF is used for localizing the body joints, which has the same purpose as the heatmaps. Formally, for every output location (i, j), the

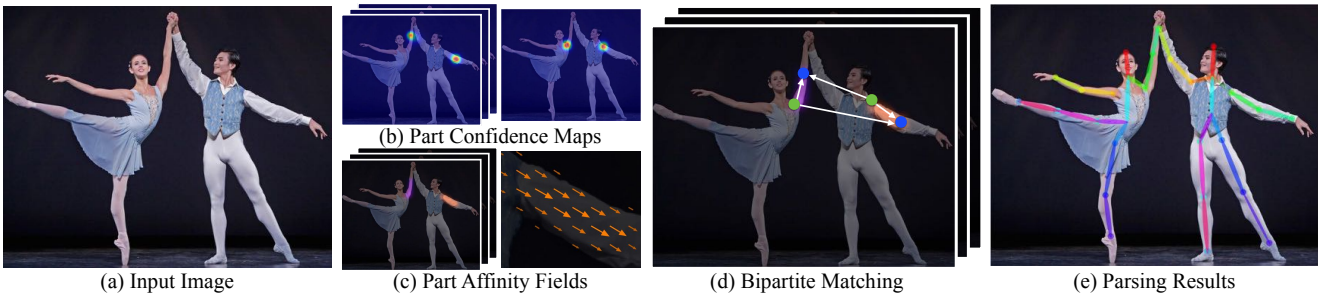

Fig. 6. Pipeline of the bottom-up approach OpenPose [72]. The (b) Part Confidence Maps is the same output as demonstrated in Fig. 1. After predicting (c) Part Affinity Fields, (d) Bipartite Matching is performed to associate body part candidates and finally get the (e) Parsing Results. Figure from [72].

Fig. 7. Part Intensity Fields (PIF) proposed in [82]. (a) shows the confidence map, which is the p_c. (b) shows the vector fields, i.e., (p_x, p_y). (c) shows the final heatmap of the left shoulder. Figure from [82].

PIF consists of

$$\{p_c^{ij}, p_x^{ij}, p_y^{ij}, p_b^{ij}, p_\delta^{ij}\}, \tag{1}$$

where p_c^{ij} is the confidence, p_x^{ij}, p_y^{ij} are the vector pointing to the joint, p_b^{ij} is called spread and used for adjusting the Laplace loss, p_δ^{ij} is the scale. An illustration of PIF is shown in Fig. 7. Moreover, PAF in [82] is different from the PAF in [21, 86]. PAF in [82] predicts two vectors and the end of the vectors are the joints, leading to two regression tasks. In Associate Embedding [87], along with a heatmap for each joint, an associative embedding tag is learned by the network. Therefore, the channel of the output in [87] is $2C$, where C is the number of keypoints. Affinity fields can also be used for pose tracking. Spatio-Temporal Affinity Fields (STAF) is proposed in [88] for consistently handle body motions of a wide range of magnitudes.

The above methods all assign the identity within a post-processing step and some works use a different different post-processing algorithm. The work in [89] design a multi-task framework and predict the bounding boxes for persons in an intermediate step. Its structure is based on the RoI pooling operation [80] and thus is closely related to Mask R-CNN [74]. The work in [90] uses dense regressions from keypoint candidates as centroids of persons to partition all keypoint detections.

3.3. Human pose estimation in videos.

Estimating the pose in videos requires the algorithm to take temporal information into consideration. The work in [91] is the first to study how to adapt deep networks for estimating human pose in videos. They use optical flow to align heatmap predictions from neighboring frames and propose a final parametric pooling layer to convert the heatmaps into a final confidence map. Song *et al.*[92] propose Thin-Slicing Network, in which the spatio-temporal relational is leveraged via a flow warping layer. The method in [93] demonstrates that fine-tuning the network on those frames with high confidence keypoints is good for overall performance. In [94], they improve the Convolutional Pose Machines [50] with an LSTM module to capture the temporal correlation of human poses among video frames. The work in [95] treats the task of learning from multiple pose dataset as a multi-domain learning task. The multi-domain based method wins the first place of PoseTrack ECCV 2018 Challenge with addition datasets COCO and MPII.

If we not only need to get the pose of each person in all frames, but also want to identify which poses are from the same person, then the task is also known as pose tracking. Prior pose tracking works [12, 96, 43] can be viewed as a two-stage scheme: first, generate the joint candidates in each frame and next connecting those joints. In [96], pose on each frame is generated by a 3D Mask R-CNN and then predictions are linked by solving a graph partition problem. They experimentally show that the Hungarian algorithm and box overlap based metric is the best practice for connecting the candidates. The matching strategy is again adopted in [43] and the detection method in [43] is a simpler version consisting of detecting each person and then predicting the pose.

Table 4. Performance of 2D human pose estimation algorithms. Note that the results listed here may not serve as a fair comparison between methods, due to different settings such as multi-scale test.

