TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/10 pp973–1001
DOI: 10.26599/TST.2021.9010068
Volume 27, Number 6, December 2022

A Survey of Human Action Recognition and Posture Prediction

Nan Ma*, Zhixuan Wu, Yiu-ming Cheung, Yuchen Guo, Yue Gao, Jiahong Li, and Beijyan Jiang

Abstract: Human action recognition and posture prediction aim to recognize and predict respectively the action and postures of persons in videos. They are both active research topics in computer vision community, which have attracted considerable attention from academia and industry. They are also the precondition for intelligent interaction and human-computer cooperation, and they help the machine perceive the external environment. In the past decade, tremendous progress has been made in the field, especially after the emergence of deep learning technologies. Hence, it is necessary to make a comprehensive review of recent developments. In this paper, firstly, we attempt to present the background, and then discuss research progresses. Secondly, we introduce datasets, various typical feature representation methods, and explore advanced human action recognition and posture prediction algorithms. Finally, facing the challenges in the field, this paper puts forward the research focus, and introduces the importance of action recognition and posture prediction by taking interactive cognition in self-driving vehicle as an example.

Key words: human action recognition; posture prediction; computer vision; human-computer cooperation; interactive cognition

1 Introduction

The development of human society in recent years is known as the "AI Era", in which the development of intelligent technology needs self-learning and self-cognition abilities[1]. The study of human action recognition and posture prediction enables machines to understand human behaviors and intentions and has been broadly applied in many fields[2–6]. Research on human action has two basic topics: Human action recognition and posture prediction.

Human action recognition involves detecting and classifying human actions from a time series (video frames, human skeleton sequences, etc.) that contains complete action execution, as shown in Fig. 1. For example, the result of human body movement can be obtained by detecting the dynamic relationship between the static characteristics of the same frame and several adjacent frames (as shown in Fig. 1, to shake hands).

Human posture prediction automatically recognizes the current posture from temporally incomplete time

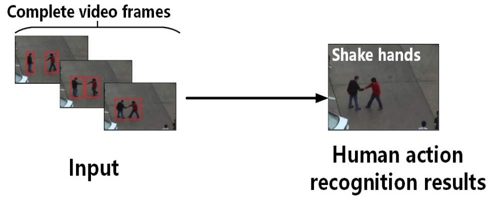


Fig. 1 Example of human action recognition. Human action recognition involves detecting and classifying human actions from a time series (video frames, human skeleton sequences, etc.) that contains complete action execution.

• Nan Ma and Zhixuan Wu are with the Beijing Key Laboratory of Information Service Engineering, the College of Robotics, Beijing Union University, Beijing 100101, China. E-mail: xxtmanan@buu.edu.cn; zhixuanwuYSLYA@163.com.
• Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong 999077, China. E-mail: ymc@comp.hkbu.edu.hk.
• Yuchen Guo is with the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: yuchen.w.guo@gmail.com.
• Yue Gao is with the School of Software, Tsinghua University, Beijing 100084, China. E-mail: gaoyue@tsinghua.edu.cn.
• Jiahong Li and Beiyan Jiang are with the College of Robotics, Beijing Union University, Beijing 100101, China. E-mail: jqrjiahong@buu. edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2021-05-31; accepted: 2021-09-06

series (video frames, human skeleton sequences, etc.), as shown in Fig. 2. For example, self-driving vehicles can predict traffic police's actions, understand police's intentions, and make a judgment in advance (as shown in Fig. 2, to traffic police change lane gesture).

The key difference between human action recognition and posture prediction is when making a judgment about an action[7]. Human action recognition is usually extrapolated from an entire video to an action tag. It is generally used in non-urgent scenarios, such as video surveillance and monitoring[8], and human action analysis[9–11]. Posture prediction is to infer the result before the action is completed, generally using to localize human body joint positions. For example, self-driving vehicles can predict pedestrian movements, conduct interactions between people and machines, understand people's intentions, and avoid dangerous accidents. It is typically used in application scenes with real-time requirements, such as human-vehicle interaction[12, 13], human parsing[14, 15], and human activity monitoring[16].

As noted above, the problems of human action recognition and posture prediction are prevalent research topics. Nevertheless, there are still great challenges for researches:

(1) **Large intra-class variation and inter-class similarity.** For example, in the traffic police dataset, "stop" and "pull over" both involve movements with the right hand raised, this similarity is also involved in other actions. This issue is one of the challenges in recognizing human action recognition. Therefore, a framework that can connect actions needs to be built to adequately identify an action.

(2) **Complex scenarios lead to reduce accuracy.** Since the motion vector is noisy and has substantially reduced resolution, these deviate accuracies. On account of the complexity of scenes, it is impossible to accurately extract the action features. In order to extract action

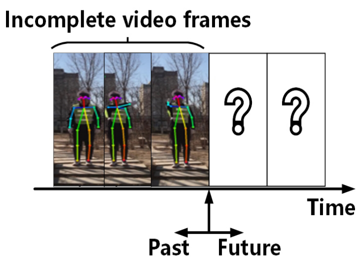


Fig. 2 Example of human posture prediction. Human posture prediction automatically recognizes the current posture from temporally incomplete time series (video frames, human skeleton sequences, etc.).

features adequately, it is also a challenge for human action recognition and posture prediction in complex scenes.

(3) **Long untrimmed sequences exist in many datasets.** Although some existing methods have introduced semi-supervised training methods to some datasets, they cannot make full use of the rich advantages of video context in some aspects and may even impair recognition accuracy if they are not properly designed for raw videos. Moreover, great differences exist in the content of real actions. Therefore, designing human action recognition algorithms that can learn actions from both marked and unmarked data is imperative.

(4) **Long-tailed distributions.** There are lots of data on some human actions (such as standing, walking, sitting, etc.) while little on other human actions (such as traffic police action), and obviously, the significant long-tail distribution is found in data distribution. To overcome the imbalance problem caused by the long-tail distribution, we need to further improve the learning of the classifier and expand the tail data.

Many relevant new ideas, frameworks, and approaches have been proposed in certain area. To better inspire future research and reveal the key trends of these fields, the study attempts to present the background, make a research overview and discuss progresses, datasets, various typical feature representation methods, and a variety of advanced human action recognition and posture prediction algorithms in recent years and other aspects. In addition, it is also pointed out that some future directions of human action recognition and posture prediction. The goal of this paper is to contribute to the field of computer vision, from theory, methodology, and system perspectives. It is believed that this survey can contribute to the field of computer vision, from theory, methodology, and system perspectives as well.

This paper is organized as follows: Section 2 presents commonly used datasets for human action recognition and posture prediction. Section 3 discusses the methods of human action feature representation and human action recognition, and summarizes the common algorithms of human action recognition. Section 4 explores the methods of human posture prediction. Section 5 provides a summary, reviews and looks forward to future research.

2 Common Datasets

From a data perspective, data can be divided into RGB and RGB-D datasets. RGB datasets contain basic color images composed of red, green, and blue channels. Compared with RGB datasets, RGB-D datasets have an extra depth data channel, which provides scene structure. A comparison of benchmark datasets for human action recognition and posture prediction is shown in Fig. 3 and the main characteristics of these datasets are summarized in Table 1. These datasets differ in the number of backgrounds, perspectives, and humans, and are widely used to compare various algorithms. Selecting appropriate datasets for model training is convenient for researchers.

2.1 RGB datasets

(1) UCF-101[25] has 13 320 video samples. It is collected from YouTube with real action videos of 101 types of actions (playing guitar, playing piano, playing violin, etc.). The 101 action categories are divided into 25 groups, and each group can contain 47 action videos. Videos from the same group may have some common characteristics, such as similar backgrounds and perspectives. This dataset is mostly used in single-person or multi-person human action recognition.

(2) J-HMDB[28] has 31 838 annotated frames, which mostly come from movies, with a small proportion coming from public databases. It includes 21 action categories, each containing a minimum of 101 clips (smiling, laughing, chewing, talking, etc.). This dataset is mostly used in single-person or multi-person human action recognition.

(3) Human3.6M[31] has 3.6 million human poses and corresponding images. This dataset is organized into 15 training scenarios including 17 types of actions (discussing, eating, sporting, greeting, etc.). And it also provides synchronized 2D and 3D data (including time of flight, high-quality image, and motion capture data), and accurate 3D human models (body surface scans) of the actors. This dataset is mostly used in 3D posture prediction.

(4) MPII[32] has about 25 000 image samples. It includes 410 types of actions (dancing, walking, running, bicycling, etc.), and more than 40 000 people with annotated human joints. The test set has rich annotations, including occlusion of body parts, 3D torso, and head orientation. This dataset is mostly used in 2D whole body, single-person or multi-person human action recognition or posture prediction.

(5) MS COCO[33] has more than 330 000 image samples. It is mainly derived from complex daily scenes, and the targets in the images are calibrated by precise segmentation. The image includes 91 types of targets (vehicle, person, sports, etc.). And it includes 328 000 videos, and 2 500 000 labels. This dataset is mostly used in 2D whole posture prediction.

(6) Charades[38] has 9848 video samples, which is from daily indoor activities collected through Amazon Mechanical Turk. It includes 157 types of actions (holding, closing door, taking, eating, etc.). The dataset contains 66 500 temporal annotations for 157 action classes, 41 104 labels for 46 object classes, and 27 847 textual descriptions of the videos. This dataset is mostly used in single-person or multi-person human action recognition.

(7) MPI-INF-3DHP[39] has more than 1 300 000 image samples. It includes 8 types of actions (walking, sitting, running, etc.). This multi-view dataset contains both true 3D annotations and a skeleton compatible with the "universal" skeleton of Human3.6M. This dataset is mostly used in posture prediction.

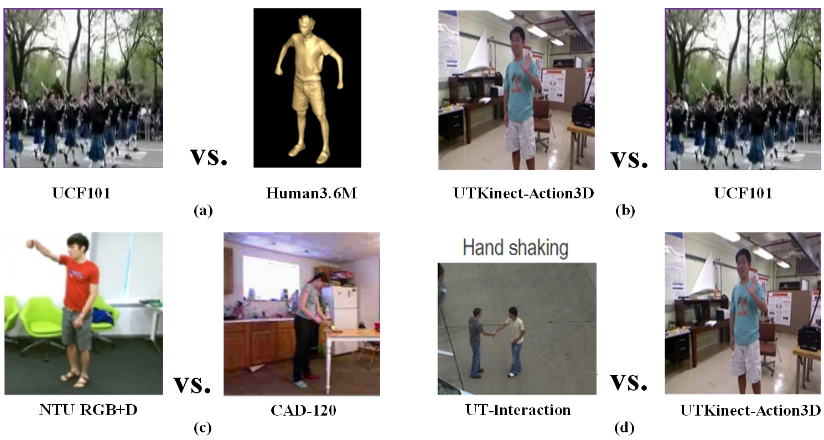


Fig. 3 Compared datasets of various types. (a) Datasets with 2D and 3D; (b) datasets of single-person scene and multi-person scene; (c) datasets of swing shooting and natural shooting scenes; (d) datasets of interactive and datasets without interactive scenes.

Table 1 Common datasets used in action recognition and posture prediction research.

Dataset	Year	Number of samples	Number of action types	Number of views	Type	Task
KTH[17]	2004	2391	6	1	RGB	Action recognition
IXMAS[18]	2006	390	13	5	RGB	Action recognition
Collective Activity[19]	2009	44	5	–	RGB	Action recognition
Hollywood2[20]	2009	3669	12	–	RGB	Action recognition
MuHAVi[21]	2010	952	17	8	RGB	Action recognition
UT-Interaction[22]	2010	20	6	–	RGB	Action recognition
CCV[23]	2011	9317	20	–	RGB	Action recognition
HMDB51[24]	2011	6849	51	–	RGB	Action recognition
UCF101[25]	2012	13 320	101	–	RGB	Action recognition
UTKinect-Action3D[26]	2012	10	10	–	RGB-D	Action recognition
CAD-120[27]	2013	120	10	–	RGB-D	Action recognition
J-HMDB[28]	2013	33 183	21	–	RGB-D	Action recognition
Florence-3D Action[29]	2013	215	9	–	RGB-D	Action recognition
Penn Action[30]	2013	2326	15		RGB	Posture prediction
Human3.6M[31]	2014	3 600 000	17	15	RGB-D	Posture prediction
MPII[32]	2014	25 000	410	–	RGB	Action recognition
MS COCO[33]	2014	328 000	–	–	RGB	Posture prediction
ActivityNet[34]	2015	27 801	203	–	RGB	Action recognition
SYSU-3D Human-Object Interaction[35]	2015	–	12	–	RGB-D	Action recognition
YouTube-8M[36]	2016	8 264 650	4800	–	RGB	Posture prediction
NTU RGB+D[37]	2016	56 880	60	–	RGB-D	Action recognition
Charades[38]	2016	9848	157	–	RGB	Action recognition
MPI-INF-3DHP[39]	2017	>1 300 000	8	14	RGB	Posture prediction
JAAD[40]	2017	346	–	–	RGB	Posture prediction
PKU-MMD[41]	2017	5 400 000	51	–	RGB-D	Action recognition
TotalCapture[42]	2017	1 892 176	4	8	RGB	Posture prediction
Kinetics-600[43]	2018	500 000	600	–	RGB	Action recognition
AVA[44]	2018	–	80	–	RGB	Action recognition
PedX[45]	2019	5000	–	2	RGB	Posture prediction
Moments-in-Time[46]	2020	1 000 000	339	–	RGB	Action recognition
Kinetics-700[47]	2020	650 317	700	–	RGB	Action recognition
NTU RGB+D-120[48]	2020	114 480	120	–	RGB-D	Action recognition
TAPOS[49]	2020	16 294	21	–	RGB	Action recognition
FineGym[50]	2020	–	10	–	RGB	Action recognition

(8) Kinetics-700[47] has 650 317 video samples. It includes 700 types of actions (digging, chasing, spraying, cutting, etc.). For an action class, all clips are from different YouTube videos. This dataset is mostly used in single-person or multi-person human action recognition.

(9) FineGym[50] has about 708 hours of video samples. It includes 10 types of actions (vault, floor exercise, uneven-bars, balance-beam, etc.). It is a new dataset built on top of gymnastics videos and records 303 competitions. This dataset is mostly used in single-person human action recognition.

2.2 RGB-D datasets

(1) UTKinect-Action3D[26] has 10 video samples. It includes 10 types of actions (walking, sitting down, standing up, etc.). Three channels were recorded: RGB, depth, and skeleton joint locations. This dataset is mostly used in single-person human action recognition.

(2) CAD-120[27] has 120 RGB-D action videos. The dataset consists of 10 action types (rinsing mouth, talking on the phone, cooking, etc.) performed by 4 subjects. The videos are captured using the Kinect sensor. Tracked skeletons, RGB images, and depth images are provided in the dataset. This dataset is mostly used in single-person human action recognition.

(3) Florence-3D[29] has 215 video samples. It includes 9 types of actions (waving, drinking from a bottle, answering phone, clapping, tying lace, sitting down, standing up, reading watch, and bowing). 3D data acquisition can be performed through a variety of

methods, including 2D images, collected sensor data and field sensors. Compared with 2D acquisition, 3D acquisition data have more information of a one-dimensional depth, which can improve the accuracy of data recognition. This dataset is mostly used in 3D whole body, single human body action recognition.

(4) NTU RGB + D action recognition[37] has 56 880 video samples. It includes 60 types of actions (reading, writing, clapping, jumping, etc.). The dataset contains RGB video, depth map sequences, 3D bone data and infrared video actions for each sample. The 3D bone data contain the 3D positions of the 25 main body joints of each frame. This dataset is mostly used in single-person or multi-person human action recognition and posture prediction.

