Pose-based Human Activity Recognition: a review

1st Sameh Neili Boualia
Université de Sousse
Ecole Nationale d'Ingénieurs de Sousse
LATIS-Laboratory of
Advanced Technology and Intelligent Systems
4023, Sousse, Tunisie
Université Tunis El Manar
Ecole Nationale d'Ingénieurs de Tunis
1002, Tunis, Tunisie
sameh.neili@gmail.com

2nd Najoua Essoukri Ben Amara
Université de Sousse
Ecole Nationale d'Ingénieurs de Sousse
LATIS-Laboratory of
Advanced Technology and Intelligent Systems
4023, Sousse, Tunisie
najoua.benamara@eniso.rnu.tn

***Abstract*—This paper serves as a survey and empirical evaluation of the state-of-the-art in activity recognition methods using still RGB images and/or videos. Understanding human activities from videos or still images is a challenging task in computer vision domain. Identifying the action or activity being accomplished automatically and then recognizing it represents the prime goal of an intelligent video system. Human Activity Recognition arises in various application domains varying from human computer interfaces, health care monitoring to surveillance and security. Despite the ongoing efforts in the domain, these tasks remained unsolved in unconstrained environments and face many challenges such as occlusions, variations in clothing and background clutter. Recently, numerous deep learning algorithms have been proposed to solve traditional artificial intelligence problems. They have shown great advances, in particular for pose estimation task since they can extract appropriate features while jointly performing discrimination. In this paper, we provide a detailed review of recent and state-of-the-art research advances in the field of human activity recognition. We propose a categorization of human activity methodologies and discuss their advantages and limitations. In particular, we divide feature representation methods into global, local and body modeling. Then, human activity classification approaches are arranged into three categories, which reflect how they model human activities: template-based, generative and discriminative. Moreover, we provide a comprehensive analysis of pose-based human activity recognition where both conventional and deep learning-based human pose estimation approaches are reported. Finally, we discuss the open-challenges in this field and endeavor to provide possible solutions.**

***Index Terms*—Human Pose Estimation, ConvNets, Deep Learning, Human Activity Recognition**

I. Introduction

Recognizing human activity from videos sequences or still images is a widely studied computer vision problem. It has many application domains such as video surveillance [1,2], human computer interaction [3] and sport performance analysis [4]. However, unless the many research works being involved, there are still many challenges ahead especially, the high number of human body freedoms degree, clothing changes and body shapes which affect limbs appearance. According to [5], Human Activity Recognition (HAR) can be divided into three levels of representation: core technology (low-level), HAR systems (mid-level) and applications (high-level) as shown in Fig.1. In traditional HAR systems, a pre-

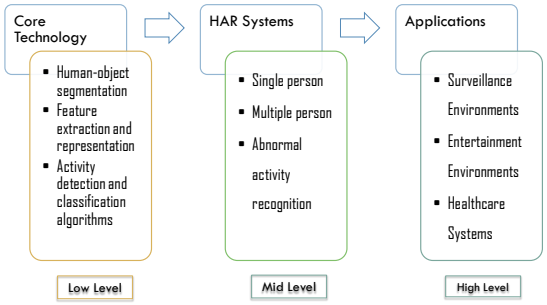

Fig. 1. Taxonomy of a HAR system [5]

processing step of human-object segmentation is obviously necessary. The object is first segmented out from each frame. The human characteristics, such as shape, silhouette, poses, or body motions are then properly extracted and represented by a set of features. Thus, an activity detection or classification algorithm is applied on the extracted features in order to recognize the various human activities.

The remainder of this paper is organized as follows. In section II, we present an overview of existing human-object segmentation methods. We discuss then different feature extraction and representation methods in section III, as well as activity detection and classification algorithms in section IV. In section V, we divide Human Pose Estimation (HPE) techniques into two main categories: i) conventional approaches and ii) deep learning-based approaches, where we evoke their contributions and limitations. Moreover, we provide a comprehensive analysis of the existing, publicly available human activity and pose estimation datasets in in section VI. Finally, we report the characteristics of future research directions in section VII and present some open issues on human activity recognition.