Method	Year	Characteristics	MPII-Single (PCKh@0.5)	MPII-Multi (mAP@0.5)	COCO test-dev (AP)	PoseTrack 2017 (mAP)
Tompson *et al.*[47]	NIPS'14	single-person	79.6	-	-	-
Tompson *et al.*[49]	CVPR'15	single-person	82.0	-	-	-
IEF [51]	CVPR'16	single-person	81.3	-	-	-
Newell *et al.*[41]	ECCV'16	single-person	90.9	-	-	-
DeeperCut [83]	ECCV'16	bottom-up	88.5	59.5	-	-
OpenPose [21, 72]	CVPR'17	bottom-up	-	75.6	64.2	-
AlphaPose [81]	ICCV'17	top-down	-	76.7	71.0	-
Mask RCNN [74]	ICCV'17	top-down	-	-	69.2	-
Associative Emb. [87]	NIPS'17	bottom-up	-	77.5	65.5	-
CPN [42]	CVPR'18	top-down	-	-	72.1	-
PersonLab [86]	ECCV'18	bottom-up	-	-	68.7	-
MultiPoseNet [89]	ECCV'18	bottom-up	-	-	70.5	-
SimpleBaseline [43]	ECCV'18	top-down	91.5	-	73.7	74.6
HRNet [44]	CVPR'19	top-down	92.3	-	75.5	74.9
PifPaf [82]	CVPR'19	bottom-up	-	-	66.7	-

3.4. 3D human pose estimation

The 3D human pose estimation task requires the algorithm to predict the depth of the joints in addition. There are basically two ways to achieve the goal: One way to predict 3D human pose is directly regressing the 3D locations; Another way is the first 2D pose and then 3D pose, which is a two-stage framework.

Estimating 3D pose directly.. The work in [97] directly regresses the 3D coordinates of each body joint via a multitask framework. Pavlakos *et al.*[98] uses the volumetric representation for 3D human pose and the network is based on Hourglass [41]. Though 3D human pose is computed from direct regression, 2D pose information can also be used for improving the 3D pose regressors [99, 17].

From 2D pose to 3D pose.. After having the 2D pose, the 3D pose can be estimated by finding the best match in a library [100, 101]. Also, we can generate 3D pose hypothesis and compare with the 2D pose [102, 103]. The work in [104] uses a fully connected residual network to predict 3D pose from a 2D pose. Here the input is 2D keypoints, making the method domain-invariant. Graph Convolutional Networks (GCN) are employed in [105] for estimating 3D pose from a sequence of 2D poses. They first construct a spatial-temporal graph on consecutive 2D poses and then apply a GCN based local-to-global network to generate 3D poses.

3.5. Human pose estimation from depth images

Predicting the 3D pose with depth images can greatly boost the performance. As reported by V2V [106], the current best result on ITOP is 88.74% (mAP), which means about 90 percent of joints can be estimated within 10cm error. As a comparison, the current best result [107] on Human3.6M is 20.8, which means that the average error of joints is 20.8cm. Though the datasets are different, the performance gap indicates that video based 3D pose estimation is not as good as depth-based.

Human pose estimation from depth images can be also formulated as a regression task. Shotton *et al.*[108] first classify each pixel into body parts with Random Forests and then use a Mean-Shift based approach to estimate locations of body joints. Recently Microsoft released a new depth sensor, named Azure Kinect, together with a CNN based skeletal tracking SDK [1] where a bottom-up approach is used for human pose estimation on the IR image. In the work of [109], the authors use Random Forests and Graph-cuts to segment human limbs in depth maps. With the rise of deep networks and end-to-end training, the work in [110] first uses CNNs to obtain a segmentation map from depth maps and then derives 3D joint locations from reconstructing segmented point-clouds. V2V-Posenet [106] uses voxels as the input and joint locations are regressed with 3D-CNN autoencoders.

Instead of treating the estimation task as a regression task, template-based estimation algorithms [111, 112, 113] utilize a library of 3D poses for fitting into the segmented depth map. The template-based algorithms have also been widely used in hand pose estimation [114, 115, 116], which is closely related to human pose estimation. Some recent depth-based works are evaluated on both hand pose estimation and human pose estimation, such as [117, 118]. In [117], the authors propose a model named Deep Depth Pose (DDP) to compute the weights for linearly combining the prototypes. An anchor-based approach named Anchor-to-Joint regression network (A2J) is proposed in [118]. In A2J, anchor points are densely set on depth image as local regressors for the joints.

[1] https://azure.microsoft.com/en-us/services/kinect-dk/

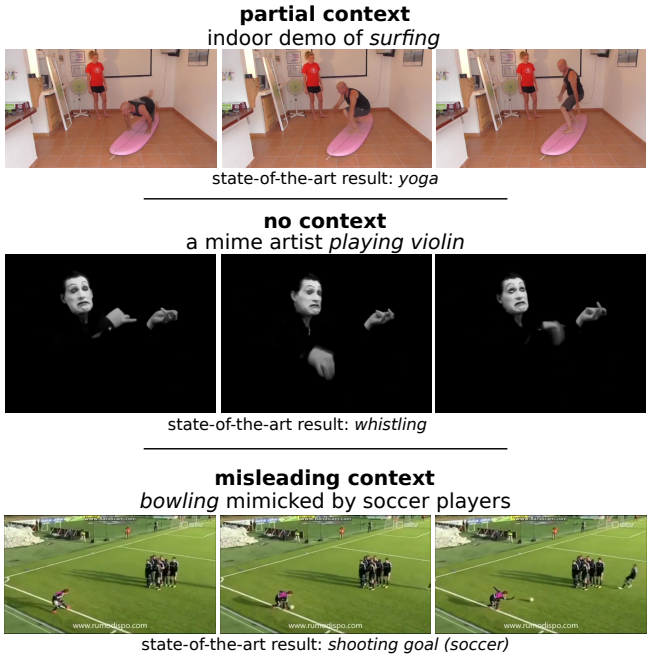

Fig. 8. Examples of how context affect recognizing an action. The first row shows the partial context in a video, in which the surf board is common but indoor scene is rare. The second row shows the absent context. The third row shows the misleading context. Figure from [126].