3 Human Action Recognition

Human action recognition methods are various, but recognition steps are roughly the same. In the process of human action recognition, on account of the diverse direction and position of human action, it is still a challenging problem to find a general and reliable solution. In modeling, the characteristics or forms of actions should exhibit strong discriminative ability to enable action that have similar temporal and spatial aspects to be distinguished[51]. Human action recognition usually includes two main parts: Human action representation and classification. The feature representation step is performed to extract representative human action information, distinguish it from action videos and convert it into feature vectors[52]. Then the action classification step is performed to identify and label human actions in a large candidate label set. In this section, the discussion will be extended to important human action representation and recognition methods.

3.1 Human action feature representation

Representation of action characteristics is the key to the accuracy in human action recognition. Many kinds of features exist in human actions, and human action feature representation methods mainly deal with the problem of a single feature incompletely describing human action features. Human action feature representation methods can be divided into global feature representation methods and local feature representation methods.

(1) Global feature representation method: The global feature representation method is based on the entire moving human body[53]. Generally, the entire human body of interest is detected by background clipping or tracking. Usually, the silhouette, optical flow, and other information are widely used, which is elaborated below.

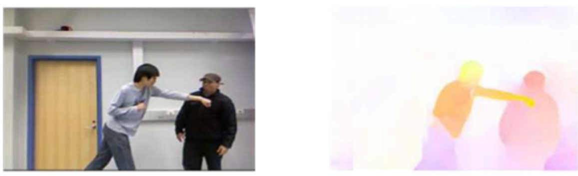
Global silhouette based feature representation methods have been used in the early papers. These methods usually detect human behavior areas by using background clipping, human contour silhouette, etc.; then they extract features for the detected area as behavioral features[54], as shown in Fig. 4. For example, Singh et al.[55] used an adaptive background foreground separation technique to extract motion information and generate human silhouettes from the input video; then, they derived directional feature vectors from contour, and clustered and distinguished different data in vector space. This method could be used for front and side views of most activities. Jiang and Tian[56] proposed a moving human target detection algorithm that combined spatio-temporal background difference and closed contour fitting. The algorithm obtained the initial target area by background difference and constructed a weighted multi-directional Gaussian filter to filter the initial target area to obtain the edge information, finally constructed the closed contour to extract the complete moving target, and marked the target position. Asumang et al.[57] proposed a seed image pruning technique, which mainly described as the maximum angle between boundaries along this contour shared by two parts, such as upper and lower arms. In 2020, Abdelbaky and Aly[58] proposed a Principal Component Analysis Network (PCANet), which used a motion energy template to appropriately represent the time information of the input video, and calculated Multiple Short-Time Motion Energy Image (ST-MEI) templates to capture human movement information. Global silhouette feature representation methods describe information in details and can easily extract the Region of Interest (RoI) in a simple background, it relies heavily on stable segmentation, which may fail in complicated scenes, such as in INRIA[59] and USC-HAD[60] datasets. However, it has difficulty extracting contour features in

Fig. 4 Global silhouette based feature representation. (a) Background image; (b) target image; (c) result image.

a complex background and has large limitations.

Human action has a strong temporal and spatial correlation, so researchers usually use optical flow based representation methods to obtain spatio-temporal features. As shown in Fig. 5, the optical flow based human feature representation methods not only contain the velocity and direction of moving objects, but also the relationship information with the surrounding environments, both of which are important information for recognizing human actions. Ali and Shah[62] proposed action features derived from optical flow for human action recognition in video. It includes divergence, vorticity, symmetric, and antisymmetric optical flow fields. These features are computed by performing Principal Component Analysis (PCA) on the spatio-temporal volumes of the kinematic features. Lu et al.[63] used optical flow information to represent the action information of human behavior, and used a 3D convolution network feature extractor to extract deep RGB features and deep optical flow features. Then, deep RGB features and deep optical flow features are cascaded and fused as a joint feature that is expected to have stronger recognition ability. Ullah et al.[64] used Convolution Neural Networks (CNNs) based optical flow model FlowNet2[65] to extract time features. They fed two consecutive frames into pretrained FlowNet2 CNN model, and then extracted the feature maps from final convolution layers of FlowNet2. In 2020, Rashwan et al.[66] proposed Histograms of Optical Flow Co-Occurrence (HOF-CO) to form the overall motion feature histogram of action. Optical flow based representation methods do not require background representation, thus it is the advantage of dealing with motion background. However, this kind of method will be affected by the noise of the dynamic environment background, which makes it work poorly in conditions with noise, multiple light sources, shadow, and occlusion.

(2) Local feature representation method: The local feature representation method is used to detect and identify parts with significant changes in moving human bodies. Generally, they extract points or blocks of interest in the human body. Unlike global feature representation, local feature representation does not require accurate human positioning and tracking, hence having better stability[67]. Local feature representation can be divided into the following two methods: Local feature detectors and local feature descriptors for human action recognition.



(a) (b)

Fig. 5 Characteristic representation of the optical flow field[61]. (a) Real human action sequence; (b) optical flow.

Local feature detectors for human action recognition construct the entire video as a distribution set of local feature points along the time dimension[68]. These methods are widely used in image retrieval, video analysis, feature matching, lipreading[69], and target recognition[70–72]. Gopalakrishna et al.[73] introduced a method based on Laplace of Gaussian (LOG) and angle-based distance similarity measurement technology for multiple moving target recognition in video sequences. This method extracted feature vectors through appropriate LogGabor approach to test the purpose of moving object image sequence, so it has better recognition accuracy. Gabor filters provide excellent spatial and frequency information for object localization in scenes, which could improve the performance of moving target detection and recognition under certain circumstances such as target occlusion. This method could be used in research on intelligent video surveillance systems. Paul et al.[74] used Harris corner detection and Scale-Invariant Feature Transform (SIFT) to feature matching. This method performed a heuristic search and generated a real nearest neighbor or data point close to it, which improved the result quality of the algorithm. In 2019, Vaghela et al.[75] proposed a Morphological Retina Keypoint Descriptor (MREAK). This method effectively matches key points and detects new key points by implementing selected morphological operations, and adopting a neighborhood sampling model. It also improves the accuracy of key point matching and reduces calculation time. Piergiovanni and Ryoo[76] proposed a learnable convolutional representation flow layer trained in an end-to-end fashion. It computes the flow on a CNN tensor with a smaller spatial size and benefits its speed. Local feature detectors mainly highlight the local particularity of the images and improve the performance of human action recognition. However, the feature points extracted by the local feature detectors are sparse, which will bring a considerable information loss.

Local feature descriptors for human action recognition involve taking the visual observation object as a whole and extracting human action features. Feature descriptors mostly adopt Histogram of Oriented Gradients (HOG), Gray-Level Co-occurrence Matrix (GLCM), Speeded-Up Robust Features (SURF), and Graphics and Intelligence based on Scripting Technology (GIST) or their deformation. The main purpose is to extract descriptors from an RGB video after background subtraction and to create the smallest bounding box around human objects[77]. Zhao et al.[78] proposed an improved SURF, which uses Binary Robust Independent Elementary Feature (BRIEF) to generate feature descriptors, determine matching points and optimize images, and then conduct feature tracking and feature extraction on images. Dusmanu et al.[79] proposed a novel approach to local feature extraction. This method uses a describe-and-detect methodology to describe higher-level information and obtain better features. Experimental results show that it could improve the real-time performance of feature extraction. In 2020, Sadhukhan et al.[77] used an effective sparse filtering method to describe the local feature points of human movement, and reduce the number of features by eliminating redundant features and assigning weight to the remaining features after elimination. Local feature descriptors can handle more complex situations such as occlusion and complex backgrounds. However, these methods gain in robustness comes at the price of higher matching time and memory consumption[79], and will also produce great differences in the same local image content rotation changes.

3.2 Human action recognition methods

After feature representation, action classification should be performed. We divided human action classification methods into two categories: Shallow learning methods and deep learning methods.

(1) Shallow learning methods: Traditional shallow learning methods are usually divided into direct recognition and sequential recognition.

Direct recognition: The method refers to the representation of the entire video sequence as a feature vector. Typical methods include template matching and Support Vector Machine (SVM)[17, 80, 81].

Template matching method aims to identify the object in a given pattern, compare the similarity with the prestored pattern in the recognition process, and select the smallest distance from the test sequence as the recognition result of the test sequence[82–84]. Bobick and Davis[85] first proposed Motion Energy Image (MEI) to describe action recognition by describing how an object moves and where it moves in space. Motion History Image (MHI) was generated based on the action energy map. MHI is a vision-based template method that represents the target action in the form of image brightness by calculating the pixel changes at the same position during a certain time period. Therefore, MHI images can characterize the recent movements of the human body during an action. Weinland et al.[18] proposed Motion History Volumes (MHV) on the basis of MHI for human behavior in multiple calibrated cameras with background subtraction. Zernetsch et al.[86] proposed a method for detecting the starting intention of a bicycle on the basis of MHIs. This method could detect the initial action in the image sequence and classify MHIs frame by frame. It is used to detect the use of a wide-angle stereo camera system at urban intersections. Common template matching methods include Dynamic Time Warping (DTW). Vajda[87] proposed an action recognition method based on fast DTW and feedforward neural network. This method used the modified FastDTW (approximate value of DTW) to classify the movements of various parts of the human body. Chang et al.[88] proposed Discriminative Differentiable Dynamic Time Warping (D3TW) algorithm which is a weakly-supervised method. This method attempts to solve sequence alignment problem and weakly supervised action alignment and segmentation in videos. In 2021, Yang et al.[89] extracted the normalized features of actions and selected inner class center features to construct a template library of actions. They used action detection, action filtering, and adaptive weight shift templates to recognize the actions in video sequences. The experimental recognition accuracy reached 96.74%. The template matching method is easy to understand. However, the algorithm calculation is relatively large, and it does not consider the time and space correlation in the actual situation.

SVM is a widely-used classifier[90–92]. On the basis of its kernel tricks, it can handle high-dimensional feature in a nonlinear space. It has achieved great success in many computer vision and machine learning tasks before deep CNNs and is also widely used in human action recognition. For example, Li[93] proposed a human action recognition method based on fuzzy SVM. Koppula et al.[27] proposed a combination of HOG descriptor and SVM recognizer by using the structure SVM method to solve the problem of joint labeled object provision and human activity in an RGB-

D video. The method described the problem as a Markov Random Field (MRF), where nodes represent objects and sub-activities, and edges represent relationships between objects, their relationships with sub-activities, and their evolution over time. Experiments showed that this method performed well in activity recognition of different marked objects. Uslu and Baydere[94] proposed an SVM-based activity detection. The method that combines feature extraction with a classifier, and proposed the idea that the best classification feature could be determined without experiments on multiple features. This method can be used to help people who need assistance in their daily lives, monitor and detect their activities, and generate their context information, so as to ensure their security. In 2020, Wang et al.[95] proposed to connect local features to form a global representation, and used these features to train Linear Support Vector Machine (LSVM) to perform action recognition using all the contexts of a video. SVM can avoid information redundancy in the process of feature extraction and is capable of accurate and fast classification, thereby accurately recognizing most human movements[96]. However, if the amount of action recognition data to be processed is large, the training time of SVM will be long, and it is difficult to solve multi classification problem.

Sequence recognition: The method uses holistic features from frames to model, and then selects an appropriate classifier on recognition. Some common methods include probabilistic Latent Semantic Analysis (pLSA)[97, 98], Hidden Markov Model (HMM)[99–101], Conditional Random Fields (CRF)[102–104], and so on. Tan et al.[98] proposed a method, which used pLSA for human action recognition. To address the inability of pLSA to guarantee the implicit topic correctness, the algorithm correlated the topic with the action label "one-to-one". And it not only obtained the topic through the supervised method, but also ensured the correctness of the topic during training. Yamato et al.[105] used HMM to determine the number of states that are most suitable for the model on the basis of the number of key poses of human action and to fully express the intrinsic correlation between features. To apply HMMs, they converted a set of time series images into image feature vector sequences, and converted the sequences into symbol sequences by vector quantization[106]. When learning human movements, they optimized the parameters of HMMs so that they can best describe the training sequences in the category. Liu et al.[107] proposed a behavior recognition method based on a human 3D skeleton and Multiple Conditional Random Fields model (MCRF). First, this method divided human action into global action, arm action, and leg action, which can form multiple types of feature sets. Second, it used a CRF model for each feature set based on 3D skeleton division. Third, it integrated all CRF models to obtain the MCRF model, and finally used for behavior recognition. Chereshnev and Kertész-Farkas[108] proposed a method of modeling the distribution of raw data in a half-second context window on the basis of dynamic Bayesian networks for mobile real-time human action recognition. In 2020, Ali and Bouguila[109] proposed variational-based Beta-Liouville hidden Markov models, which considers the prior knowledge, under fitting and over fitting in the training process for human action recognition.

Others: Numerous shallow learning methods such as machine learning exist. Many works utilize unsupervised learning or Semi-Supervised Learning (SSL) framework for human action recognition, which can significantly reduce the labeling effort. Generally, unsupervised learning is when the input data are unlabeled; that is, no corresponding output variable exists. Unlabeled data are used to classify the observations. Shi et al.[110] proposed conditional Variational Auto-Encoder (VAE) to learn model the class-agnostic frame-wise probability conditioned on the frame attention of human actions and solved the action-context confusion issue.

SSL is a learning method that combines supervised learning and unsupervised learning[111, 112]. SSL uses limited labeled samples and a large number of unlabeled samples for model learning. Unlabeled samples can also provide extra knowledge about the data and thus improving model performance because of some probabilistic or geometric information in unlabeled samples or between labeled and unlabeled samples. SSL is also widely used in human action recognition. In fact, because almost unlimited unlabeled video data exist, SSL could be a good way to achieve good accuracy with limited labeled data. Tang et al.[113] proposed a human action recognition method based on Multiview Semi-supervised Learning (MVSL). In this paper, they proposed three kinds of view data, which are skeleton joint point view data, RGB color image view data, and depth image view data. This method used the complementary expression ability of views to comprehensively represent human action, and used the classifier level fusion technology and the prediction ability of three views to effectively solve the problem of unmarked sample confidence evaluation. Pikramenos et al.[114] proposed a semi-supervised automatic retrieval adaptive skeleton method, which not only improved the accuracy of action recognition, but also realized data enhancement. SSL is very advantageous in making full use of unlabeled data, but the current method is not suitable for long-term skeleton sequence learning.

Some researchers have studied sensor based human action recognition methods that can establish links between different data[115–118]. Lei et al.[119] studied the use of RGB-D cameras for fine-grained recognition of kitchen activities. This method is set to combine shape and appearance to locate hands and track changes in object motion to identify objects and their functions. Ranjan et al.[120] confirmed that Radio Frequency Identification (RIFD) is used for location-based behavior recognition, which has higher accuracy for people moving at home. Killijian et al.[121] introduced a new technique for capturing hypotheses about the behavior of human groups. The framework provides a customizable layered approach that allows comparison and inference of models and tracking. Jeong et al.[122] proposed a method of classifying walking activities using eight-foot pressure sensors embedded in smart shoes. These methods have high requirements for sensors, and cannot fully achieve outdoor real-time recognition.

(2) Deep learning methods: Deep learning methods are an abstract representation based on the multilayer representation of the complex relationship between learning data. In these methods, continuity is used to express the close degree between the extracted features and the semantic space. Therefore, the gap between observation, representation, and semantic spaces would be decreased[123]. Deep learning has achieved remarkable results in image recognition, object detection, scene recognition, and other fields with its excellent performance[124–127]. Therefore, many researchers attempted to combine deep learning with human action recognition. Human action recognition is also a video-based computer vision task, and deep learning is expected to achieve promising performance. In most cases, convolutional neural networks are used for visual feature extraction and classification. Recurrent neural networks are also utilized to model the temporal dynamics.