978-1-5386-7747-6/19/$31.00 ©2019 IEEE

II. Human-object segmentation methods

The segmentation process aims to understand the visual content of an image. Assigning predefined object or scene labels to each pixel, this technique translates the image input into a segmentation map output. Most state-of-the-art segmentation techniques share a similar two-stage paradigm: first, a Fully Convolutional Network (FCN) stage is used to segment the foreground object; second, based on the first frame, this FCN is fine-tuned for several hundred of iterations to adapt the model to the whole video. In [6], the authors proposed a novel segmentation approach based on a meta neural network called: *modulator*. It consists of three basic components: a segmentation network and two spatial and visual modulators that produce a set of parameters manipulating the feature maps to adapt the network model to segment a specific object. The advantage of the proposed work is that it uses a single forward pass when adapting the model to the appearance of a specific object. Unlike traditional approaches, Zhang et al. [7] presented a brand new segmentation framework based on human poses rather than proposal region detection. Later on, the authors of [8] presented a CNN architecture for semi-automatic segmentation that turns extreme clicking annotations into accurate object masks named DEXTR. Cai et al. [9] also, exploited a modified CNN architecture as an encoder-decoder block to generate robust salient feature maps for object segmentation. They called the proposed end-to-end generic model as Metric Expression Network (MEnet), used to resolve a specific occlusion problem: distortion. Recently, Chen et al. [10] proposed a novel segmentation technique named: Atrous Spatial Pyramid Pooling (ASPP). As its name points to, this technique is based on a new type of convolution: *Atrous*, becoming the most popular one in real-time segmentation domain. Based on dilation rate, it allows to explicitly control the resolution where feature responses are computed with DCNN. Therefore, objects are robustly segmented at multiple scales.

III. Feature extraction and representation methods

Feature extraction and representation process is a fundamental step not only in HAR domain, but in many other computer vision tasks. At first, extracted characteristics must be represented in some form of features. Principally, representation forms can be categorized into three main groups as shown in Fig.2: i) global, ii) local and iii) body modeling representations.

For global representations, space-time volume (STV) methods [11] were first proposed. As indicated by their name, they are based on space-time information. Features are built by concatenating the consecutive silhouette of objects along the time axis. Observing that the frequency can add some information to represent features, the Discrete Fourier Transform (DFT) has been widely used to represent the geometric structure of the object [12]. Thus, global features provide a proper way to combine different types of information (spatial, temporal, frequency). However, they are gradually outdated since they