4. Human pose estimation for action recognition

Human pose estimation and action recognition are two closely related tasks, as both of them are concerned with understanding the human motion. Before deep learning models prevailing in vision tasks, several early works directly use the estimated pose for action recognition [119, 120], or pose based features [7, 121]. Also, some works try to tackle both of the tasks at the same time [122, 123] with a multi-task framework. Later, deep learning-based methods greatly improve the performance of action recognition algorithms and the philosophy of end-to-end training is widely used in the deep network-based models. Generally, for recognizing the action from a video, the gap between pose based methods and those end-to-end trained is obvious. Despite better results, the interpretability of end-to-end trained networks is relatively poor. That is, learning with the background scenes information may lead to overfitting and thus affect the real generalization [124, 125]. As illustrated in Fig. 8, the context may not always be helpful or sometimes be misleading. In [126], a new dataset is constructed to help evaluate methods on the out-of-context actions.

4.1. Overview of action recognition methods

Recognizing action from videos has been widely studied and can be roughly broken down into two directions: video-based and skeleton-based. Skeleton-based action recognition is more concerned with human motion, without taking the environment or object context into consideration. Inputs with more environmental context generally contain more information, therefore video-based methods may achieve better performance and handle more complex actions.

Representative methods for video based methods can be group into three categories: 3D convolutional networks based methods [127], long short-term memory (LSTM) based methods [128] and two-stream convolutional networks based methods [129]. A 3D convolutional network concatenates the frames in a video and then use the 3D convolution method to learn from the 3D space-time structure. 3D convolutions are computationally expensive, so some works study how to extend 2D convolutions [130, 131] to deal with 3D structure inputs. In LSTM based methods, the frames from a video are taken as a sequence of inputs [132, 133]. In contrast, a two-stream convolutional network [134] usually leverages the optical flow information, which is computed from a sequence of frames. Also, there are two-stream based methods that do not require optical flow and use other alternatives instead, such as SlowFast Network [135].

For skeleton-based action recognition, the input becomes a sequence of the body joint locations. Since the inputs are sequential data, recurrent neural networks are widely used, such as bi-RNNs [136], AGC-LSTM [137], Bayesian GC-LSTM [138]. Moreover, because the locations are not as structural as images, graph models are also used, such as the graph convolutional network used in [139].

4.2. Pose estimated from images for action recognition

Though methods that rely on pose estimation to perform action recognition are not common in the literature, using the pose directly estimated from images has been tested in previous works. Kinetics-400 [25] is a large scale dataset widely adopted for evaluating video-based action recognition algorithms. In [139], the authors use OpenPose [21] to extract the pose on each frame and then test their skeleton-based action recognition. Such a setting is quickly followed by later skeleton-based action recognition methods. This evaluation protocol is later followed by many skeleton-based action recognition methods.

In Tab. 5, we demonstrate the performance comparison between video-based methods and skeleton-based methods. Obviously, on the Kinectics-400 dataset, skeleton-based methods are not comparable to the video-based methods. For the methods shown in Tab. 5, Yan *et al.*[139] propose to model temporal dynamic with a spatial-temporal graph convolutional network (ST-GCN). The workflow of ST-GCN is shown in Fig. 9, from which we can see that the most important part is how to learn with the unstructured pose data. Following the GCN framework, the work in [140] proposes a part-based graph convolutional network (PB-GCN) to learn the correlations between parts. Also, the work in [141] proposes a two-stream adaptive graph convolutional network (2s-GCN) to leverage the second-order information (the lengths and directions of bones) of the skeleton data.

Generally speaking, the skeleton-based methods shown in Tab. 5 requires a fairly good pose results. Although the videos with more environmental context generally contain more information, lacking the surrounding information is not enough for explaining the big gap. We are of the opinion that there are two more reasons behind the big gap: The first reason is that at present pose based action recognition algorithms are less powerful and not robust enough to the noise contained in the

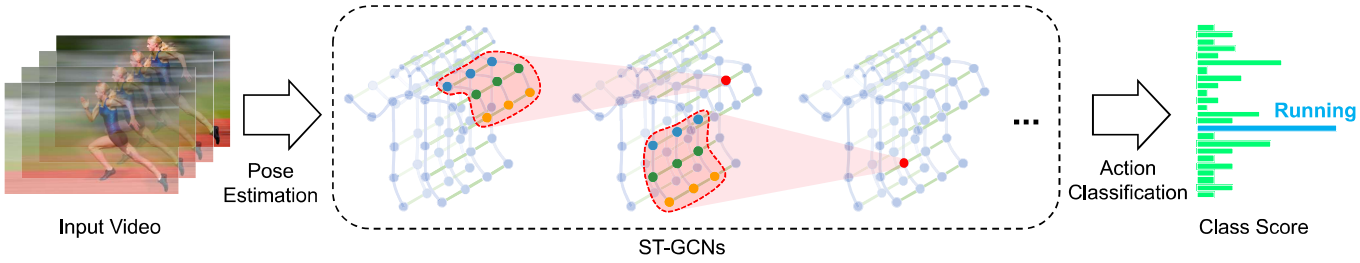

Fig. 9. Using graph convolutional networks to process the skeleton data extracted from videos. Figure from [139].

skeleton annotations; The second reason is that the dataset contains action categories that are not concerned with human body motion. Yan *et al.*[139] selects a subset of 30 classes from Kinectics. These classes are strongly related to human motion and the subset is named as Kinectics-Motion. When trained with the full dataset, STGCN achieves 72.4% on Kinectics-Motion, while video-based method Flow CNN [25] achieves 72.8%. These results suggest that the human motion related subset Kinectics-Motion is a good dataset for evaluating the pose based action recognition methods. Most recently, a large scale dataset named HIE [142] with well-labeled poses in outdoor scenes for each action. Unlike previous datasets, HIE is a large scale and also contains both poses and actions, therefore it can be a good dataset for evaluating human-centric action recognition in outdoor scenes.