Convolutional neural networks: Simonyan and Zisserman[128] proposed a two-stream convolutional network architecture that incorporates spatial and temporal networks. The spatial stream performs human action recognition from still video frames, while the temporal recognizes human actions in the form of dense optical streams, and then combines the two through post fusion, which is achieved a good recognition effect. In Ref. [129], the traditional CNNs were extended to 3D-CNN with temporal information, and feature calculation was performed on the temporal and spatial dimensions of the video data. The feature map in the convolution process was connected with the data in several consecutive frames[129]. This paper[129] also compared methods based on manual features and methods based on deep learning (RNNs and CNNs), the experimental results show that 3D-CNN is more effective as a representation of spatiotemporal information. Huynh-The and Kim[130] proposed an efficient skeleton action recognition method based on CNNs which used image encoder to convert skeleton coordinate data into image forming data. Li et al.[131] designed a two-dimensional CNNs, which extracted the action mode through the action vector. As a supplement to the pseudo three-dimensional CNNs, CNNs made up for the information lost in the RGB image. In 2020, Yang et al.[132] resolved the costly multi-branch network problem and proposed a generic Temporal Pyramid Network (TPN) at the feature-level. In 2021, Jiang et al.[133] used 3D convolutional neural networks as baseline to recognize action, which includes efficient and temporal efficient two attention modules, and these attention modules could effectively model actions in spatial and temporal. Kumawat et al.[134] proposed spatio-temporal Short-Term Fourier Transform (STFT) convolutional neural networks to reduce parameters for action recognition, which is better than the conventional 3D convolutional layer and its variants by experiments. Liu et al.[135] proposed to a two-stream convolution neural network to recognize single person behavior and interaction behavior, which could improve the accuracy.

Recurrent neural networks (RNNs): RNNs are suitable for temporal problems. Thus, the human action recognition network based on Long-short Term Memory (LSTM) is developed[136, 137]. Doahue et al.[138] proposed a Long-term Recursive Convolutional Neural network (LRCN), which combined CNNs and LSTM network to perform feature representation on video data. Single-frame image information obtained features through CNNs. The output of the CNNs was passed through the LSTM in chronological order, so that the video data are finally characterized in the spatial and temporal dimensions. The network can deal with little input preprocessing and no manual design features. In Ref. [139], CNNs were to obtain the global description. With parameters being shared in time series, both feature aggregation and LSTM architecture were kept as a function of video length. Li et al.[140] proposed an adaptive learning framework based on the RNN tree (RNN-T) for bone based human action recognition. This method used RNN-T model and its associated action category hierarchy was used to distinguish fine-grained action classes that are difficult to handle with a single network, and extend existing models to accommodate new action classes[141]. Liu et al.[141] proposed the global context aware attention LSTM network, which could selectively focus on the information nodes in each frame by using the global context memory unit. They also introduced a recursive attention mechanism, which could gradually improve the attention performance of the network. In 2020, Ji et al.[142] proposed Action Genome to enhance the correlation of movement time characteristics and capture changes between objects and their pairwise relationships while an action occurs. Ullah et al.[143] proposed a Conflux LSTMs Network to recognize actions from multi-view cameras. Compared with the latest data, the experimental results of the benchmark dataset show that the northwest UCLA and MCAD datasets increased by 3% and 2%, respectively. These methods can be used in intelligent video surveillance, human-computer interaction, video retrieval and other applications[1, 144–146]. In 2021, Wang et al.[147] proposed a recurrent neural network for spatiotemporal predictive learning (PredRNN) to learn sequential actions.

3.3 Summary

Table 2 summarizes different improved algorithms mentioned. Findings show that researchers tend to focus on deep learning, but this does not mean that shallow learning is not good. As for the mainstream algorithms for human action recognition, different algorithms have their own structure and datasets. Therefore, different algorithms require different feature representation methods. The applications of the algorithms also have certain differences. The latest methods (Unsupervised Domain Adaptation (UDA)[172], TPN[132], Action Genome[142], Symbiotic Graph Neural Networks (Sym-GNN)[6], etc.) have been used well in action recognition.

At present, action recognition is divided into the following research directions:

(1) **Spatio-temporal networks for action recognition**. The most remarkable feature of human action is that it contains not only static information in the spatial but also motion information in the temporal. Recent works[142–167] improved the understanding of temporal from this task. Some works[173, 174] proposed innovative two-stream fusion schemes, and some studies[175, 176] set up pipelines to connect spatial and temporal information. Others[177–179] studied the spatial hierarchy and temporal series characteristics of skeleton. All in all, these works aim at recognizing the actions of interest that present in both space and time.

(2) **Recognize specific action segments for untrimmed action video.** Action recognition models have been widely studied, most of which are based on trimmed videos, while many video datasets are untrimmed. Therefore, in recent years, weakly supervised learning has been successfully exploited for recognition in untrimmed videos[110, 180].

(3) **Interaction for action recognition.** In real world applications, it includes interactions between humans, between human and objects, and between human and environment. Many existing works are observed to attempt to explore interactions in videos[181, 182].

(4) **Joints correlation in skeleton-based human action recognition.** Human skeleton information is a kind of graph structure data, human actions are usually dependent on two or more neighbor-connected joints. Therefore, it is of great significance for us to explore the dependency information of skeleton based action recognition. Some researchers[183–186] have constructed a more effective graph structure on human skeleton information and achieved great performance improvement.

After years of research on human action recognition, there are still some problems, which we summarize as the following six points and they are sorted out in future work.

(1) Spatio-temporal learning is still an urgent problem. Many works isolate spatial learning and temporal learning, which is why a spatial and temporal fusion occurs at the last level. A loss occurs each time the spatial and temporal features are extracted separately[173]. An effective simulation module can provide valuable clues by integrating motion modeling into the whole spatial-temporal feature learning method.

(2) Using weakly supervised learning method to learn untrimmed dataset need to be further improved. For many applications large amount of video data need to be

Table 2 Common algorithms used in action recognition research.

Dataset	Year	Author	Method	Accuracy (%)
KTH	2011	Zhang et al.[148]	Boosted co-EM (shallow learning)	94.50
KTH	2013	Tan et al. [98]	pLSA (shallow learning)	91.50
KTH	2013	Wang et al.[149]	HMM (shallow learning)	94.17
KTH	2014	Wang et al.[150]	Semi-Supervised (shallow learning)	88.40
KTH	2019	Al-Obaidi and Adhayaratne[151]	Time saliency (deep learning)	99.06
KTH	2019	Almaadeed al.[152]	3DCNN+MHI (deep learning)	99.80
KTH	2020	Basha et al.[153]	3D-CNN (deep learning)	95.27
Weizmann	2012	Zhao et al.[102]	CRF (shallow learning)	91.70
Weizmann	2013	Tan et al.[98]	pLSA (shallow learning)	97.00
Weizmann	2019	Al-Obaidi and Adhayaratne[151]	Time saliency (deep learning)	99.65
Weizmann	2020	Basha et al.[153]	3D-CNN (deep learning)	95.86
UCF-101	2015	Donahue et al.[138]	LRCNs (deep learning)	87.60
UCF-101	2015	Ng et al.[139]	CNNs (deep learning)	88.60
UCF-101	2015	Wu et al.[154]	CNNs and LSTM (deep learning)	91.30
UCF-101	2019	Yeh et al.[155]	Optical Flow (deep learning)	73.60
UCF-101	2019	Shou et al.[156]	DMC-Net (deep learning)	92.30
UCF-101	2019	Zhang et al.[157]	LT3D-CFN (deep learning)	92.87
UCF-101	2019	Li et al.[131]	CNNs (deep learning)	94.30
UCF-101	2020	Alwassel et al.[158]	Cross-Modal Deep Clustering (deep learning)	95.50
UCF-101	2021	Kumawat et al.[134]	T-STFT (deep learning)	94.70
NTU-RGB+D	2014	Vemulapalli et al.[159]	Lie Group (shallow learning)	52.80 (CV), 50.10 (CS)
NTU RGB+D	2018	Liu et al.[141]	GCA-LSTM (deep learning)	84.00 (CV), 76.10 (CS)
NTU RGB+D	2019	Li et al.[160]	AS-GCN (deep learning)	94.20 (CV), 86.80 (CS)
NTU RGB+D	2019	Si et al.[161]	AGC-LSTM (deep learning)	95.00 (CV), 89.20 (CS)
NTU RGB+D	2020	Cheng et al.[162]	4s Shift-GCN (deep learning)	96.50 (CV), 90.70 (CS)
NTU RGB+D	2021	Chen et al.[163]	CTR-GCN (deep learning)	96.80 (CV), 92.40 (CS)
NTU RGB+D	2021	Duan et al.[164]	PoseC3D (deep learning)	97.00 (CV), 99.60 (CS)
HMDB-51	2015	Tran et al.[165]	C3D (deep learning)	51.60
HMDB-51	2019	Shou et al.[156]	DMC-Net (deep learning)	71.80
HMDB-51	2019	Jiang et al.[166]	STM (deep learning)	72.20
HMDB-51	2020	Li et al.[167]	TEA (deep learning)	73.30
HMDB-51	2020	Duan et al.[168]	OmniSource (deep learning)	83.80
HMDB-51	2020	Gowda et al.[169]	SMART (deep learning)	84.36
HMDB-51	2021	Kumawat et al.[134]	T-STFT (deep learning)	71.50
Kinetics	2019	Li et al.[160]	AS-GCN (deep learning)	34.80 (Top-1)
Kinetics	2019	Li et al.[170]	CoST (deep learning)	77.50 (Top-1)
Kinetics	2021	Chen and Huang[171]	ER-ZSAR (deep learning)	42.10 (Zero-Shot) (Top-1)

Note: pLSA, probabilistic Latent Semantic Analysis; HMM, Hidden Markov Model; 3DCNN+MHI, 3-Dimensional Convolution Neural Network + Motion History Images; CRF, Conditional Random Fields; LRCN, Long-term Recursive Convolutional Neural network; DMC, Discriminative Motion Cues; LT3D-CFN, Long-term 3D Convolutional Fusion Network; T-STFT, spatio-Temporal Short-Term Fourier Transform; GCA-LSTM, Global Context-Aware Attention LSTM; AS-GCN, Actional-Structural Graph Convolutional Networks; 4s Shift-GCN, shift graph convolutional network; CTR-GCN, Channel-wise Topology Refinement Graph Convolution Network; DMC-Net, Discriminative Motion Cues; STM, SpatioTemporal and Motion Encoding; SMART, Sampling through Multi-frame Attention and Relations in Time; CoST, Collaborative SpatioTemporal; ER-ZSAR, Elaborative Concepts-Zero-Shot Action Recognition.

analyzed, however, annotating each frame in a video is cumbersome and costly. The previous weakly supervised approaches only provide transcripts. Although the video text can be obtained from script or subtitles, the cost of obtaining these texts is still very high[187]. The spatial and temporal segmentation of untrimmed action videos is processed to develop more robust and efficient action recognition approaches that can automatically learn from unlabeled videos.

(3) The existing methods have the problem of focusing on the interaction between people during recognition. Recent work has exploited human-human interaction

in event, object, and scene modeling, but most works focus on human-human relation recognition in images. Methods that use temporal convolution have very limited temporal reception due to resource challenges. Long-term interaction is important but hard to detect[182] and reduces the accuracy. Finding an appropriate method is necessary to identify interactions correctly in video and use them for action recognition and capturing human-human (human-objects, human-environments) spatial-temporal features and more precise details.

(4) Construct the high-order semantic relationship between joint points[188, 189]. For the higher-order association between joints, such as the association between multi-view joints, but the current methods appropriate modeling methods. We need to design an effective feature extraction method that can consider the coupling relationship between joint points.

(5) Different semantic in different environments for the same action. For example, "waving" can be expressed as "no" when answering questions, and "goodbye" when people are separated. We need to design a reasonable model to analyze and recognize actions in different scenes.

(6) The efficacy of action recognition is directly correlated to the complexity of the network and the computational cost. Despite impressive results on commonly used benchmark datasets, the method consumes a large amount of time and computation costs[162, 190]. A light-weight network needs to be designed to improve the accuracy and speed of identification. For example, specific modules are designed to handle missing bone points to improve accuracy. To reduce the computation, attention mechanism can be used for action recognition. The significant feature map is calculated, and the candidate area of the image is extracted according to the significant area, so as to fully capture the spatial and temporal characteristics of the candidate area of the video, thereby effectively reducing the computational burden of the network.

4 Human Posture Prediction Methods

Unlike human action recognition, the human posture prediction methods are to infer continuous or intermittent actions and predict the whole action before the action completed. In many real scenes (such as rollover), the system can predict the action and make corresponding response, which can effectively reduce the occurrence of accidents. For example, human posture prediction provides an important guarantee for the safe and stable operation of intelligent driving system in the process of self-driving[4]. It can judge the pedestrian's intention (such as walking, jogging, running) and make corresponding decisions. Therefore, human posture prediction is worth studying, and accurate decisions must be made in incomplete movements.

4.1 Skeleton-based human posture prediction methods

Researches usually use skeleton to predict action, for example, Ke et al.[191] proposed a method of skeleton-based action prediction, which aims to predict actions from partial skeleton sequence. Liu et al.[192] focused on streaming 3D skeleton sequences, and proposed dilated convolutional network for online action prediction. Rout et al.[193] used posture analysis and mathematical modeling of the position of adjacent joints of muscles, so that this method could predict and optimize the posture of weight lifting assembly operations. Therefore, we introduce skeleton-based human posture prediction methods at first. The first skeleton based human posture prediction method uses pictorial structures, which has great limitations. This method represents the target object as a collection of "parts", and the combination of these sets can be deformed. This part-based model can well simulate joints. However, this simulation is achieved at the cost of limited expressive power, and global information cannot be used[194]. Through scholarly research, the emergence of CNNs has prompted research on human posture prediction to evolve from traditional methods to deep learning. The location and number of people in an image are usually unknown, which is why we typically use two methods: Top-down and bottom-up.

4.1.1 Top-down human posture prediction

The top-down method first detects people, then estimates each person's parts, and finally calculates each person's posture, as shown in Fig. 6. The representative algorithms are G-RMI, Coarse-Fine Network (CFN), Coarse Proposal Network (CPN), Mask R-CNN, and Regional Multi-person Pose Estimation (RMPE).

(1) G-RMI: G-RMI[195] acquires the bounding box, including single person, through Faster R-CNN detection, and then estimates the posture of a single person. In 2019, Kreiss et al.[196] used this method, and proposed some methods that are particularly suitable for city movements, such as self-driving vehicles and delivery robots. They used a Partial Correlation Field

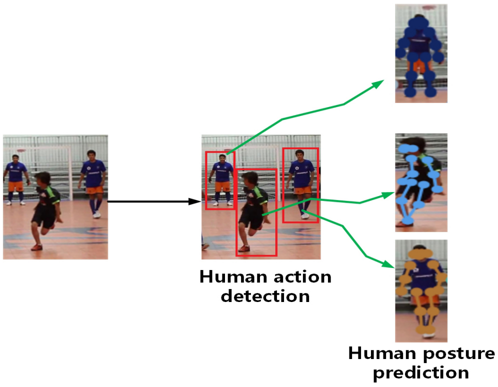


Fig. 6 Top-down human posture prediction.

(PAF) to correlate body parts to form a complete human posture. The G-RMI method pays more attention to the geometric relationship and the output representation of the network, which can be used to predict structured images.

(2) CFN: CFN[197, 198] is better for low resolution human images. In Ref. [197], multi-level monitoring is used to locate key points. Each coarse detector branch is based on CNNs' feature layer, while the fine detector branch is based on multiple feature layers. This method can be used for benchmark testing of multiple tasks, including partial aerial view and human posture prediction. In 2019, Zhang et al.[157] proposed a 3D Convolutional Fusion Network (LT3D-CFN), which could extract features from the spatial and temporal dimensions of a video clip.