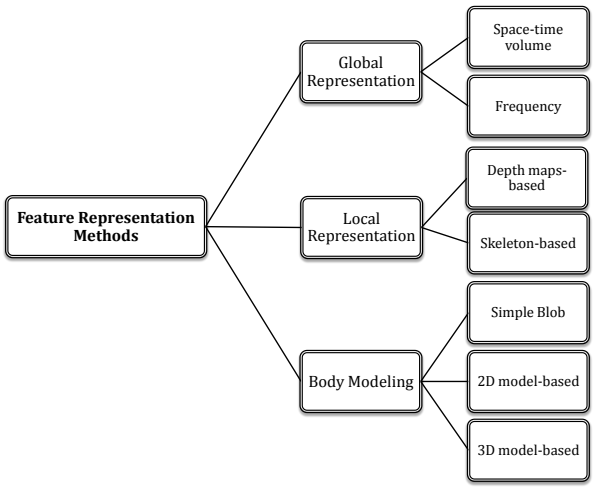

Fig. 2. Principal HAR feature representation methods

are sensitive to viewpoint changing, occlusions and noise. To square up to those problems, some research work were conducted to exploit depth maps. Compared with conventional color images, depth frames contain additional coordinates and are therefore more informative. As example, Xia et al. [13] proposed a skeleton-based representation named HOJ3D: the spherical histograms of 3D locations of selected joints. The encoded features are fed to a Hidden Markov model (HMM) for classification. Yang and Tian [14] proposed a new type of feature named EigenJoints. The 3D joint positions were employed to characterize three kinds of activity information including posture, motion and offset features. To reduce redundancy and noise, a Principal Component Analysis (PCA) was further employed and the efficient eigen-vectors were selected. Besides, Li et al. [15] employed the action graph model, where activities were represented using several salient postures. All activities shared the same posture sets and each posture was characterized as a bag of 3D points from the depth maps. However, including all the 3D points has a high computational cost. Therefore, the authors have taken refuge to sample the representative 3D points achieving over 90% recognition accuracy by sampling about 1% points of their own database containing twenty actions. Later on, Jalal et al. [16] have extracted multi-features from depth videos through 3D human silhouettes. Recently, in order to recognize human actions, Seddik et al. [17] have addressed the problem of efficiently combining three modalities which were joints, RGB frames and depth maps. They combined the performance of local descriptors with the strength of global bags-of-visual-words representations. They have also studied the performance of multiple fusion schemes guided by different features concatenations, Fisher vectors representations and iterative scores fusion.

Besides global and local representation methods, some research work were proposed to benefit from human body modeling in order to offer more efficient and discriminative

features. Usually, body modeling can be classified into three main groups: simple blob modeling, 2D model-based and 3D model-based methods. At this stage, since we have to calculate 2D/3D joint positions of human body, pose estimation problem is normally invoked. Simple blob representation is a model-free method. Indeed, human body is simply represented by ellipse [18], or stick figures [19]. This class relies on extensive training using ground truth (GT) data obtained generally by commercial motion systems [20]. Freifeld et al. [21] defined a new *contour person* model of the human body that has the expressive power of a detailed 3D model and the computational benefits of a simple 2D part-based model. The model was used for pose estimation and segmentation tasks. Later on, Chan et al. [22] presented a novel 2D model for human motion estimation with a focus on simple three-body-segment components and a 2D stick model construction as the motion posture resemblance. The model was applied to short temporal daily activities including walking, running, and jumping obtained from publicly available video and experimental captures of Yoga motion activity. For 3D model-based approaches, Lillo et al. [23] recognized complex activities composed of sequential or simultaneous atomic actions characterized by body motions of a single actor. They tackled this problem by introducing a hierarchical compositional 3D human body model. Their results have shown the benefits of using a hierarchical model that exploits the sharing of body poses into atomic actions, and atomic actions into activities. Recently, Rahmani et al. [24] proposed a Robust Non-Linear Knowledge Transfer Model (R-NKTM) for HAR from novel views. The proposed R-NKTM was a deep fully-connected neural network that transferred knowledge of human actions from any unknown view to a shared high-level virtual view. The R-NKTM was learned from 2D projections of dense trajectories of synthetic 3D human models fitted to real motion capture data and generalized to real videos of human actions.

IV. Activity Detection and Classification Approaches

In the following, we are interested on the third component of core technology level of a general HAR taxonomy: activity detection/classification. Generally, those approaches can be considered as a three-level categorization as shown in Fig.3: template-based methods, generative model-based, and discriminative model-based approaches.

For template-based methods, as Dynamic Time Warping (DTW) or Template-matching, they are generally based on similarity measure between two input sequences. Despite the simplicity of application, those methods are considered computationally expensive since their need to extensive templates. In order to overcome this cost issue, model-based approaches are then proposed. They can be divided into two types: generative and discriminative model-based approaches. On one hand, generative model-based techniques are essentially based on probability computation. First, they learn a model of the joint represented by the probability P(X,Y), where X is the input observations and Y represents the result label. Then, $P(X|Y)$ is calculated using Bayes rules and the most likely label Y is picked. The most widely used methods are HMM [25] and Dynamic Bayesian Networks (DBN) [26]. On the other hand, discriminative model-based methods determine the result label Y by learning the conditional probability distribution $P(Y|X)$ directly from a given observed fixed-lenght feature vector X. They can be illustrated basically with two types: i) classic approaches: such as: SVM [27], Relevance Vector Machine (RVM) [28], Artificial Neural Networks (ANN) and ii) deep learning architecture-based approaches. Basically, the deep learning architectures can be classified into four groups, namely Deep Neural Networks (DNNs), Convolutional Neural Networks (ConvNets/CNNs), Recurrent Neural Networks (RNNs), and some emergent architectures [29].