Apart from using the pose keypoints explicitly extracted from images, other types of encoding the pose are also verified by recent work. Liu *et al.*[143] proposes to use the pose heatmap estimated from RGB images inputs to enhance the skeleton-based action recognition. As the workflow shown in Fig. 10, the final prediction of the action label is based on the estimated pose keypoints and heatmaps. In their experiments, for the Cross-Subject metric on NTU RGB+D, results with pose generated from depth can reach 80.5% for 2D inputs and 82.3% for 3D inputs, while the best result generated from RGB images is 78.8%. The gap between the performance demonstrates that the quality of poses estimated from the depth and RGB images are different. Also on the same dataset (NTU RGB+D), the experiments in [126] show that STGCN with the pose estimated by OpenPose [21] is 79.8%, while with the pose given by the Kinect sensor the result is 81.5%. The work in [144] represents the motions of joints with spatio-temporal activations, which are extracted by a stack of pose estimation layers. Next, the activations are reprojected in space and time using a stack of 3D convolutions for action recognition.

4.3. Multi-task training of pose estimation and action recognition

Estimating the body pose and recognizing the human action form videos can help infer more semantic information about the human in a video. The work in [123] uses a spatial-temporal And-Or graph model to decompose the action into poses. Liu *et al.*[150] present a pose feature describing human action as the aggregation of a sequence of joint estimation maps. Iqbal *et al.*[151] propose a pictorial structure model and empirically

Table 5. Performance of video based and skeleton based methods on Kinetics-400. For skeleton-based methods, the pose is generated by OpenPose [21].

Method	Year	Top-1 Acc	Top-5 Acc
Video based			
I3D [130]	CVPR'17	71.6	90.0
R(2+1)D [145]	CVPR'18	73.9	90.9
SlowFast [135]	ICCV'19	79.8	93.9
Skeleton based			
Feature Enc [146]	CVPR'15	14.9	25.8
Deep LSTM [28]	CVPR'16	16.4	35.3
Temporal Conv [147]	CVPRW'17	20.3	40.0
ST-GCN [139]	AAAI'18	30.7	52.8
BPLHM [148]	TNNLS'19	33.4	56.2
AS-GCN [149]	CVPR'19	34.8	56.5
2s-AGCN [141]	CVPR'19	36.1	58.7

demonstrate that information about human actions can improve pose estimation. Luvizon *et al.*[152] design a network that estimates 2D and 3D pose from still images and recognizes action from videos. Their method is evaluated on both pose estimation datasets and action recognition datasets. The work in [153] explicitly learns the pose motions via a temporal pose convolution and therefore learns spatio-temporal pose representations for action recognition. Action Machine proposed in [154] uses person bounding boxes for instance-level action analysis. To achieve this, based on I3D [130], a branch for human pose estimation and a 2D CNN for pose-based action recognition are constructed. Part-aligned Pose-guided Recurrent Network (P2RN) [155] is proposed to use the information on body-parts to enhance the performance of an action recognition system. For pose and action related to hands, Tekin *et al.*[156] and Yang *et al.*[157] recently propose unified frameworks for jointly estimating poses and predicting actions. In [156], the information in the temporal domain is merged and propagated to infer interactions between hand and object trajectories and recognize actions. In [157], the authors exploit joint-aware features for gesture recognition and 3D hand pose estimation.

5. Discussion

5.1. Why do we need human pose based action recognition?

The most improtant reason is that context in videos may be misleading sometimes for recognizing action. There are a

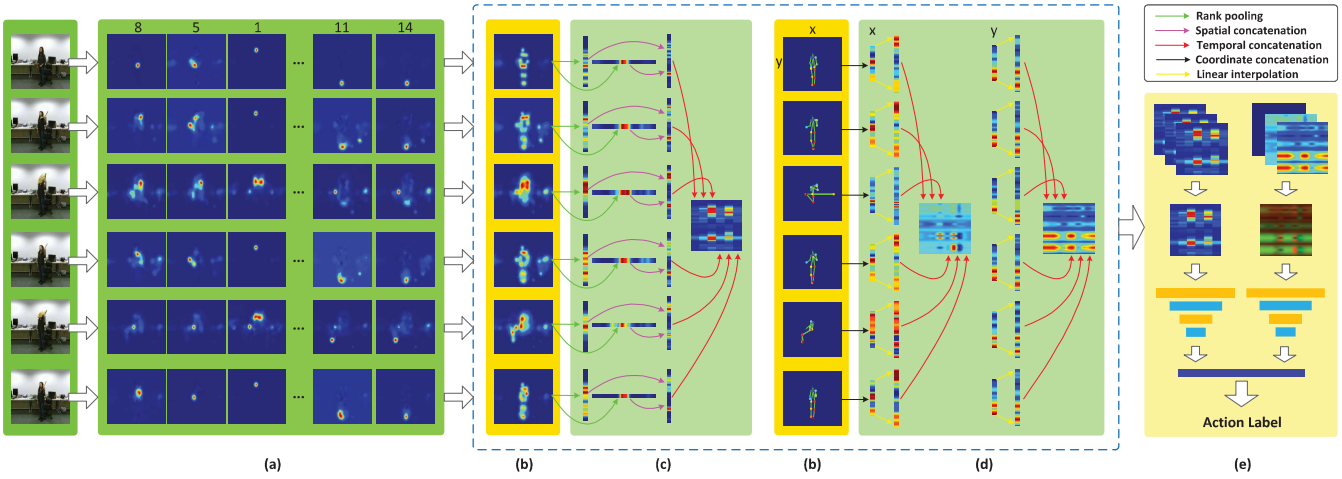

Fig. 10. Workflow of using the pose headmap for action recognition. Six steps of the method: (a) Predicting headmap for body parts; (b) Aggregation of the body parts to form the pose heatmap and the pose keypoints; (c) Describing the evolution of heatmaps with spatial rank pooling; (d) Describing the evolution of poses with body guided sampling; (e)Predicting the action label. Figure from [143].