(3) CPN: CPN[199] first uses the pedestrian detection framework, then uses the CPN network to regress the key points of each detected pedestrian candidate frame, and finally outputs the results. CPN can solve the problem of multi-person attitude prediction. In 2020, Long et al.[200] proposed a novel Coarse-to-Fine Temporal Proposal (CFTP) which can be combined with CPN, a temporal Convolutional Anchor Network (CAN) and a Proposal Reranking Network (PRN). They conducted extensive experiments on two action benchmarks (THUMOS14 and ActivityNet v1.3) and showed the superior performance of this method.

(4) Mask R-CNN: Mask R-CNN[201] is an extension of Faster R-CNN. For each target of Faster R-CNN, FCN is used for semantic segmentation. The segmentation task is performed at the same time as location and recognition, as shown in Fig. 7. Mask R-CNN predicts segmentation masks on each RoI by adding a small FCN applied to each RoI. In 2019, Huang et al.[203] proposed the network block, which combined

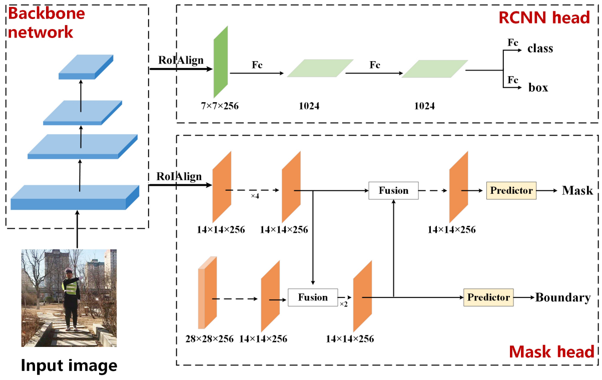


Fig. 7 Mask R-CNN for police gesture posture prediction. Originally shown in Ref. [202].

the instance features with the corresponding prediction mask, and regressed the mask IOU. They improved the quality of generating the prediction mask by accelerating information flow and integrating features of different levels. This method is effective and easy to implement in the instance segmentation mask task. Dabral et al.[204] proposed a mask R-CNN which is based on HG-RCNN. The network took advantage of the hourglass structure in multi-person 3D human pose prediction. First, they estimated 2D key points in each RoI. Then they promoted the estimated key points to 3D. Finally, they placed the estimated 3D pose in a camera coordinate system by using the weak perspective projection hypothesis and the joint optimization of focal length and root translation. In 2020, Tian et al.[202] improved mask R-CNN, which is called boundary-preserving Mask R-CNN (BMask R-CNN). It could improve mask positioning accuracy and the performance better than Mask R-CNN on coco dataset.

(5) RMPE: RMPE[2] is first used to obtain the region frame of the human body through the target detection algorithm. Then, the region box is input into the Space Transformation Network (STN) and single-person pose estimator (SPPE) module to detect the human posture automatically. Then, the training was carried out in the parametric pose non-maximum- suppression (PP-NMS). In the training process, SPPE is used to avoid local optimization and further improve the effect of Symmetric Spatial Transformer Network (SSTN), as shown in Fig. 8. This topic is discussed in Refs. [205, 206], which tried to estimate human posture from RGB. In 2019, Qiao et al.[207] used the RMPE framework to improve the top-down process by adding attention mechanism, that is, to extract features from human posture prediction through an associated network. It also revealed the important

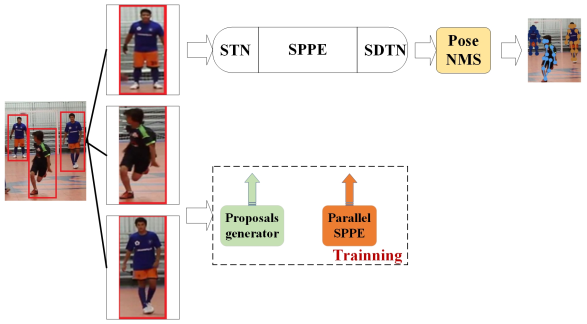


Fig. 8 RMPE. Originally shown in Ref. [2].

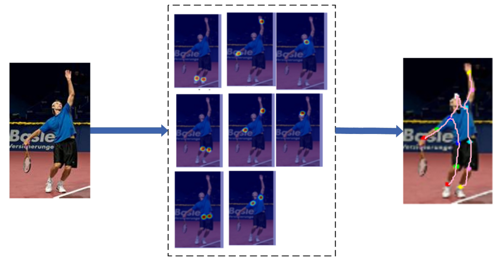


Fig. 9 Bottom-up human posture prediction.

role of joint extraction in human posture prediction. RMPE is universal, and its attention mechanism is suitable for other computer vision tasks, such as semantic segmentation and pedestrian recognition.

Many top-down prediction methods for human body posture prediction are available. Pishchulin et al.[208] proposed a top-down prediction method of a joint model, which generates reasonable posture changes by using a large amount of action capture data. Eichner and Ferrari[209] proposed a new multi-person posture prediction framework, which is based on the predictor that automatically detects the occlusion of human position in the image. The paper extended the graph structure, integrated the occlusion predictor and mutual exclusion, and blocked body parts from different people in the same image area. Reference [2] proposed a top-down method for estimating the pose of multiple people in a complex environment. The top-down SSTN can extract a single region. The top-down method can deal with inaccurate boundary frame and redundant detection, and finally predict everyone's posture.

The research above shows that the top-down human posture prediction method is bound to be constrained by the task of target detection task. Some of these methods have high accuracy but poor real-time performance and are limited by computing resources.

4.1.2 Bottom-up human posture prediction

The bottom-up method detects each part of each person in the image, associates these parts with the examples, and realizes human posture prediction, as shown in Fig. 9. Its representative algorithms are OpenPose, DeepCut, associative embedding, and part segmentation.

(1) OpenPose: OpenPose is one of the most popular bottom-up multi-person posture prediction methods[210, 211]. Reference [212] proposed a bottom-up method for limited detection of multi-person 2D poses in images. They selected the bipartite matching of adjacent joint positions by detecting the appropriate affinity fields of human joint parts and parts respectively, and finally realized the human posture prediction. This paper also proved that the greedy algorithm is enough to generate high quality body posture analysis, even if the number of people in the image increases, efficiency can be maintained. Kato et al.[213] improved the algorithm of the bottom-up method for key points in human body and used the label correction of the teacher model to improve the accuracy by modifying OpenPose. OpenPose can be applied to target detection, semantic segmentation and spatial correlation capture. In 2020, Slembrouck et al.[214] used 2D joint detections per view based on OpenPose to estimate their corresponding 3D positions and solve association problem, so as to allow multiple persons to be tracked at the same time.

(2) Deepcut: Deepcut[215] is also a bottom-up method for estimating human posture. Reference [216] used the distance between candidate nodes to determine whether they are the same important nodes, so as to compress the nodes of various candidate regions into fewer nodes. This method can be used to predict the pose of single and multiple human bodies by using integer linear programming.

(3) Associative embedding: Associative embedding[217, 218] implements end-to-end joint detection and grouping. In this paper, they proposed a CNN monitoring method for detection and grouping. The network outputs the detection and allocation results simultaneously, thus achieving pixel level prediction. This method can solve the problems of machine vision, including multi-person posture prediction, instance segmentation, and multi-target tracking.

(4) Part segmentation: Part segmentation[116, 219] gives a scene and divides it into different categories. Reference [219] proposed a joint solution to deal with semantic object and part segmentation simultaneously, obtained a set of compact segments from the Semantic Compositional Parts (SCP), and constructed an effective Fully Connected conditional Random Field (FCRF) to jointly predict the final object and part label. Jackson et al.[220] proposed CNNs cascade structure. According to a series of positioning, this structure obtained specific posture information of human body. Then, this information was taken as input, and part segmentation was performed.

Human posture can be predicted from the bottom-up, for example, Rangesh and Trivedi[221] proposed a pipeline structure that combines articulated human posture prediction, which used a particle filter with Gaussian Process Dynamics Model (GPDM) to track the joint posture of pedestrians reliably through image sequence, so as to reduce driving accidents of intelligent vehicles. Lin et al.[222] designed a scale perception network jointly trained in a semi supervised way. They predicted pedestrians of a specific scale by matching the perception field of pedestrians with the target scale and using the most appropriate feature maps, which could ensure a large tradeoff between accuracy and speed. Anderson et al.[223] trained the Depth Neural Network (DNN) using scene information on the synthetic datasets, simulated the real pedestrian trajectory, and evaluated the prediction results on various pedestrian trajectory reference datasets. In 2020, Cheng et al.[224] proposed a novel bottom-up human pose estimation method (HigherHRNet), which solved the scale change challenge in multi-person pose estimation and located key points more accurately.

4.2 Time-series-based interaction methods for human posture prediction

Time-series-based interaction methods are also used for human posture prediction[225–227]. This prediction is based on incomplete actions to infer the future behavior of actions, which usually use LSTM or graph neural network to represent time-series prediction or interaction between human and environments respectively. Human posture prediction involves a series of actions in a specific scene. As shown in Fig. 10, the pedestrian crossing behavior prediction is made according to the actions in time periods between T_{n-t} and T_n, and predict the behavior in the later time period T_{n+t}. It is necessary to learn the dependencies between global and local contexts in order to better predict actions, that means this requires semantic analysis of the context and interaction with the surrounding environments. Vondrick et al.[228] proposed a method to anticipate concepts in the future by learning from unlabeled video, and anticipated actions one second in the future and objects five seconds in the future in experiments. Ke et al.[229] proposed a leveraging structural context models, which is used LSTM to process a sequence of global and local interaction contexts, and it is used for human-human interaction prediction. Xue et al.[230] proposed Bi-Prediction, which used bidirectional LSTM to predict trajectory, and it is usually used in crowed scene. Xue et al.[231] proposed hierarchical LSTM to obtain person, social, and scene scale information, which can predict pedestrian postures. Gujjar and Vaughan[232] proposed a method for inferring pedestrian crossing intention, which used a binary action classifier network. Furnari and Farinella[233] focused on the egocentric action anticipation, and proposed an architecture able to anticipate actions at multiple temporal scales using two LSTMs. Saleh et al.[234] proposed spatio-temporal DenseNet to predict pedestrians' intended actions, which could use temporal subsequent frames to predict. Yau et al.[235] proposed a Graph-based Spatiotemporal Interaction Modelling (Graph-SIM) to predict pedestrian crossing action, which used bird's-eye-view to obtain features and model interactions between pedestrians and surrounding traffic environments. Zhang et al.[236] proposed a novel Intuition-Analysis Integrated (IAI) framework inspired by psychological research, which could mitigate the visual gap problem via capturing statistical correlations between past and future. Jaouedi et al.[225] proposed a deep learning model to predict human activities, which is improved RNN (containing LSTM and GRU), because of learning long-term features from sequential and temporal data in RNN. It is not difficult to find that it is a challenge for time-series models to capture the correlation between the past and the future at the visual level and enable the model to predict postures like humans.

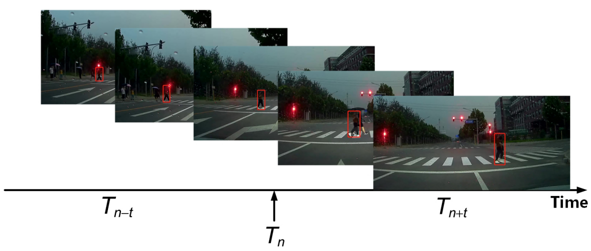


Fig. 10 Crossing scenario for posture prediction.

4.3 Summary

As indicated in Table 3 and the mainstream human posture prediction algorithms, no single algorithm can be applied to all posture prediction problems. Recently, efforts to produce accurate and natural action

Table 3 Common algorithms used in posture prediction research.

Dataset	Year	Author	Method	Value (evaluation metric)
MS COCO	2017	Fang et al.[2]	RMPE (top-down)	61.80% (AP)
MS COCO	2017	He et al.[201]	Mask R-CNN (top-down)	63.10% (AP)
MS COCO	2017	Newell et al.[217]	Associative Embedding (bottom-up)	65.50% (AP)
MS COCO	2017	Huang et al.[197]	CFN (top-down)	72.60% (AP)
MS COCO	2018	Kocabas et al.[237]	PRN (bottom-up)	69.60% (AP)
MS COCO	2018	Chen et al.[199]	CPN (top-down)	73.00% (AP)
MS COCO	2018	Xiao et al.[238]	Simple Baseline (bottom-up)	73.70% (AP)
MS COCO	2021	Cao et al.[212]	OpenPose (bottom-up)	60.50% (AP)
MS COCO	2019	Kreiss et al.[196]	PifPaf (top-down)	66.70% (AP)
MS COCO	2019	Li et al.[239]	MSPN (top-down)	76.10% (AP)
MS COCO	2019	Sun et al.[240]	HRNet-W48 (bottom-up)	77.00% (AP)
MS COCO	2021	Liu et al.[241]	UDP-Pose-PSA (bottom-up)	79.50% (AP)
MPII	2016	Pishchulin et al.[216]	DeepCut (bottom-up)	54.10% (pckh-0.5)
MPII	2016	Insafutdinov et al.[242]	DeeperCut (bottom-up)	59.40% (pckh-0.5)
MPII	2016	Wei et al.[243]	CPM (bottom-up)	87.95% (pckh-0.5)
MPII	2016	Newell et al.[244]	Stacked Hourglass Networks (bottom-up)	90.90% (pckh-0.5)
MPII	2017	Newell et al.[217]	Associative Embedding (bottom-up)	77.50% (mAP)
MPII	2017	Fang et al.[2]	RMPE (top-down)	82.10% (pckh-0.5)
MPII	2017	Chu et al.[245]	CRF (bottom-up)	91.50% (pckh-0.5)
MPII	2021	Cao et al.[212]	OpenPose (bottom-up)	76.50% (AP)
MPII	2019	Sun et al.[240]	HRNet-W48 (bottom-up)	90.80% (pckh-0.5)
MPII	2021	Groos et al.[246]	EfficientPose IV (bottom-up)	91.20% (pckh-0.5)
Human3.6M	2018	Kanazawa et al.[247]	HMR (bottom-up)	56.80 mm (average MPJPE)
Human3.6M	2019	Xu et al.[248]	DenseRaC (bottom-up)	48.00 mm (average MPJPE)
Human3.6M	2019	Zhao et al.[249]	SemGCN (bottom-up)	43.80 mm (average MPJPE)
Human3.6M	2020	Huang et al.[250]	DeepFuse (bottom-up)	37.50 mm (average MPJPE)
Human3.6M	2021	Shan et al.[251]	Pose3D-RIE (bottom-up)	30.10 mm (average MPJPE)
Human3.6M	2021	Reddy et al.[252]	TesseTrack (bottom-up)	18.70 mm (average MPJPE)
JAAD	2019	Gujjar and Vaughan[232]	Res-EnDec (deep learning)	81.14% (AP)
PePScenes	2021	Yau et al.[235]	Graph-SIM (deep learning)	94.40% (accuracy)
3D Pedstria Trajectory	2020	Zhong et al.[253]	SocialGAN (bottom-up)	71.60% (prediction error)

Note: AP, Average Precision; RMPE, Regional Multi-Person Pose Estimation; CFN, Coarse-Fine Network; PRN, Pose Residual Network; CPN, Cascaded Pyramid Network; MSPN, Multi-Stage Pose Estimation Network; IEF, Iterative Error Feedback; CPM, Convolutional Pose Machines; CRF, Conditional Random Field; MPJPE, Mean Per Joint Position Error; HMR, Human Mesh Recovery; PePScenes, Pedestrian Prediction on nuScenes.

sequences have failed. To address this issue, the latest methods (PoseTrack[254], HRNer[240], Exploiting temporal context[255], HigherHRNet[224], Efficient human pose estimation (EfficientPose)[246], Graph-SIM[227], etc.) attempt to improve accuracy in tracking, resolution, context and so on.