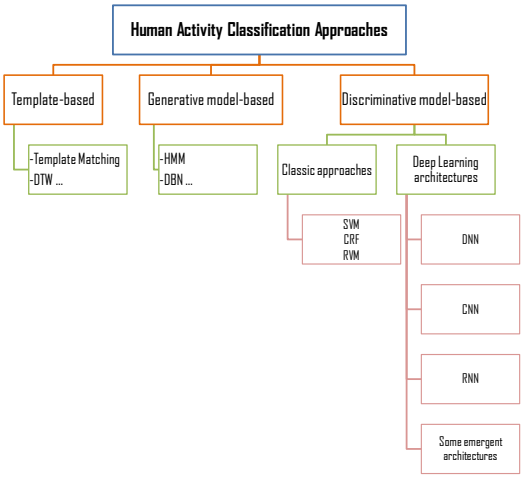

Fig. 3. Our proposed categorization of human activity classification approaches

V. Pose-based HAR

Human poses are important cues for analysis of videos including people and there is a strong evidence that representations based on body pose are highly effective for a variety of tasks such as activity recognition, content retrieval and social signal processing. Scientifically speaking, HPE refers to the process of estimating the configuration of the human body parts (3D pose) or their projection onto image plane (2D HPE). It includes all the human body related problems ranging from the whole human body pose parsing to the detailed body parts localization. Generally, HPE approaches can be divided into two main groups, as presented in Fig.4: conventional HPE approaches, and deep learning based approaches. For conventional approaches, they can be classified into three main groups: i) generative, ii) discriminative and iii) hybrid approaches. However, deep learning based approaches can be divided into two classes according to the handling manner of input images: holistic processing or part-based processing.

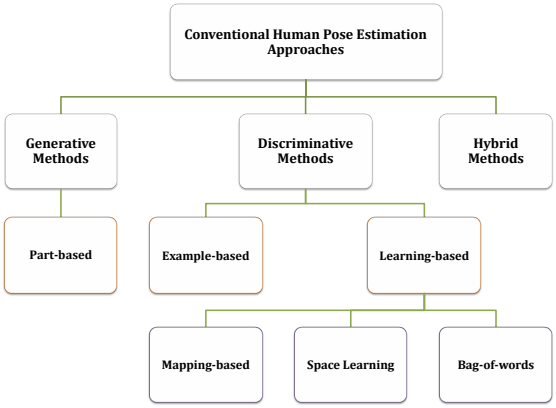

Fig. 4. Our proposed categorization of HPE approaches

A. Conventional HPE Approaches

Previous work on HPE were based essentially on hand-crafted features extracted from raw data. At first, the proposed approaches have often used graphical structure models for recovering the human pose using image-based observations. These models were composed of joints and rigid parts. Ferrari et al. [30] proposed an approach based on a generic detector using a weak model for upper-body pose estimation from TV and movie video shots. Furthermore, they exploited belief propagation technique to refine estimated poses from individual frames and a grab-cut initialized from detected regions to reduce the full pose search space. However, the proposed detector seems to be sensitive to self-occlusions. To resolve this problem, Shotton et al. [31] proposed a new technique to map pose estimation problem into a simpler per-pixel classification problem. It was based on a human body-part segmentation from depth synthetic data. As a classifier, they exploited a randomized decision forest as it reduced over-fitting problem. Later on, Chen et al. [32] presented a graphical model for HPE with image dependent pairwise relations. They used the local image measurements not only to detect joints, but also to predict the spatial relationships between them. This aimed to learn conditional probabilities for the presence of parts and their spatial relationships. It consisted of both unitary appearance terms and binary deformation terms between the parts. Yet, these binary terms were observation dependent, thus they do not rely solely on anthropomorphic priors. Furthermore, relation between the parts could have different types, based on their relative location with respect to each other. Types of part relations were learned with K-Means clustering in the experiments and govern spatial connections between the parts. Besides, a new approach was proposed in [33] using puppets. The solution was to estimate the pose of the body only at one frame and then, use the optical flow technique to check its performance in neighboring frames.
Generally, conventional learning methods seem to be limited. They cannot reach an optimal balance between the discriminability of the extracted features and the robustness of the chosen classifier [34]. The reason is that the choice of features is highly problem-dependent in real-world scenarios. Therefore, stronger approaches are needed to characterize and transform this type of data automatically into discriminate feature vectors.