bunch of related works [124, 125, 126] presenting examples of helpless context in videos. Removing context by extracting poses can fundamentally avoid the missleading context. In addition, although videos in general contains more information, the network may not learn the right feature for classification (e.g. learns more from background instead of person movement), thus would fail in testing case. Skeleton based approach focuses on person motion, thus can better generalize to new action video. Moreover, recognizing actions by human pose provides potential future understanding of human behaviors from the perspective of body motions.

To summarize, human pose estimation plays a critical role for reliable and explainable human action recognition, with broad applications in industrial applications like human-computer interaction and video surveillance.

5.2. How can we improve human pose based action recognition?

Despite promising progress that has been made in the literature, there still exist several challenges that should be well addressed in future work.

First of all, human pose algorithms may fail to differentiate each person in a crowd. Even the bottom-up approaches which are supposed to well locate each human part, the assembling process is also challenging in the crowd cases. Second, the human pose estimation algorithms usually require large backbone models as well as high feature resolution to improve the localization performance, which leads to intensive computational speed. This prohibits the wide applications of these algorithms running on embedded devices and mobile phones. Thus, inference speed is a critical but challenging problem to optimize. Third, although there are a few great works devoted to the 3D human pose estimation problem, most of the existing benchmarks are captured in the constrained environment with limited subjects and poses. Therefore, an unconstrained 3D human pose benchmark with diverse subjects and action poses is inevitable to further boost the research on the 3D human pose estimation problem. Last but not the least, the skeleton provides an important clue for the human pose estimation problem but other information like optical flow and RGB appearance are usually complementary. Thus, the effective fusion of these cues can serve as a good foundation to understand human actions.

Besides the four challenges discussed above, there also exist a lot of interesting problems that should be well explored for the topic of human pose estimation and its application to human action recognition. The development of human pose estimation and action recognition should contribute a lot to the computer vision industry in the next decade.

Acknowledgements

This work is supported in part by the start-up funds from University at Buffalo.

References

[1] B. Ren, M. Liu, R. Ding, H. Liu, A survey on 3d skeleton-based action recognition using learning method, arXiv preprint arXiv:2002.05907 (2020).

[2] Z. Liu, J. Zhu, J. Bu, C. Chen, A survey of human pose estimation: the body parts parsing based methods, Journal of Visual Communication and Image Representation 32 (2015) 10–19.

[3] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, E.-h. Zahzah, Human pose estimation from monocular images: A comprehensive survey, Sensors 16 (2016) 1966.

[4] S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, in: British Machine Vision Conference, 2010.

[5] S. Johnson, M. Everingham, Learning effective human pose estimation from inaccurate annotation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1465–1472.

[6] M. Dantone, J. Gall, C. Leistner, L. Van Gool, Human pose estimation using body parts dependent joint regressors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3041–3048.

[7] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M. J. Black, Towards understanding action recognition, in: IEEE International Conference on Computer Vision, 2013, pp. 3192–3199.

[8] B. Sapp, B. Taskar, MODEC: multimodal decomposable models for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3674–3681.

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

[10] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.

[11] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, Y. Wang, AI challenger : A large-scale dataset for going deeper in image understanding, CoRR abs/1711.06475 (2017).

[12] U. Iqbal, A. Milan, J. Gall, Posetrack: Joint multi-person pose estimation and tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4654–4663.

[13] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, C. Lu, Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10863–10872.

[14] L. Sigal, A. O. Balan, M. J. Black, Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, International Journal of Computer Vision 87 (2010) 4.

[15] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2013) 1325–1339.

[16] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, F. Li, Towards viewpoint invariant 3d human pose estimation, in: European Conference on Computer Vision, 2016, pp. 160–177.

[17] 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: International Conference on 3D Vision, 2017, pp. 506–516.

[18] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: IEEE International Conference on Computer Vision, 2011.

[19] L. D. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: IEEE International Conference on Computer Vision, 2009, pp. 1365–1372.

[20] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, B. Schiele, Deepcut: Joint subset partition and labeling for multi person pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4929–4937.

[21] Z. Cao, T. Simon, S. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1302–1310.

[22] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, B. Schiele, Posetrack: A benchmark for human pose estimation and tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5167–5176.

[23] L. Pishchulin, M. Andriluka, B. Schiele, Fine-grained activity recognition with holistic and pose based features, in: German Conference on Pattern Recognition, Springer, 2014, pp. 678–689.

[24] K. Soomro, A. R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv:1212.0402 (2012).

[25] 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, arXiv preprint arXiv:1705.06950 (2017).

[26] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600, arXiv preprint arXiv:1808.01340 (2018).

[27] J. Carreira, E. Noland, C. Hillier, A. Zisserman, A short note on the kinetics-700 human action dataset, arXiv preprint arXiv:1907.06987 (2019).