At present, human posture prediction is divided into the following research directions:

(1) **Coordinate representation in posture prediction.** The process of decoding the predicted final joint coordinates in the original image space is surprisingly significant for human posture prediction performance. Coordinate[256] and heatmap[257] are two common coordinate representation designs in human posture prediction. Human posture prediction is needed to identify the fine-grained joint coordinates to predict the human posture.

(2) **Predicting poses of multiple humans in real-time.** For example, using multiple cameras to capture the same scene[258] and updating iteratively via cross-view multi-human tracking can efficiently solve the correspondence problem and predict multiple human postures.

(3) **Occlusion problem in human posture prediction.** The performance of many existing methods drops when the target person is occluded by other objects, or the motion is too fast/slow relative to the scale and speed of the training data. To address the

problem, some studies (such as Ref. [259]) proposed a series of methods for human posture prediction.

(4) **Node weight allocation.** The flexibility of each node is different. Grouping the key points and providing a certain weight can help posture prediction. Human posture is predicted by the motion of key points with different weights[205].

(5) **3D pose prediction.** The 3D datasets reduce the learning pressure of the model in 2D attitude estimation, and can form a simple network structure, which occupies less memory of the video card[260]. Thus, some works gradually shifted to research on 3D datasets for human posture prediction.

(6) **Context semantic relation.** Human posture prediction needs to predict the action of the next frame through the global information of the previous frame, which is helpful to realize accurate long-term behavior prediction. For example, some works[192, 261] used context information to enrich temporal and spatial correlation, so as to predict human posture.

After years of research on human posture prediction, some problems remain, which are summarized into seven points and future works below.

(1) The problem of coordinate encoding and decoding (i.e., denoted as coordinate representation) has attracted little attention[257]. However, the method of directly taking the coordinates lacks spatial and contextual information, and heat maps are usually very noisy and incomplete which are reduced in use. A suitable coordinate representation method needs to be found. An interesting task is to explore how to coordinate representation from image models for human posture prediction.

(2) The computational complexity load and the network complexity increase exponentially with the number of cameras used. This condition affects the prediction of human posture. A reasonable way to control the relationship between the amount of calculation and the number of cameras needs to be determined[262]. For example, computational complexity varies only linearly as the number of cameras changes, enabling the applications on large-scale camera systems.

(3) Ambiguous appearance in posture prediction. The accuracy is limited by a number of factors such as ambiguous appearance[263]. The detected joints are ambiguous because the posture prediction is imperfect. Methods such as image fusion can be applied to obtain accurate predictions even when occlusion occurs. Thus, the proposed methods should solve the problem of reducing the accuracy caused by ambiguous appearance.

(4) Many researches focused on detection for skeleton based on human posture prediction. The prediction of human posture infers a human action from temporally incomplete video data, but many papers focused only on detection[211, 264]. In the follow-up work, human posture can be predicted in advance based on the detection of human joints, laying a foundation for the implementation of the application.

(5) Many parameters. Some models[265] have a large number of parameters. Some papers[266, 267] adopt the method of weight sharing, which not only reduces the number of parameters, but also reduces the amount of network calculation. However, when some networks[268] are trained in multi task mode, weight sharing will have a negative impact on each other. How to reduce the number of parameters while ensuring the network quality is a challenging research.

(6) Existing methods perform lower in real scenarios. Many studies[269] predicted postures in specific situations, and the accuracy in real situations is significantly reduced. The accuracy of human posture prediction in real scenarios needs to be improved. For example, the prediction of human posture during self-driving requires high accuracy and real-time performance. Therefore, the accuracy of realistic scenes needs to be improved.

(7) Learning long-term time correlation. Context semantic modeling plays an important role in human posture prediction. Usually, in the task of posture prediction, it is necessary to analyze not only the surrounding environment, but also the previous posture, so as to realize the interaction between posture and environments[7], and then improve the accuracy of human posture prediction.

The human posture prediction effect can be enhanced by choosing the right algorithm through the selection of different feature conditions and application ranges.

5 Conclusion

We surveyed more than 200 papers with over 40 papers coming from the International Conference on Computer Vision and Pattern Recognition (CVPR), and we also cited articles in IEEE International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), and other related conferences or journals, which introduced human action recognition and posture prediction. Subsequently, we expand upon these papers to gather more relevant work. In the last two years, methods (such as UDA[172], TPN[132], Action Genome[142], Sym-GNN[6]) have been used in video understanding tasks, action analysis, and other relevant action recognition fields. For the past two years, methods (such as HRNer[240], Exploiting temporal context[255], HigherHRNet[224], Efficient human pose estimation (EfficientPose)[246], Graph-SIM[227]) have tried to improve accuracy in tracking, resolution, and context, and are applied to human-object interaction detection, human parsing, and other relevant posture prediction. After the emergence of deep learning techniques, researchers have tended to focus on deep learning, whereas previous approaches focused on shallow learning. For example, the multi-stream LSTM derived from LSTM has a higher recognition accuracy than single SVM in nearly two years top meetings. Deep learning methods are also improving. However, differences still exist, even in deep learning. For example, two-stream adaptive graph convolutional network (2S-GCN)[270], Dynamic Directed Graph Convolutional Network (DDGCN)[177], PoseC3D[164], and Channel-wise Topology Refinement Graph Convolution Network (CTR-GCN)[163] were used in action recognition on the NTU-RGB+D dataset, but their accuracy was different, PoseC3D notably outperforms state-of-the-art methods on the NTU RGB+D. Tremendous progress has been made in this area. On the basis of the literature review, this work summarizes the development and practical applications in this field, with mainly helping readers understand human action recognition and posture prediction.

Although human action recognition and posture prediction have been completed through various methods, several research areas may still need to be explored in the future. Research hotspots of human action recognition and posture prediction will focus on the following aspects:

(1) **The importance of data in human action recognition and posture prediction.** At present, many studies are based on a certain sample for training and learning, but the labeled data are limited in reality, and the workload of self-labeling is large. Recently, weak supervised learning and unsupervised learning methods are used to learn unlabeled data. In addition, the number of existing datasets is lack in some requirements. Some studies used GAN-based learning method to expand the dataset. However, how to effectively fill the gap in the field remains unsolved, which is an urgent problem.

(2) **Incapable of effectively modeling the intricate correlations among regions of interest, especially in the case of misalignment and occlusion.** At present, two ways can be used to solve the problem of perspective. One is to use geometric means to normalize the perspective of the feature, and the other is to use multi-view target recognition. However, many false detection results still occur in more complex scenarios. In particular, pedestrian action recognition and prediction need to be more accurate to ensure the safety of self-driving. Solving complex data problems at the scene is a direction that requires future efforts.

(3) **Recognition of unknown human posture.** The type of human poses is an indefinite number. For example, common human poses cannot be used for training in the process of self-driving vehicles research. We can identify and predict human poses that do not exist in the training library through transfer learning, which can be used for future research direction.

(4) **Enhance the research on scene semantic understanding for human action recognition and posture prediction.** For example, in the process of self-driving interactive cognition research, the meaning of pedestrians reaching out at the roadside and in the middle of the road are “taxi” and “stop”, respectively. Because the meaning of action is different in different semantic scenes, how to effectively recognize human action in complex scenes and play a positive role in the interaction between humans and vehicles. Only in this way, self-driving vehicles are no longer a “ghost”, but an interactive wheeled robot. Realtime human action recognition and posture prediction can also be used in the fields of interactive cognition between intelligent robots and humans. How to effectively use human action recognition and posture prediction in an interactive environment is an issue that requires researchers’ constant attention.

Human action recognition and posture prediction are the focus of current computer vision research, especially in intelligent interactive cognition, which have practical application requirements and good application prospects. This paper covers existing work in this area and identifies several related issues that deserve further investigation.

Acknowledgment

The authors wish to thank Dian’en Zhang and Wenjuan Li from Beijing Union University, Beijing, China. We really thank anonymous reviewers’ constructive suggestions. This work was supported by the National Natural Science Foundation of China (Nos. 61871038 and 61931012), the Premium Funding Project for Academic Human Resources Development of Beijing Union University (No. BPHR2020AZ02), and the Generic Pre-research Program of the Equipment Development Department in Military Commission (No. 41412040302).

References

[1] D. Y. Li, N. Ma, and Y. Gao, Future vehicles: Learnable wheeled robots, *Sci. China Inf. Sci.*, vol. 63, no. 9, p. 193201, 2020.

[2] H. S. Fang, S. Q. Xie, Y. W. Tai, and C. W. Lu, RMPE: Regional multi-person pose estimation, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 2353–2362.

[3] H. P. Liu, Y. P. Wu, and F. C. Sun, Extreme trust region policy optimization for active object recognition, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2253–2258, 2018.

[4] L. Chen, N. Ma, P. Wang, J. H. Li, P. F. Wang, G. L. Pang, and X. J. Shi, Survey of pedestrian action recognition techniques for autonomous driving, *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 458–470, 2020.

[5] X. Y. Zhang, C. S. Li, H. C. Shi, X. B. Zhu, P. Li, and J. Dong, AdapNet: Adaptability decomposing encoder-decoder network for weakly supervised action recognition and localization, *IEEE Trans. Neural Netw. Learn. Syst.*, doi: 10.1109/TNNLS.2019.2962815.

[6] M. S. Li, S. H. Chen, X. Chen, Y. Zhang, Y. F. Wang, and Q. Tian, Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction, *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2021.3053765.

[7] Y. Kong and Y. Fu, Human action recognition and prediction: A survey, arXiv preprint arXiv: 1806.11230, 2018.

[8] N. Khalid, M. Gochoo, A. Jalal, and K. Kim, Modeling two-person segmentation and locomotion for stereoscopic action identification: A sustainable video surveillance system, *Sustainability*, vol. 13, no. 2, p. 970, 2021.

[9] T. S. Kim and A. Reiter, Interpretable 3D human action analysis with temporal convolutional networks, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Honolulu, HI, USA, 2017, pp. 1623–1631.

[10] Y. Q. Zhao, W. W. T. Fok, and C. W. Chan, Video-based violence detection by human action analysis with neural network, in *Proc. SPIE* 11321, 2019 *Int. Conf. Image and Video Processing*, *and Artificial Intelligence*, Shanghai, China, 2019, p. 113212N.

[11] G. Li and C. Y. Li, Learning skeleton information for human action analysis using Kinect, *Signal Process.: Image Commun.*, vol. 84, p. 115814, 2020.

[12] M. Mahmood, A. Jalal, and K. Kim, WHITE STAG model: Wise human interaction tracking and estimation (WHITE) using spatio-temporal and angular-geometric (STAG) descriptors, *Multimed. Tools Appl.*, vol. 79, no. 11, pp. 6919–6950, 2020.

[13] L. Vianello, J. B. Mouret, E. Dalin, A. Aubry, and S. Ivaldi. Human posture prediction during physical human-robot interaction, *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 6046–6053, 2021.

[14] Y. R. Bin, X. Cao, X. Y. Chen, Y. H. Ge, Y. Tai, C. J. Wang, J. L. Li, F. Y. Huang, C. X. Gao, and N. Sang, Adversarial semantic data augmentation for human pose estimation, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 606–622.

[15] R. Schnürer, A. C. Öztireli, M. Heitzler, R. Sieber, and L. Hurni, Instance segmentation, body part parsing, and pose estimation of human figures in pictorial maps, *Int. J. Cartogr.*, doi: 10.1080/23729333.2021.1949087.

[16] E. S. L. Ho, J. C. P. Chan, D. C. K. Chan, H. P. H. Shum, Y. M. Cheung, and P. C. Yuen, Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments, *Comput. Vis. Image Underst.*, vol. 148, pp. 97–110, 2016.

[17] A. Schuldt, I. Laptev, and B. Caputo, Recognizing human actions: A local SVM approach, in *Proc. 17th Int. Conf. Pattern Recognition*, Cambridge, UK, 2004, pp. 32–36.

[18] D. Weinland, R. Ronfard, and E. Boyer, Free viewpoint action recognition using motion history volumes, *Comput. Vis. Image Underst.*, vol. 104, nos. 2&3, pp. 249–257, 2006.

[19] W. Choi, K. Shahid, and S. Savarese, What are they doing?: Collective activity classification using spatio-temporal relationship among people, in *Proc. 12th Int. Conf. Computer Vision Workshops*, *ICCV Workshops*, Kyoto, Japan, 2009, pp. 1282–1289.

[20] M. Marszalek, I. Laptev, and C. Schmid, Actions in context, in *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 2929–2936.

[21] S. Singh, S. A. Velastin, and H. Ragheb, MuHAVi: A multicamera human action video dataset for the evaluation of action recognition methods, in *Proc. 7th IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, Boston, MA, USA, 2010, pp. 48–55.

[22] M. S. Ryoo and J. K. Aggarwal, UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA), in *Proc. IEEE Int. Conf. Pattern Recognition Workshops*, Zurich, Switzerland, 2010.

[23] Y. G. Jiang, G. N. Ye, S. F. Chang, D. Ellis, and A. A. Loui, Consumer video understanding: A benchmark database and an evaluation of human and machine performance, in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, Trento, Italy, 2011, pp. 1–8.

[24] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, HMDB: A large video database for human motion recognition, in *Proc. 2011 Int. Conf. Computer Vision*, Barcelona, Spain, 2011, pp. 2556–2563.

[25] K. Soomro, A. R. Zamir, and M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv: 1212.0402, 2012.

[26] L. Xia, C. C. Chen, and J. K. Aggarwal, View invariant

human action recognition using histograms of 3D joints, in *Proc. 2012 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, 2012, pp. 20–27.

[27] H. S. Koppula, R. Gupta, and A. Saxena, Learning human activities and object affordances from RGB-D videos, *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, 2013.

[28] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, Towards understanding action recognition, in *Proc. 2013 IEEE Int. Conf. Computer Vision*, Sydney, NSW, Australia, 2013, pp. 3192–3199.

[29] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses, in *Proc. 2013 IEEE Conf. Computer Vision and Pattern Recognition Workshops*, Portland, OR, USA, 2013, pp. 479–485.

[30] W. Y. Zhang, M. L. Zhu, and K. G. Derpanis, From actemes to action: A strongly-supervised representation for detailed action understanding, in *Proc. 2013 IEEE Int. Conf. Computer Vision*, Sydney, Australia, 2013, pp. 2248–2255.

[31] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, 2014.

[32] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, 2D human pose estimation: New benchmark and state of the art analysis, in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 3686–3693.

[33] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, Microsoft COCO: Common objects in context, in *Proc. 13th European Conf. Computer Vision*, Zurich, Switzerland, 2014, pp. 740–755.

[34] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, ActivityNet: A large-scale video benchmark for human activity understanding, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Boston, MA, USA, 2015, pp. 961–970.

[35] J. F. Hu, W. S. Zheng, J. H. Lai, and J. G. Zhang, Jointly learning heterogeneous features for RGB-D activity recognition, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Boston, MA, USA, 2015, pp. 5344–5352.

[36] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, YouTube-8M: A large-scale video classification benchmark, arXiv preprint arXiv: 1609.08675, 2016.

[37] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 1010–1019.

[38] G. A. Sigurdsson, G. Varol, X. L. Wang, A. Farhadi, I. Laptev, and A. Gupta, Hollywood in homes: Crowdsourcing data collection for activity understanding, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 510–526.

[39] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. P. Xu, and C. Theobalt, Monocular 3D human pose estimation in the wild using improved CNN supervision, in *Proc. 2017 Int. Conf. 3D Vision* (*3DV*), Qingdao, China, 2017, pp. 506–516.

[40] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, Agreeing to cross: How drivers and pedestrians communicate, in *Proc. 2017 IEEE Intelligent Vehicles Symp.* (*IV*), Los Angeles, CA, USA, 2017, pp. 264–269.