[28] A. Shahroudy, J. Liu, T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3d human activity analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.

[29] 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 Transactions on Pattern Analysis and Machine Intelligence (2019).

[30] B. G. Fabian Caba Heilbron, Victor Escorcia, J. C. Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.

[31] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, Youtube-8m: A large-scale video classification benchmark, CoRR abs/1609.08675 (2016).

[32] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2010, pp. 9–14.

[33] L. Xia, C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2012, pp. 20–27.

[34] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2012, pp. 1290–1297.

[35] J. Wang, X. Nie, Y. Xia, Y. Wu, S. Zhu, Cross-view action modeling, learning, and recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2649–2656.

[36] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, volume 1, IEEE, 2005, pp. 886–893.

[37] L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: IEEE International Conference on Computer Vision, IEEE, 2009, pp. 1365–1372.

[38] L. Bourdev, S. Maji, T. Brox, J. Malik, Detecting people using mutually consistent poselet activations, in: European Conference on Computer Vision, Springer, 2010, pp. 168–181.

[39] P. F. Felzenszwalb, D. P. Huttenlocher, Pictorial structures for object recognition, International journal of computer vision 61 (2005) 55–79.

[40] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1385–1392.

[41] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision, 2016, pp. 483–499.

[42] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103–7112.

[43] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: European Conference on Computer Vision, 2018, pp. 472–487.

[44] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.

[45] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, C. Bregler, Learning human pose estimation features with convolutional networks, in: International Conference on Learning Representations, 2014.

[46] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.

[47] J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: Advances in Neural Information Processing Systems, 2014, pp. 1799–1807.

[48] A. Jain, J. Tompson, Y. LeCun, C. Bregler, Modeep: A deep learning framework using motion features for human pose estimation, in: Asian Conference on Computer Vision, 2014, pp. 302–315.

[49] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 648–656.

[50] S. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

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

[52] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, X. Wang, Multi-context attention for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5669–5678.

[53] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, in: IEEE International Conference on Computer Vision, 2017, pp. 1290–1299.

[54] S. Huang, M. Gong, D. Tao, A coarse-fine network for keypoint localization, in: IEEE International Conference on Computer Vision, 2017, pp. 3047–3056.

[55] X. Nie, J. Feng, J. Xing, S. Xiao, S. Yan, Hierarchical contextual refinement networks for human pose estimation, IEEE Transactions on Image Processing 28 (2018) 924–936.

[56] G. Moon, J. Y. Chang, K. M. Lee, Posefix: Model-agnostic general human pose refinement network, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7773–7781.

[57] M. R. Ronchi, P. Perona, Benchmarking and error diagnosis in multi-instance pose estimation, in: IEEE International Conference on Computer Vision, 2017, pp. 369–378.

[58] K. Su, D. Yu, Z. Xu, X. Geng, C. Wang, Multi-person pose estimation with enhanced channel-wise and spatial information, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5674–5682.

[59] X. Peng, Z. Tang, F. Yang, R. S. Feris, D. N. Metaxas, Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2226–2234.

[60] J. Huang, Z. Zhu, F. Guo, G. Huang, The devil is in the details: Delving into unbiased data processing for human pose estimation, in: IEEE International Conference on Computer Vision, 2019.

[61] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, J. Wang, Human pose estimation using global and local normalization, in: IEEE International Conference on Computer Vision, 2017, pp. 5600–5608.

[62] Y. Chen, C. Shen, X. Wei, L. Liu, J. Yang, Adversarial posenet: A structure-aware convolutional network for human pose estimation, in: IEEE International Conference on Computer Vision, 2017, pp. 1221–1230.

[63] W. Tang, P. Yu, Y. Wu, Deeply learned compositional models for human pose estimation, in: European Conference on Computer Vision, 2018, pp. 197–214.

[64] L. Ke, M. Chang, H. Qi, S. Lyu, Multi-scale structure-aware network for human pose estimation, in: European Conference on Computer Vision, 2018, pp. 731–746.

[65] X. Nie, J. Feng, S. Yan, Mutual learning to adapt for joint human parsing and pose estimation, in: European Conference on Computer Vision, 2018, pp. 519–534.

[66] X. Nie, J. Feng, Y. Zuo, S. Yan, Human pose estimation with parsing induced learner, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2100–2108.

[67] W. Tang, Y. Wu, Does learning specific features for related parts help human pose estimation?, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1107–1116.

[68] A. Bulat, G. Tzimiropoulos, Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources, in: IEEE International Conference on Computer Vision, 2017, pp. 3726–3734.

[69] U. Rafi, B. Leibe, J. Gall, I. Kostrikov, An efficient convolutional network for human pose estimation, in: British Machine Vision Conference, 2016.

[70] F. Zhang, X. Zhu, M. Ye, Fast human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3517–3526.

[71] X. Nie, J. Zhang, S. Yan, J. Feng, Single-stage multi-person pose machines, in: IEEE International Conference on Computer Vision, 2019.

[72] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, Y. A. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 1–1.

[73] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, CoRR abs/1904.07850 (2019).

[74] K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[75] M. Sun, S. Savarese, Articulated part-based model for joint object detection and pose estimation, in: IEEE International Conference on Computer Vision, 2011, pp. 723–730.

[76] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, B. Schiele, Articulated people detection and pose estimation: Reshaping the future, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3178–3185.

[77] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, IEEE Tansactions on Pattern Analysis and Machine Intelligence 35 (2012) 2878–2890.

[78] G. Gkioxari, B. Hariharan, R. B. Girshick, J. Malik, Using k-poselets for detecting people and localizing their keypoints, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3582–3589.