[41] C. H. Liu, Y. Y. Hu, Y. H. Li, S. J. Song, and J. Y. Liu, PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding, arXiv preprint arXiv: 1703.07475, 2017.

[42] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, Total capture: 3D human pose estimation fusing video and inertial sensors, in *Proc. British Machine Vision Conf.*, 2017, vol. 2, no. 5, pp. 1–13.

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

[44] C. H. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Q. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al., AVA: A video dataset of spatio-temporally localized atomic visual actions, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6047–6056.

[45] W. Kim, M. S. Ramanagopal, C. Barto, M. Y. Yu, K. Rosaen, N. Goumas, R. Vasudevan, and M. Johnson-Roberson, PedX: Benchmark dataset for metric 3-D pose estimation of pedestrians in complex urban intersections, *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1940–1947, 2019.

[46] M. Monfort, A. Andonian, B. L. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. F. Fan, D. Gutfreund, C. Vondrick, et al., Moments in time dataset: One million videos for event understanding, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 502–508, 2020.

[47] J. Monfort, A. Andonian, B. L. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. F. Fan, D. Gutfreund, C. Vondrick, et al., Moments in time dataset: One million videos for event understanding, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 502–508, 2020.

[48] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot, NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, 2020.

[49] D. Shao, Y. Zhao, B. Dai, and D. H. Lin, Intra-and inter-action understanding via temporal action parsing, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 727–736.

[50] D. Shao, Y. Zhao, B. Dai, and D. H. Lin, FineGym: A hierarchical video dataset for fine-grained action understanding, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 2613–2622.

[51] C. Y. Ding, K. Liu, F. Cheng, and E. Belyaev, Spatio-temporal attention on manifold space for 3D human action recognition, *Appl. Intell.*, vol. 51, no. 1, pp. 560–570,

2021.
[52] H. B. Zhang, Y. X. Zhang, B. N. Zhong, Q. Lei, L. J. Yang, J. X. Du, and D. S. Chen, A comprehensive survey of vision-based human action recognition methods, *Sensors*, vol. 19, no. 5, p. 1005, 2019.
[53] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa, Approaches for global-based action representations for games and action understanding, in *Proc. 2011 IEEE Int. Conf. Automatic Face & Gesture Recognition*, Santa Barbara, CA, USA, 2011, pp. 753–758.
[54] Y. Zhu, J. K. Zhao, Y. N. Wang, and B. B. Zheng, A review of human action recognition based on deep learning, (in Chinese), *Acta Autom. Sin.*, vol. 42, no. 6, pp. 848–857, 2016.
[55] M. Singh, A. Basu, and M. K. Mandal, Human activity recognition based on silhouette directionality, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 9, pp. 1280–1292, 2008.
[56] J. F. Jiang and S. S. Tian, Human detection based on background subtraction and closed contour fitting, (in Chinese), *Comput. Eng. Appl.*, vol. 51, no. 14, pp. 198–202, 2015.
[57] E. K. N. Asumang, X. Zuo, S. Zheng, and H. L. Yu, Human pose estimation based on evidence supporting and sub-graph pruning, in *Proc. 32nd Youth Academic Ann. Conf. Chinese Association of Automation (YAC)*, Hefei, China, 2017, pp. 20–27.
[58] A. Abdelbaky and S. Aly, Human action recognition using short-time motion energy template images and PCANet features, *Neural Comput. Appl.*, vol. 32, no. 16, pp. 12561–12574, 2020.
[59] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in *Proc. 2005 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, pp. 886–893.
[60] M. Zhang and A. A. Sawchuk, USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors, in *Proc. 2012 ACM Conf. Ubiquitous Computing*, Pittsburgh, PA, USA, 2012, pp. 1036–1043.
[61] C. Peng, H. Z. Huang, A. C. Tsoi, S. L. Lo, Y. Liu, and Z. Y. Yang. Motion boundary emphasised optical flow method for human action recognition, *IET Comput. Vis.*, vol. 14, no. 6, pp. 378–390, 2020.
[62] S. Ali and M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, 2010.
[63] T. W. Lu, S. H. Ai, Y. Y. Jiang, Y. D. Xiong and F. Min, Deep optical flow feature fusion based on 3D convolutional networks for video action recognition, in *Proc. 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Guangzhou, China, 2018, pp. 1077–1080.
[64] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, Activity recognition using temporal optical flow convolutional features and multilayer LSTM, *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9692–9702, 2019.
[65] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1647–1655.
[66] H. A. Rashwan, M. A. Garcia, S. Abdulwahab, and D. Puig, Action representation and recognition through temporal co-occurrence of flow fields and convolutional neural networks, *Multimed. Tools Appl.*, vol. 79, no. 45, pp. 34141–34158, 2020.
[67] Y. Zhu, X. Y. Li, C. H. Liu, M. Zolfaghari, Y. J. Xiong, C. R. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, A comprehensive study of deep video action recognition, arXiv preprint arXiv: 2012.06567, 2020.
[68] I. Laptev, On space-time interest points, *Int. J. Comput. Vis.*, vol. 64, no. 2, pp. 107–123, 2005.
[69] X. Liu, Y. M. Cheung, M. Li, and H. L. Liu, A lip contour extraction method using localized active contour model with automatic parameter selection, in *Proc. 20th Int. Conf. Pattern Recognition*, Istanbul, Turkey, 2010, pp. 4332–4335.
[70] L. C. Zhu and Y. Yang, ActBERT: Learning global-local video-text representations, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 8743–8752.
[71] C. I. Patel, D. Labana, S. Pandya, K. Modi, H. Ghayvat, and M. Awais, Histogram of oriented gradient-based fusion of features for human action recognition in action video sequences, *Sensors*, vol. 20, no. 24, p. 7299, 2020.
[72] Z. X. Zheng, G. Y. An, D. P. Wu, and Q. Q. Ruan, Global and local knowledge-aware attention network for action recognition, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 334–347, 2021.
[73] M. T. Gopalakrishna, M. Ravishankar, and D. R. Rameshbabu, Multiple moving object recognitions in video based on log Gabor-PCA approach, in *Recent Advances in Intelligent Informatics*, S. M. Thampi, A. Abraham, S. K. Pal, and J. M. C. Rodriguez, eds. Champaign, IL, USA: Springer, 2014, pp. 93–100.
[74] J. Paul, W. Stechele, M. Kröhnert, and T. Asfour, Resource-aware programming for robotic vision, arXiv preprint arXiv: 1405.2908, 2014.
[75] H. Vaghela, M. Oza, and S. Bagul, MREAK: Morphological retina keypoint descriptor, in *Proc. 2019 Int. Conf. Artificial Intelligence and Information Technology (ICAIIT)*, Yogyakarta, Indonesia, 2019, pp. 10–15.
[76] A. J. Piergiovanni and M. S. Ryoo, Representation flow for action recognition, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 9937–9945.
[77] S. Sadhukhan, S. Mallick, P. K. Singh, R. Sarkar, and D. Bhattacharjee, A comparative study of different feature descriptors for video-based human action recognition, in *Intelligent Computing: Image Processing Based Applications*, J. K. Mandal and S. Banerjee, eds. Singapore: Springer, 2020, pp. 35–52.

[78] H. Zhao, J. W. Dang, S. Wang, Y. P. Wang, and D. C. Gao, Dense trajectory action recognition algorithm based on improved SURF, *IOP Conf. Ser.*: *Earth Environ. Sci.*, vol. 252, no. 3, p. 032179, 2019.

[79] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, D2-Net: A trainable CNN for joint description and detection of local features, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 8084–8093.

[80] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, Learning realistic human actions from movies, in *Proc. 2008 IEEE Conf. Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[81] S. R. Mishra, K. D. Krishna, G. Sanyal, and A. Sarkar, A feature weighting technique on SVM for human action recognition, *J. Sci. Ind. Res.*, vol. 79, no. 7, pp. 626–630, 2020.

[82] V. Bloom, D. Makris, and V. Argyriou, Clustered spatio-temporal manifolds for online action recognition, in *Proc. 22nd Int. Conf. Pattern Recognition*, Stockholm, Sweden, 2014, pp. 3963–3968.

[83] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa, Motion history image: Its variants and applications, *Mach. Vis. Appl.*, vol. 23, no. 2, pp. 255–281, 2012.

[84] A. F. Bobick and J. W. Davis, The recognition of human movement using temporal templates, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.

[85] A. Bobick and J. Davis, Real-time recognition of activity using temporal templates, in *Proc. 3rd IEEE Workshop on Applications of Computer Vision. WACV'96*, Sarasota, FL, USA, 1996, pp. 39–42.

[86] S. Zernetsch, V. Kress, B. Sick, and K. Doll, Early start intention detection of cyclists using motion history images and a deep residual network, in *Proc. 2018 IEEE Intelligent Vehicles Symp.* (*IV*), Changshu, China, 2018, pp. 1–6.

[87] T. Vajda, Action recognition based on fast dynamic-time warping method, in *Proc. 5th Int. Conf. Intelligent Computer Communication and Processing*, Cluj-Napoca, Romania, 2009, pp. 127–131.

[88] C. Y. Chang, D. A. Huang, Y. N. Sui, L. Fei-Fei, and J. C. Niebles, D3TW: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 3541–3550.

[89] X. Yang, D. J. D. Liu, J. Liu, F. R. Yan, P. P. Chen, and Q. Niu, Follower: A novel self-deployable action recognition framework, *Sensors*, vol. 21, no. 3, p. 950, 2021.

[90] X. Z. Wang and S. X. Lu, Improved fuzzy multicategory support vector machines classifier, in *Proc. 2006 Int. Conf. Machine Learning and Cybernetics*, Dalian, China, 2006, pp. 3585–3589.

[91] V. Parameswari and S. Pushpalatha, Human activity recognition using SVM and deep learning, *Int. European Journal of Molecular & Clinical Medicine.*, vol. 7, no. 4, pp. 1984–1990, 2020.

[92] P. Hristov, A. Manolova, and O. Boumbarov, Deep learning and SVM-based method for human activity recognition with skeleton data, in *Proc. 28th National Conf. Int. Participation* (*TELECOM*), Sofia, Bulgaria, 2020, pp. 49–52.

[93] K. Li, Human action recognition based on fuzzy support vector machines, in *Proc. 5th Int. Symp. Computational Intelligence and Design*, Hangzhou, China, 2012, pp. 45–48.

[94] G. Uslu and S. Baydere, Support Vector Machine based activity detection, in *Proc. 21st Signal Processing and Communications Applications Conf.* (*SIU*), Haspolat, Turkey, 2013, pp. 1–4.

[95] H. G. Wang, Z. J. Song, W. Q. Li, and P. C. Wang. A hybrid network for large-scale action recognition from RGB and depth modalities, *Sensors*, vol. 20, no. 11, p. 3305, 2020.

[96] J. Q. Zhou and M. Zhi, A human action recognition method based on MHI and support vector machine, (in Chinese), *Softw. Guide*, vol. 16, no. 2, pp. 36–38, 2017.

[97] L. Chen and H. C. Lu, A new object recognition method based on ML-pLSA model, (in Chinese), *J. Electron. Inf. Technol.*, vol. 33, no. 12, pp. 2909–2915, 2011.

[98] L. Z. Tan, L. M. Xia, J. X. Huang, and S. P. Xia, Human action recognition based on pLSA model, (in Chinese), *J. Natl. Univ. Def. Technol.*, vol. 35, no. 5, pp. 102–108, 2013.

[99] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, Activity recognition and abnormality detection with the switching hidden semi-Markov model, in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition* (*CVPR'05*), San Diego, CA, USA, 2005, pp. 838–845.

[100] C. Sminchisescu, A. Kanaujia, and D. Metaxas, Conditional models for contextual human motion recognition, *Comput. Vis. Image Underst.*, vol. 104, nos. 2&3, pp. 210–220, 2006.

[101] J. W. Xu and Q. Luo, Human action recognition based on mixed Gaussian hidden Markov model, *MATEC Web Conf.*, vol. 336, p. 06004, 2021.

[102] L. Zhao, L. Guo, J. S. Xie, and H. Liu, Video abnormal target description based on CRF model, in *Proc. 2012 Int. Conf. Audio, Language and Image Processing*, Shanghai, China, 2012, pp. 519–524.

[103] K. Liu, L. Gao, N. M. Khan, L. Qi, and L. Guan. A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action recognition, *IEEE Trans. Multimed.*, vol. 23, pp. 64–76, 2020.

[104] T. L. Liu, X. D. Dong, Y. Z. Wang, X. B. Dai, Q. Z. You, and J. B. Luo, Double-layer conditional random fields model for human action recognition, *Signal Process.: Image Commun.*, vol. 80, p. 115672, 2020.

[105] J. Yamato, J. Ohya, and K. Ishii, Recognizing human action in time-sequential images using hidden Markov model, in *Proc. 1992 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Champaign, IL, USA, 1992, pp. 379–385.

[106] P. Zhang, M. Ito, S. I. Ito, and M. Fukumi, Implementation

of EOG mouse using Learning Vector Quantization and EOG-feature based methods, in *Proc. 2013 IEEE Conf. Systems, Process & Control (ICSPC)*, Kuala Lumpur, Malaysia, 2013, pp. 88–92.

[107] H. Liu, L. Guo, B. Yi, and G. Z. Wang, Human activity recognition based on 3D skeletons and MCRF model, (in Chinese), *J. Univ. Sci. Technol. China*, vol. 44, no. 4, pp. 285–291, 2014.

[108] R. Chereshnev and A. Kertész-Farkas, RapidHARe: A computationally inexpensive method for real-time human activity recognition from wearable sensors, *J. Ambient Intell. Smart Environ.*, vol. 10, no. 5, pp. 377–391, 2018.

[109] S. Ali and N. Bouguila, Multimodal action recognition using variational-based Beta-Liouville hidden Markov models, *IET Image Process.*, vol. 14, no. 17, pp. 4785–4794, 2020.

[110] B. F. Shi, Q. Dai, Y. D. Mu, and J. D. Wang, Weakly-supervised action localization by generative attention modeling, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 1006–1016.

[111] H. H. Chen, B. B. Jiang, and X. Yao, Semisupervised negative correlation learning, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5366–5379, 2018.

[112] S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim, and K. M. Lee, Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images, *IEEE Trans. Med. Imaging*, vol. 38, no. 3, pp. 762–774, 2019.

[113] C. Tang, W. J. Wang, X. F. Wang, C. Zhang, and L. Zou, Human action recognition based on multi-view semi-supervised learning, (in Chinese), *Pattern Recognit. Artif. Intell.*, vol. 32, no. 4, pp. 376–384, 2019.

[114] G. Pikramenos, E. Mathe, E. Vali, I. Vernikos, A. Papadakis, E. Spyrou, and P. Mylonas, An adversarial semi-supervised approach for action recognition from pose information, *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17181–17195, 2020.

[115] C. Chen, R. Jafari, and N. Kehtarnavaz, Improving human action recognition using fusion of depth camera and inertial sensors, *IEEE Trans. Hum.-Mach. Syst.*, vol. 45, no. 1, pp. 51–61, 2015.

[116] L. T. Law and Y. M. Cheung, Color image segmentation using rival penalized controlled competitive learning, in *Proc. Int. Joint Conf. Neural Networks*, Portland, OR, USA, 2003, pp. 108–112.

[117] C. Chen, R. Jafari, and N. Kehtarnavaz, A real-time human action recognition system using depth and inertial sensor fusion, *IEEE Sens. J.*, vol. 16, no. 3, pp. 773–781, 2016.

[118] N. Dawar and N. Kehtarnavaz, Action detection and recognition in continuous action streams by deep learning-based sensing fusion, *IEEE Sens. J.*, vol. 18, no. 23, pp. 9660–9668, 2018.

[119] J. N. Lei, X. F. Ren, and D. Fox, Fine-grained kitchen activity recognition using RGB-D, in *Proc. 2012 ACM Conf. Ubiquitous Computing*, Pittsburgh, PA, USA, 2012, pp. 208–211.

[120] J. Ranjan, Y. Yao, E. Griffiths, and K. Whitehouse, Using mid-range RFID for location based activity recognition, in *Proc. 2012 ACM Conf. Ubiquitous Computing*, Pittsburgh, PA, USA, 2012, pp. 647–648.

[121] M. O. Killijian, M. Roy, G. Trédan, and C. Zanon, SOUK: Social observation of human kinetics, in *Proc. 2013 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing*, Zurich, Switzerland, 2013, pp. 193–196.

[122] G. M. Jeong, P. H. Truong, and S. I. Choi, Classification of three types of walking activities regarding stairs using plantar pressure sensors, *IEEE Sens. J.*, vol. 17, no. 9, pp. 2638–2639, 2017.

[123] M. Koohzadi and N. M. Charkari, Survey on deep learning methods in human action recognition, *IET Comput. Vis.*, vol. 11, no. 8, pp. 623–632, 2017.

[124] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in *Proc. CVPR 2011*, Colorado Springs, CO, USA, 2011, pp. 3361–3368.

[125] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, Large-scale video classification with convolutional neural networks, in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1725–1732.

[126] K. Tong, Y. Q. Wu, and F. Zhou, Recent advances in small object detection based on deep learning: A review, *Image Vis. Comput.*, vol. 97, p. 103910, 2020.

[127] W. G. Wang, Q. X. Lai, H. Z. Fu, J. B. Shen, H. B. Ling, and R. G. Yang, Salient object detection in the deep learning era: An in-depth survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2021.3051099.

[128] K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in videos, arXiv preprint arXiv: 1406.2199, 2014.

[129] Z. W. Ding, P. C. Wang, P. O. Ogunbona, and W. Q. Li, Investigation of different skeleton features for CNN-based 3D action recognition, in *Proc. 2017 IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW)*, Hong Kong, China, 2017, pp. 617–622.

[130] T. Huynh-The and D. S. Kim, Data augmentation for CNN-based 3D action recognition on small-scale datasets, in *Proc. 17^{th} Int. Conf. Industrial Informatics (INDIN)*, Helsinki, Finland, 2019, pp. 239–244.

[131] S. Li, Z. C. Zhao, and F. Su, A spatio-temporal hybrid network for action recognition, in *Proc. 2019 IEEE Visual Communications and Image Processing (VCIP)*, Sydney, Australia, 2019, pp. 1–4.

[132] C. Y. Yang, Y. H. Xu, J. P. Shi, B. Dai, and B. L. Zhou, Temporal pyramid network for action recognition, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 588–597.

[133] G. H. Jiang, X. Y. Jiang, Z. J. Fang, and S. S. Chen, An efficient attention module for 3d convolutional neural networks in action recognition, *Appl. Intell.*, vol. 51, no. 10, pp. 7043–7057, 2021.

[134] S. Kumawat, M. Verma, Y. Nakashima, and S. Raman, Depthwise spatio-temporal STFT convolutional neural networks for human action recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2021.3076522.

[135] C. C. Liu, J. Ying, H. M. Yang, X. Hu, and J. Liu, Improved human action recognition approach based on two-stream convolutional neural network model, *Vis. Comput.*, vol. 37, no. 6, pp. 1327–1341, 2021.

[136] Z. F. Zhang, Z. M. Lv, C. Q. Gan, and Q. Y. Zhu, Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions, *Neurocomputing*, vol. 410, pp. 304–316, 2020.

[137] M. Majd and R. Safabakhsh, Correlational convolutional LSTM for human action recognition, *Neurocomputing*, vol. 396, pp. 224–229, 2020.

[138] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, Long-term recurrent convolutional networks for visual recognition and description, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Boston, MA, USA, 2015, pp. 2625–2634.

[139] J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, Beyond short snippets: Deep networks for video classification, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Boston, MA, USA, 2015, pp. 4694–4702.

[140] W. B. Li, L. Y. Wen, M. C. Chang, S. N. Lim, and S. W. Lyu, Adaptive RNN tree for large-scale human action recognition, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 1453–1461.

[141] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva, and A. C. Kot, Skeleton-based human action recognition with global context-aware attention LSTM networks, *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, 2018.

[142] J. W. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, Action genome: Actions as compositions of spatio-temporal scene graphs, in *Proc. 2020 IEEE/CVF Conf. ComputerVision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 10233–10244.

[143] A. Ullah, K. Muhammad, T. Hussain, and S. W. Baik, Conflux LSTMs network: A novel approach for multi-view action recognition, *Neurocomputing*, vol. 435, pp. 321–329, 2021.

[144] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan, and A. A. Abbasi, Human action recognition using fusion of multiview and deep features: An application to video surveillance, *Multimed. Tools Appl.*, doi: 10.1007/s11042-020-08806-9.

[145] J. M. Llaurado-Fons, A. Martinez, F. A. Pujol-López, and H. Mora, An architecture for human action recognition in smart cities video surveillance systems, in *Proc. Int. Research and Innovation Forum 2020*: *Disruptive Technologies in Times of Change*. Champaign, IL, USA: Springer International Publishing, 2021, pp. 51–56.

[146] N. Ma, Y. Gao, J. H. Li, and D. Y. Li, Interactive cognition in self-driving, (in Chinese), *Sci. Sin. Inform.*, vol. 48, no. 8, pp. 1083–1096, 2018.

[147] U. Wang, H. X. Wu, J. J. Zhang, Z. F. Gao, J. M. Wang, P. S. Yu, and M. S. Long. PredRNN: A recurrent neural network for spatiotemporal predictive learning, arXiv preprint arXiv: 2103.09504, 2021.

[148] T. Z. Zhang, S. Liu, C. S. Xu, and H. Q. Lu, Boosted multi-class semi-supervised learning for human action recognition, *Pattern Recognit.*, vol. 44, nos. 10&11, pp. 2334–2342, 2011.

[149] Y. Y. Wang and B. Wang, The conditional random fields method for human action recognition, (in Chinese), *J. Chongqing Univ. Technol.* (*Nat. Sci.*), vol. 27, no. 6, pp. 93–99&105, 2013.

[150] S. Wang, Z. G. Ma, Y. Yang, X. Li, C. Y. Pang, and A. G. Hauptmann, Semi-supervised multiple feature analysis for action recognition, *IEEE Trans. Multimed.*, vol. 16, no. 2, pp. 289–298, 2014.

[151] S. Al-Obaidi and C. Abhayaratne, Temporal salience based human action recognition, in *Proc. 2019 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, UK, 2019, pp. 2017–2021.

[152] N. Almaadeed, O. Elharrouss, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, A novel approach for robust multi human action recognition and summarization based on 3D convolutional neural networks, arXiv preprint arXiv: 1907.11272, 2019.

[153] S. H. S. Basha, V. Pulabaigari, and S. Mukherjee, An information-rich sampling technique over spatio-temporal CNN for classification of human actions in videos, arXiv preprint arXiv: 2002.02100, 2020.

[154] Z. X. Wu, X. Wang, Y. G. Jiang, H. Ye, and X. Y. Xue, Modeling spatial-temporal clues in a hybrid deep learning framework for video classification, in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, Australia, 2015, pp. 461–470.

[155] T. H. Yeh, C. Kuo, A. S. Liu, Y. H. Liu, Y. H. Yang, Z. J. Li, J. T. Shen, and L. C. Fu, ResFlow: Multi-tasking of sequentially pooling spatiotemporal features for action recognition and optical flow estimation, in *Proc. 2019 IEEE/RSJ Int. Conf. Intelligent Robots and Systems* (*IROS*), Macau, China, 2019, pp. 2835–2840.

[156] Z. Shou, X. D. Lin, Y. Kalantidis, L. Sevilla-Lara, M. Rohrbach, S. F. Chang, and Z. C. Yan, DMC-Net: Generating discriminative motion cues for fast compressed video action recognition, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 1268–1277.

[157] Y. Y. Zhang, K. R. Hao, X. S. Tang, B. Wei, and L. H. Ren, Long-term 3D convolutional fusion network for action recognition, in *Proc. 2019 IEEE Int. Conf. Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China, 2019, pp. 216–220.

[158] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, Self-supervised learning by cross-modal audio-video clustering, arXiv preprint arXiv: 1911.12667, 2020.

[159] R. Vemulapalli, F. Arrate, and R. Chellappa, Human action recognition by representing 3D skeletons as points in a

lie group, in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 588–595.

[160] M. S. Li, S. H. Chen, X. Chen, Y. Zhang, Y. F. Wang, and Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 3590–3598.

[161] C. Y. Si, W. T. Chen, W. Wang, L. Wang, and T. N. Tan, An attention enhanced graph convolutional LSTM network for skeleton-based action recognition, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 1227–1236.

[162] K. Cheng, Y. F. Zhang, X. Y. He, W. H. Chen, J. Cheng, and H. Q. Lu, Skeleton-based action recognition with shift graph convolutional network, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 180–189.

[163] Y. X. Chen, Z. Q. Zhang, C. F. Yuan, B. Li, Y. Deng, and W. M. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, arXiv preprint arXiv: 2107.12213, 2021.

[164] H. D. Duan, Y. Zhao, K. Chen, D. Shao, D. H. Lin, and B. Dai, Revisiting skeleton-based action recognition, arXiv preprint arXiv: 2104.13586, 2021.

[165] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in *Proc. 2015 IEEE Int. Conf. Computer Vision*, Santiago, Chile, 2015, pp. 4489–4497.

[166] B. Y. Jiang, M. M. Wang, W. H. Gan, W. Wu, and J. J. Yan, STM: Spatiotemporal and motion encoding for action recognition, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 2000–2009.

[167] Y. Li, B. Ji, X. T. Shi, J. G. Zhang, B. Kang, and L. M. Wang, TEA: Temporal excitation and aggregation for action recognition, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 906–915.

[168] H. D. Duan, Y. Zhao, Y. J. Xiong, W. T. Liu, and D. H. Lin, Omni-sourced webly-supervised learning for video recognition, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 670–688.

[169] S. N. Gowda, M. Rohrbach, and L. Sevilla-Lara, SMART frame selection for action recognition, arXiv preprint arXiv: 2012.10671, 2020.

[170] C. Li, Q. Y. Zhong, D. Xie, and S. L. Pu, Collaborative spatiotemporal feature learning for video action recognition, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 7864–7873.

[171] S. Z. Chen and D. Huang, Elaborative rehearsal for zero-shot action recognition, arXiv preprint arXiv: 2108.02833, 2021.

[172] J. Munro and D. Damen, Multi-modal domain adaptation for fine-grained action recognition, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 119–129.

[173] Q. Q. Xiong, J. J. Zhang, P. Wang, D. D. Liu, and R. X. Gao, Transferable two-stream convolutional neural network for human action recognition, *J. Manuf. Syst.*, vol. 56, pp. 605–614, 2020.

[174] A. Abdelbaky and S. Aly, Two-stream spatiotemporal feature fusion for human action recognition, *Vis. Comput.*, vol. 37, no. 7, pp. 1821–1835, 2021.

[175] X. T. Yang, X. D. Yang, M. Y. Liu, F. Y. Xiao, L. S. Davis, and J. Kautz, STEP: Spatio-temporal progressive learning for video action detection, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 264–272.

[176] Y. X. Li, W. Y. Lin, J. See, N. Xu, S. G. Xu, K. Yan, and C. Yang, CFAD: Coarse-to-fine action detector for spatiotemporal action localization, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 510–527.

[177] M. Korban and X. Li, DDGCN: A dynamic directed graph convolutional network for action recognition, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 761–776.

[178] W. Peng, J. G. Shi, T. Varanka, and G. Y. Zhao, Rethinking the ST-GCNs for 3D skeleton-based human action recognition, *Neurocomputing*, vol. 454, pp. 45–53, 2021.

[179] W. Peng, J. G. Shi, and G. Y. Zhao, Spatial temporal graph deconvolutional network for skeleton-based human action recognition, *IEEE Signal Process. Lett.*, vol. 28, pp. 244–248, 2021.

[180] J. Li, P. Lei, and S. Todorovic, Weakly supervised energy-based learning for action segmentation, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 6242–6250.

[181] T. F. Zhou, W. G. Wang, S. Y. Qi, H. B. Ling, and J. B. Shen, Cascaded human-object interaction recognition, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 4262–4271.

[182] J. J. Tang, J. Xia, X. Z. Mu, B. Pang, and C. W. Lu, Asynchronous interaction aggregation for action detection, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 71–87.

[183] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, Exploiting deep residual networks for human action recognition from skeletal data, *Comput. Vis. Image Underst.*, vol. 170, pp. 51–66, 2018.

[184] P. F. Zhang, C. L. Lan, J. L. Xing, W. J. Zeng, J. R. Xue, and N. N. Zheng, View adaptive neural networks for high performance skeleton-based human action recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, 2019.

[185] W. Peng, X. P. Hong, H. Y. Chen, and G. Y. Zhao, Learning graph convolutional network for skeleton-based human action recognition by neural searching, in *Proc. the 34th AAAI Conf. Artificial Intelligence*, 2020, vol. 34, no. 3, pp. 2669–2676.

[186] Z. Y. Liu, H. W. Zhang, Z. H. Chen, Z. Y. Wang, and W. L. Ouyang, Disentangling and unifying graph

convolutions for skeleton-based action recognition, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 140–149.

[187] M. Fayyaz and J. Gall, SCT: Set constrained temporal transformer for set supervised action segmentation, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 498–507.

[188] J. Q. Dong, Y. B. Gao, H. J. Lee, H. Zhou, Y. F. Yao, Z. J. Fang, and B. Huang, Action recognition based on the fusion of graph convolutional networks with high order features, *Appl. Sci.*, vol. 10, no. 4, p. 1482, 2020.

[189] J. M. Zhou, K. Y. Lin, H. X. Li, and W. S. Zheng, Graph-based high-order relation modeling for long-term action recognition, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 8980–8989.

[190] S. Sudhakaran, S. Escalera, and O. Lanz, Gate-shift networks for video action recognition, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 1099–1108.

[191] Q. H. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, Learning latent global network for skeleton-based action prediction, *IEEE Trans. Image Process.*, vol. 29, pp. 959–970, 2019.

[192] J. Liu, A. Shahroudy, G. Wang, L. Y. Duan, and A. C. Kot, Skeleton-based online action prediction using scale selection network, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1453–1467, 2020.

[193] B. Rout, R. R. Dash, and D. Dhupal, Posture prediction and optimization for a manual assembly operation involving lifting of weights, *Int. J. Simul. Multidisci. Des. Optim.*, vol. 11, p. 1, 2020.

[194] H. Y. Luan, Y. Xiong, J. L. Zhou, and T. P. Qian, From DeepNet to HRNet, here is a full guide to in-depth learning "human posture estimation", blog.nanonets, http://blog.itpub.net/31562039/viewspace-2643565/, 2019.

[195] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, Towards accurate multi-person pose estimation in the wild, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 3711–3719.

[196] S. Kreiss, L. Bertoni, and A. Alahi, PifPaf: Composite fields for human pose estimation, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 11969–11978.

[197] S. L. Huang, M. M. Gong, and D. C. Tao, A coarse-fine network for keypoint localization, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 3047–3056.

[198] F. Richoux and J. F. Baffier, Automatic cost function learning with interpretable compositional networks, arXiv preprint arXiv: 2002.09811, 2021.

[199] Y. L. Chen, Z. C. Wang, Y. X. Peng, Z. Q. Zhang, G. Yu, and J. Sun, Cascaded pyramid network for multi-person pose estimation, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7103–7112.

[200] F. C. Long, T. Yao, Z. F. Qiu, X. M. Tian, T. Mei, and J. B. Luo, Coarse-to-fine localization of temporal action proposals, *IEEE Trans. Multimed.*, vol. 22, no. 6, pp. 1577–1590, 2020.