[79] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, Towards accurate multi-person pose estimation in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3711–3719.

[80] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1137–1149.

[81] H. Fang, S. Xie, Y. Tai, C. Lu, RMPE: regional multi-person pose estimation, in: IEEE International Conference on Computer Vision, 2017, pp. 2353–2362.

[82] S. Kreiss, L. Bertoni, A. Alahi, Pifpaf: Composite fields for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11977–11986.

[83] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, Deepercut: A deeper, stronger, and faster multi-person pose estimation model, in: European Conference on Computer Vision, 2016, pp. 34–50.

[84] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[85] G. Hidalgo, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, Y. Sheikh, Single-network whole-body pose estimation, in: IEEE International Conference on Computer Vision, 2019.

[86] G. Papandreou, T. Zhu, L. Chen, S. Gidaris, J. Tompson, K. Murphy, Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model, in: European Conference on Computer Vision, 2018, pp. 282–299.

[87] A. Newell, Z. Huang, J. Deng, Associative embedding: End-to-end learning for joint detection and grouping, in: Advances in Neural Information Processing Systems, 2017, pp. 2277–2287.

[88] Y. Raaj, H. Idrees, G. Hidalgo, Y. Sheikh, Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4620–4628.

[89] M. Kocabas, S. Karagoz, E. Akbas, Multiposenet: Fast multi-person pose estimation using pose residual network, in: European Conference on Computer Vision, 2018, pp. 437–453.

[90] X. Nie, J. Feng, J. Xing, S. Yan, Pose partition networks for multi-person pose estimation, in: European Conference on Computer Vision, 2018, pp. 705–720.

[91] T. Pfister, J. Charles, A. Zisserman, Flowing convnets for human pose estimation in videos, in: IEEE International Conference on Computer Vision, 2015, pp. 1913–1921.

[92] J. Song, L. Wang, L. V. Gool, O. Hilliges, Thin-slicing network: A deep structured model for pose estimation in videos, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5563–5572.

[93] J. Charles, T. Pfister, D. R. Magee, D. C. Hogg, A. Zisserman, Personalizing human video pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3063–3072.

[94] Y. Luo, J. S. J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, L. Lin, LSTM pose machines, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5207–5215.

[95] H. Guo, T. Tang, G. Luo, R. Chen, Y. Lu, L. Wen, Multi-domain pose network for multi-person pose estimation and tracking, in: European Conference on Computer Vision Workshops, 2018, pp. 209–216.

[96] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, D. Tran, Detect-and-track: Efficient pose estimation in videos, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 350–359.

[97] S. Li, A. B. Chan, 3d human pose estimation from monocular images with deep convolutional neural network, in: Asian Conference on Computer Vision, 2014, pp. 332–347.

[98] G. Pavlakos, X. Zhou, K. G. Derpanis, K. Daniilidis, Coarse-to-fine volumetric prediction for single-image 3d human pose, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1263–1272.

[99] S. Park, J. Hwang, N. Kwak, 3d human pose estimation using convolutional neural networks with 2d pose information, in: European Conference on Computer Vision, Springer, 2016, pp. 156–169.

[100] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, B. Chen, Synthesizing training images for boosting human 3d pose estimation, in: International Conference on 3D Vision, 2016, pp. 479–488.

[101] H. Yasin, U. Iqbal, B. Krüger, A. Weber, J. Gall, A dual-source approach for 3d pose estimation from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4948–4956.

[102] E. Jahangiri, A. L. Yuille, Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections, in: IEEE International Conference on Computer Vision, 2017, pp. 805–814.

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

[104] J. Martinez, R. Hossain, J. Romero, J. J. Little, A simple yet effective baseline for 3d human pose estimation, in: IEEE International Conference on Computer Vision, 2017, pp. 2659–2668.

[105] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, N. M. Thalmann, Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks, in: IEEE International Conference on Computer Vision, 2019, pp. 2272–2281.

[106] G. Moon, J. Y. Chang, K. M. Lee, V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5079–5088.

[107] K. Iskakov, E. Burkov, V. Lempitsky, Y. Malkov, Learnable triangulation of human pose, in: IEEE International Conference on Computer Vision, 2019.

[108] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1297–1304.

[109] A. Hernández-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, S. Escalera, Graph cuts optimization for multi-limb human segmentation in depth maps, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 726–732.

[110] A. Shafaei, J. J. Little, Real-time human motion capture with multiple depth cameras, in: Conference on Computer and Robot Vision, 2016, pp. 24–31.

[111] M. Ye, X. Wang, R. Yang, L. Ren, M. Pollefeys, Accurate 3d pose estimation from a single depth image, in: IEEE International Conference on Computer Vision, 2011, pp. 731–738.

[112] M. Ye, R. Yang, Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2353–2360.

[113] T. Sharp, C. Keskin, D. P. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. W. Fitzgibbon, S. Izadi, Accurate, robust, and flexible real-time hand tracking, in: ACM Conference on Human Factors in Computing Systems, 2015, pp. 3633–3642.

[114] I. Oikonomidis, N. Kyriazis, A. A. Argyros, Efficient model-based 3d tracking of hand articulations using kinect, in: British Machine Vision Conference, 2011, pp. 1–11.

[115] A. Tkach, M. Pauly, A. Tagliasacchi, Sphere-meshes for real-time hand modeling and tracking, ACM Transactions on Graphics 35 (2016) 1–11.

[116] A. Sinha, C. Choi, K. Ramani, Deephand: Robust hand pose estimation by completing a matrix imputed with deep features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4150–4158.