[201] K. M. He, G. Gkioxari, P. Dollár, and R. Girshick, Mask R-CNN, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 2980–2988.

[202] Z. Tian, C. H. Shen, and H. Chen, Conditional convolutions for instance segmentation, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 282–298.

[203] Z. J. Huang, L. C. Huang, Y. C. Gong, C. Huang, and X. G. Wang, Mask scoring R-CNN, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 6402–6411.

[204] R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, and A. Jain, Multi-person 3D human pose estimation from monocular images, in *Proc. 2019 Int. Conf. 3D Vision* (*3DV*), Quebec City, Canada, 2019, pp. 405–414.

[205] S. Jin, W. T. Liu, E. Z. Xie, W. H. Wang, C. Qian, W. L. Ouyang, and P. Luo, Differentiable hierarchical graph grouping for multi-person pose estimation, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 718–734.

[206] W. A. Mao, Z. Tian, X. L. Wang, and C. H. Shen, FCPose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 9030–9039.

[207] H. X. Qiao, Y. Xu, Z. J. Zhao, J. H. Tian, J. H. Zhang, and C. B. Peng, The network improvement and connection refinement for multi-person pose estimation, in *Proc. 2nd Int. Conf. Artificial Intelligence and Big Data* (*ICAIBD*), Chengdu, China, 2019, pp. 414–418.

[208] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, Articulated people detection and pose estimation: Reshaping the future, in *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 3178–3185.

[209] M. Eichner and V. Ferrari, We are family: Joint pose estimation of multiple persons, in *Proc. 11th European Conf. Computer Vision*, Heraklion, Greece, 2010, pp. 228–242.

[210] Z. Y. Huang, Y. Liu, Y. J. Fang, and B. K. P. Horn, Video-based fall detection for seniors with human pose estimation, in *Proc. 4th Int. Conf. Universal Village* (*UV*), Boston, MA, USA, 2018, pp. 1–4.

[211] A. Viswakumar, V. Rajagopalan, T. Ray, and C. Parimi, Human gait analysis using OpenPose, in *Proc. 5th Int. Conf. Image Information Processing* (*ICIIP*), Shimla, India, 2019, pp. 310–314.

[212] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021.

[213] N. Kato, T. Q. Li, K. Nishino, and Y. Uchida, Improving multi-person pose estimation using label correction, arXiv preprint arXiv: 1811.03331, 2018.

[214] M. Slembrouck, H. Luong, J. Gerlo, K. Schütte, D. Van Cauwelaert, D. De Clercq, B. Vanwanseele, P. Veelaert, and W. Philips, Multiview 3D markerless human pose estimation from OpenPose skeletons, in *Proc. 20th Int. Conf. Advanced Concepts for Intelligent Vision Systems*, Auckland, New Zealand, 2020, pp. 166–178.

[215] M. Rajchl, M. C. H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. J. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, et al., DeepCut: Object segmentation from bounding box annotations using convolutional neural networks, *IEEE Trans. Med. Imaging*, vol. 36, no. 2, pp. 674–683, 2017.

[216] L. Pishchulin, E. Insafutdinov, S. Y. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, DeepCut: Joint subset partition and labeling for multi person pose estimation, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4929–4937.

[217] A. Newell, Z. A. Huang, and J. Deng, Associative embedding: End-to-end learning for joint detection and grouping, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 2274–2284.

[218] Z. H. Yu, J. Zheng, D. Z. Lian, Z. H. Zhou, and S. H. Gao, Single-image piece-wise planar 3D reconstruction via associative embedding, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 1029–1037.

[219] P. Wang, X. H. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, Joint object and part segmentation using deep learned potentials, in *Proc. 2015 IEEE Int. Conf. Computer Vision* (*ICCV*), Santiago, Chile, 2015, pp. 1573–1581.

[220] A. S. Jackson, M. Valstar, and G. Tzimiropoulos, A CNN cascade for landmark guided semantic part segmentation, in *Proc. European Conf. Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 143–155.

[221] A. Rangesh and M. M. Trivedi, When vehicles see pedestrians with phones: A multicue framework for recognizing phone-based activities of pedestrians, *IEEE Trans. Intell. Vehicles*, vol. 3, no. 2, pp. 218–227, 2018.

[222] C. Z. Lin, J. W. Lu, and J. Zhou, Multi-grained deep feature learning for robust pedestrian detection, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3608–3621, 2019.

[223] C. Anderson, X. X. Du, R. Vasudevan, and M. Johnson-Roberson, Stochastic sampling simulation for pedestrian trajectory prediction, in *Proc. 2019 IEEE/RSJ Int. Conf. Intelligent Robots and Systems* (*IROS*), Macau, China, 2019, pp. 4236–4243.

[224] B. W. Cheng, B. Xiao, J. D. Wang, H. H. Shi, T. S. Huang, and L. Zhang, HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 5385–5394.

[225] N. Jaouedi, F. J. Perales, J. M. Buades, N. Boujnah, and M. S. Bouhlel, Prediction of human activities based on a new structure of skeleton features and deep learning model, *Sensors*, vol. 20, no. 17, p. 4944, 2020.

[226] L. Xu, X. Ma, and J. Yan, Scene-perception graph convolutional networks for human action prediction, in *Proc. 2021 Int. Joint Conf. Neural Networks* (*IJCNN*), Shenzhen, China, 2021, pp. 1–8.

[227] Y. Tang, L. Zhao, Z. L. Yao, C. Gong, and J. Yang, Graph-based motion prediction for abnormal action detection, in *Proc. 2nd ACM Int. Conf. Multimedia in Asia*, Singapore, 2021, p. 63.

[228] C. Vondrick, H. Pirsiavash, and A. Torralba, Anticipating visual representations from unlabeled video, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 98–106.

[229] Q. H. Ke, M. Bennamoun, S. J. An, F. Sohel, and F. Boussaid, Leveraging structural context models and ranking score fusion for human interaction prediction, *IEEE Trans. Multimed.*, vol. 20, no. 7, pp. 1712–1723, 2018.

[230] H. Xue, D. Q. Huynh, and M. Reynolds, Bi-Prediction: Pedestrian trajectory prediction based on bidirectional LSTM classification, in *Proc. 2017 Int. Conf. Digital Image Computing*: *Techniques and Applications* (*DICTA*), Sydney, NSW, Australia, 2017, pp. 1–8.

[231] H. Xue, D. Q. Huynh, and M. Reynolds, SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction, in *Proc. 2018 IEEE Winter Conf. Applications of Computer Vision* (*WACV*), Lake Tahoe, NV, USA, 2018, pp. 1186–1194.

[232] P. Gujjar and R. Vaughan, Classifying pedestrian actions in advance using predicted video of urban driving scenes, in *Proc. 2019 Int. Conf. Robotics and Automation* (*ICRA*), Montreal, Canada, 2019, pp. 2097–2103.

[233] A. Furnari and G. Farinella, What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 6251–6260.

[234] K. Saleh, M. Hossny, and S. Nahavandi, Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet, in *Proc. 2019 Int. Conf. Robotics and Automation* (*ICRA*), Montreal, Canada, 2019, pp. 9704–9710.

[235] T. Yau, S. Malekmohammadi, A. Rasouli, P. Lakner, M. Rohani, and J. Luo, Graph-SIM: A graph-based spatiotemporal interaction modelling for pedestrian action prediction, in *Proc. 2021 IEEE International Conference on Robotics and Automation* (*ICRA*), Philadelphia, PA, USA, 2021, pp. 8580–8586.

[236] T. Y. Zhang, W. Q. Min, Y. Zhu, Y. Rui, and S. Q. Jiang, An egocentric action anticipation framework via fusing intuition and analysis, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 402–410.

[237] M. Kocabas, S. Karagoz, and E. Akbas, MultiPoseNet: Fast multi-person pose estimation using pose residual network, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 437–453.

[238] B. Xiao, H. P. Wu, and Y. C. Wei, Simple baselines for human pose estimation and tracking, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 472–487.

[239] W. B. Li, Z. C. Wang, B. Y. Yin, Q. X. Peng, Y. M. Du, T. Z. Xiao, G. Yu, H. T. Lu, Y. C. Wei, and J. Sun, Rethinking on multi-stage networks for human pose estimation, arXiv preprint arXiv: 1901.00148, 2019.

[240] K. Sun, B. Xiao, D. Liu, and J. D. Wang, Deep high-resolution representation learning for human pose estimation, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 5686–5696.

[241] H. J. Liu, F. Q. Liu, X. Y. Fan, and D. Huang, Polarized self-attention: Towards high-quality pixel-wise regression, arXiv preprint arXiv: 2107.00782, 2021.

[242] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, DeeperCut: A deeper, stronger, and faster multi-person pose estimation model, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 34–50.

[243] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, Convolutional pose machines, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 4724–4732.

[244] A. Newell, K. Y. Yang, and J. Deng, Stacked hourglass networks for human pose estimation, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 483–499.

[245] X. Chu, W. Yang, W. L. Ouyang, C. Ma, A. L. Yuille, and X. G. Wang, Multi-context attention for human pose estimation, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 5669–5678.

[246] D. Groos, H. Ramampiaro, and E. A. F. Ihlen, EfficientPose: Scalable single-person pose estimation, *Appl. Intell.*, vol. 51, no. 4, pp. 2518–2533, 2021.

[247] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, End-to-end recovery of human shape and pose, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7122–7131.

[248] Y. L. Xu, S. C. Zhu, and T. Tung, DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 7759–7769.

[249] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, Semantic graph convolutional networks for 3D human pose regression, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 3420–3430.

[250] F. Y. Huang, A. L. Zeng, M. H. Liu, Q. X. Lai, and Q. Xu, DeepFuse: An IMU-aware network for real-time 3D human pose estimation from multi-view image, in *Proc. 2020 IEEE Winter Conf. Applications of Computer Vision*, Snowmass, CO, USA, 2020, pp. 418–427.

[251] W. K. Shan, H. P. Lu, S. S. Wang, X. F. Zhang, and W. Gao, Improving robustness and accuracy via relative information encoding in 3D human pose estimation, arXiv preprint arXiv: 2107.13994, 2021.

[252] N. D. Reddy, L. Guigues, L. Pishchulin, J. Eledath, S. G. Narasimhan, C. M. University, and Amazon, TesseTrack: End-to-end learnable multi-person articulated 3D pose tracking, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 15190–15200.

[253] J. Q. Zhong, H. Sun, W. M. Cao, and Z. H. He, Pedestrian motion trajectory prediction with stereo-based 3D deep pose estimation and trajectory learning, *IEEE Access*, vol. 8, pp. 23480–23486, 2020.

[254] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, PoseTrack: A benchmark for human pose estimation and tracking, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 5167–5176.

[255] A. Arnab, C. Doersch, and A. Zisserman, Exploiting temporal context for 3D human pose estimation in the wild, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 3390–3399.

[256] X. Sun, B. Xiao, F. Y. Wei, S. Liang, and Y. C. Wei, Integral human pose regression, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 536–553.

[257] F. Zhang, X. T. Zhu, H. B. Dai, M. Ye, and C. Zhu, Distribution-aware coordinate representation for human pose estimation, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 7091–7100.

[258] H. Chen, P. F. Guo, P. F. Li, G. H. Lee, and G. Chirikjian, Multi-person 3D pose estimation in crowded scenes based on multi-view geometry, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 541–557.

[259] Z. Zhang, C. Y. Wang, W. H. Qin, and W. J. Zeng, Fusing wearable IMUs with multi-view images for human pose estimation: A geometric approach, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 2197–2206.

[260] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, 3D Human pose estimation: A review of the literature and analysis of covariates, *Comput. Vis. Image Underst.*, vol. 152, pp. 1–20, 2016.

[261] Y. Kong, Z. Q. Tao, and Y. Fu, Deep sequential context networks for action prediction, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 3662–3670.

[262] A. Jalal, I. Akhtar, and K. Kim, Human posture estimation and sustainable events classification via pseudo-2D stick model and K-ary tree hashing, *Sustainability*, vol. 12, no. 23, p. 9814, 2020.

[263] M. Hassan, V. Choutas, D. Tzionas, and M. Black, Resolving 3D human pose ambiguities with 3D scene constraints, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 2282–2292.

[264] J. N. Zhen, Q. Fang, J. M. Sun, W. T. Liu, W. Jiang, H. J.

Bao, and X. W. Zhou, SMAP: Single-shot multi-person absolute 3D pose estimation, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 550–566.

[265] H. Y. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, GHUM & GHUML: Generative 3D human shape and articulated pose models, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 6183–6192.

[266] W. Yang, S. Li, W. L. Ouyang, H. S. Li, and X. G. Wang, Learning feature pyramids for human pose estimation, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 1290–1299.

[267] K. Huang, T. Q. Sui, and H. Wu. 3D human pose estimation with multi-scale graph convolution and hierarchical body pooling, *Multimed. Syst.*, doi: 10.1007/s00530-021-00808-3.

[268] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, 3D human pose estimation in video with temporal convolutions and semi-supervised training, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 7745–7754.

[269] H. S. Koppula and A. Saxena, Anticipating human activities using object affordances for reactive robotic response, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, 2016.

[270] L. Shi, Y. F. Zhang, J. Cheng, and H. Q. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 12018–12027.

Nan Ma received the PhD degree from University of Science & Technology Beijing, Beijing, China, in 2013. She is now a professor at the College of Robotics, Beijing Union University, Beijing, China. Her current research interests include interactive cognition, human posture detection, knowledge discovery, and intelligent system. She is an IEEE Senior Member and a CAAI Senior Member. She has published more than 50 papers in international journals and conferences.

Zhixuan Wu is pursuing the BS degree in software engineering at the College of Robotics, Beijing Union University, Beijing, China. Her current research interests include human action recognition and posture prediction.

Yiu-ming Cheung received the PhD degree from Chinese University of Hong Kong, Hong Kong, China, in 2000. He is a full professor of the Department of Computer Science and an associate director of the Institute of Computational and Theoretical Studies at Hong Kong Baptist University (HKBU), China. He is an IEEE Fellow, IET/IEE Fellow, British Computer Society (BCS) Fellow, Fellow of the Royal Society of Arts (RSA), and Distinguished Fellow of International Engineering and Technology Institute (IETI), Hong Kong, China; as well as the "Chu Tian Scholars" in China. He has published over 250 articles in the high-quality conferences and journals. His research interests include artificial intelligence, intelligent visual computing, pattern recognition, data mining, watermarking, and optimization.

Yuchen Guo received the PhD degree from Tsinghua Univeristy, Beijing, China, in 2018. He is now a research associate in Tsinghua University, Beijing, China. His current research interests inlcude artificial intelligence and large-scale image classification.

Yue Gao received the PhD degree from Tsingua University, Beijing, China, in 2012. He is currently an associate professor in School of Software, Tsinghua University, Beijing, China. His research falls in the field of computer vision, medical image analysis, and machine learning and its applications. He has published more than 90 papers in premier journals and conferences like *TIP*, *TMI*, *TMM*, *TCSVT*, *TNNLS*, *TOMM*, *TGRS*, *TIE*, *Human Brain Mapping*, MICCAI, CVPR, IJCAI, AAAI, ECCV, and ACM Multimedia.

Jiahong Li received the PhD degree from Beijing Institute of Technology, Beijing, in 2017. Currently, he is a lecturer at the College of Robotics, Beijing Union University, Beijing, China. He is a member of the IEEE. His research interests include interactive cognition, distributed estimation fusion, and multi-agent decision theory with a focus on applications for autonomous vehicles.

Beiyan Jiang received the PhD degree from Beijing University of Chemical Technology, Beijing, China, in 2015. She is now a lecturer at Beijing Union University, Beijing, China. Her current research interests include artificial intelligence and unmanned driving.