[117] M. J. Marin-Jimenez, F. J. Romero-Ramirez, R. Muñoz-Salinas, R. Medina-Carnicer, 3d human pose estimation from depth maps using a deep combination of poses, Journal of Visual Communication and Image Representation 55 (2018) 627–639.

[118] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, J. Yuan, A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image, in: IEEE International Conference on Computer Vision, 2019.

[119] S. Maji, L. D. Bourdev, J. Malik, Action recognition from a distributed representation of pose and appearance, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3177–3184.

[120] C. Wang, Y. Wang, A. L. Yuille, An approach to pose-based action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 915–922.

[121] L. Pishchulin, M. Andriluka, B. Schiele, Fine-grained activity recognition with holistic and pose based features, in: German Conference on Pattern Recognition, 2014, pp. 678–689.

[122] A. Yao, J. Gall, L. V. Gool, Coupled action recognition and pose estimation from multiple views, International Journal of Computer Vision 100 (2012) 16–37.

[123] B. X. Nie, C. Xiong, S. Zhu, Joint action recognition and pose estimation from video, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1293–1301.

[124] Y. Li, Y. Li, N. Vasconcelos, Resound: Towards action recognition without representation bias, in: European Conference on Computer Vision, 2018, pp. 513–528.

[125] J. Choi, C. Gao, J. C. Messou, J.-B. Huang, Why can't i dance in the mall? learning to mitigate scene bias in action recognition, in: Advances in Neural Information Processing Systems, 2019, pp. 851–863.

[126] P. Weinzaepfel, G. Rogez, Mimetics: Towards understanding human actions out of context, CoRR abs/1912.07249 (2019).

[127] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

[128] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, K. Saenko, Long-term recurrent convolutional networks for visual recognition and description, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[129] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[130] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4724–4733.

[131] J. Zhu, Z. Zhu, W. Zou, End-to-end video-level representation learning for action recognition, in: International Conference on Pattern Recognition, 2018, pp. 645–650.

[132] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 816–833.

[133] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, Spatio-temporal attention-based LSTM networks for 3d action recognition and detection, IEEE Transactions on Image Processing 27 (2018) 3459–3471.

[134] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. V. Gool, Temporal segment networks: Towards good practices for deep action recognition, in: European Conference on Computer Vision, 2016, pp. 20–36.

[135] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: IEEE International Conference on Computer Vision, 2019, pp. 6202–6211.

[136] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.

[137] C. Si, W. Chen, W. Wang, L. Wang, T. Tan, An attention enhanced graph convolutional LSTM network for skeleton-based action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1227–1236.

[138] R. Zhao, K. Wang, H. Su, Q. Ji, Bayesian graph convolution lstm for skeleton based action recognition, in: IEEE International Conference on Computer Vision, 2019, pp. 6882–6892.

[139] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI Conference on Artificial Intelligence, 2018, pp. 7444–7452.

[140] K. C. Thakkar, P. J. Narayanan, Part-based graph convolutional network for action recognition, in: British Machine Vision Conference, 2018, p. 270.

[141] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12026–12035.

[142] W. Lin, H. Liu, S. Liu, Y. Li, G.-J. Qi, R. Qian, T. Wang, N. Sebe, N. Xu, H. Xiong, et al., Human in events: A large-scale benchmark for human-centric video analysis in complex events, arXiv preprint

arXiv:2005.04490 (2020).

[143] M. Liu, J. Yuan, Recognizing human actions as the evolution of pose estimation maps, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1159–1168.

[144] W. J. McNally, A. Wong, J. McPhee, Star-net: Action recognition using spatio-temporal activation reprojection, in: Conference on Computer and Robot Vision, IEEE, 2019, pp. 49–56.

[145] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.

[146] B. Fernando, E. Gavves, J. O. M., A. Ghodrati, T. Tuytelaars, Modeling video evolution for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5378–5387.

[147] T. S. Kim, A. Reiter, Interpretable 3d human action analysis with temporal convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1623–1631.

[148] X. Zhang, C. Xu, X. Tian, D. Tao, Graph edge convolutional neural networks for skeleton-based action recognition, IEEE Transactions on Neural Networks and Learning Systems (2019) 1–14.

[149] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.

[150] M. Liu, F. Meng, C. Chen, S. Wu, Joint dynamic pose image and space time reversal for human action recognition from videos, in: AAAI Conference on Artificial Intelligence, 2019, pp. 8762–8769.

[151] U. Iqbal, M. Garbade, J. Gall, Pose for action-action for pose, in: IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2017, pp. 438–445.

[152] D. C. Luvizon, D. Picard, H. Tabia, 2d/3d pose estimation and action recognition using multitask deep learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5137–5146.

[153] A. Yan, Y. Wang, Z. Li, Y. Qiao, PA3D: pose-action 3d machine for video recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7922–7931.

[154] J. Zhu, W. Zou, Z. Zhu, L. Xu, G. Huang, Action machine: Toward person-centric action recognition in videos, IEEE Signal Processing Letters 26 (2019) 1633–1637.

[155] L. Huang, Y. Huang, W. Ouyang, L. Wang, Part-aligned pose-guided recurrent network for action recognition, Pattern Recognition 92 (2019) 165–176.

[156] B. Tekin, F. Bogo, M. Pollefeys, H+o: Unified egocentric recognition of 3d hand-object poses and interactions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4511–4520.

[157] S. Yang, J. Liu, S. Lu, M. H. Er, A. C. Kot, Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis, in: European Conference on Computer Vision, Springer, 2020, pp. 769–786.