B. Deep HPE approaches

In this section, we focus on HPE methods based on deep learning approaches. In the last few years, there is a considerable amount of work on deep HPE not only on data representation, but also on human model design. The most widespread deep learning approach is CNN/ConvNets. Indeed, CNN obviate the need for the extensive pre-processing steps, which makes the algorithms faster and more computationally efficient. Moreover, they provide several layers of feature extractors that make it easier to learn implicitly the patterns of each joint. Therefore, this new type of technologies does not need hand-crafted feature anymore. Several methods have been proposed for challenging tasks such as object detection and classification [35], pedestrian detection [36], segmentation [37], hand pose estimation [38], gesture detection and localization [39], video action classification [40], [41] and more. Besides classification, ConvNets have been also trained for regression tasks as HPE [42,43], facial landmark detection [44] and depth prediction [45]. For HPE problem, Gkioxari [46] used a CNN architecture for both pose estimation and action detection. It was a good example of task invariance of the features. In order to determine various human attributes, a collection of CNNs were trained in [47], where each one learned a poselet [48] from a set of image patches. The image was then represented by a collection of part-based deep representations which were then concatenated to obtain a full representation. The architecture consisted of four stages of convolution - normalization - pooling layers, one fully connected layer and finally a logistic regression layer utilized as a classifier. Poselets were commonly used in conjunction with deep CNN for people detection and pose estimation as well. Toshev et al. [49] followed a more direct approach and employed a cascade of DNN regressors to handle the pose estimation task. Later on, in [50], the authors introduced a new top-down procedure called Iterative Error Feedback, which allowed error predictions to be fed back in the CNN and progressively change the initial solution. This new technique made the model self-corrected and more expressive in terms of features. A recent study [51] proposed to apply the convolutions and pooling steps in a way to allow the image being processed repeatedly in a bottom-up and top-down manner with intermediate supervision. Just recently, Belagiannis and Zisserman [52] contributed by introducing a CNN for pose estimation that combined feed forward and recurrent modules able to suppress false detection progressively. In [50], the authors proposed to integrate a consensus voting scheme within a CNN, where votes gathered from every location per keypoint were aggregated to obtain a probability distribution for each keypoint location. Another CNN was trained to infer 3D human pose from uncertainty maps of 2D joint estimates

[51]. To estimate human pose in videos, [52] exploited the ability of ConvNets to benefit from temporal context which was established by combining information between successive time frames using optical flow. The output of this network was a conditional probability distribution to find the best pose configuration explaining the image evidence.
All the research work reviewed in this section and the advances in GPU hardware are pointing to the fact that deep learning is becoming the industrial and academic standard for almost all the computer vision tasks. But one problem is standing as a drawback: if there is not a very large dataset available for training, these deep networks are cursed with the problem of over-fitting, considering the very large number of parameters. Another point to note is that the choice of the number as well as the positions of the key joints play an important role in pose estimation task. Different works, which have been discussed previously, have exploited different numbers/positions of joints. Some of them used just the upper-body joints [52,53]. Others predicted full-body joints [39,42]. This choice depends essentially on the fixed task.

VI. Common Datasets

An ideal HAR/HPE dataset should take into account several important issues:

- Sufficient amount of input data which can include still images and/or videos
- Input media quality (resolution, grayscale, RGB, depth...)
- Large number of subjects performing an action/activity
- Different type of classes of action or poses
- Illumination variations and occlusions
- Complex backgrounds

However, most of existing datasets are recorded in controlled environments performing a set of specific actions. Those limitations are making HAR/HPE algorithms unable to deal with realistic scenarios and therefore do not cover real-world situations.

A. HPE Datasets

Due to the large variations in different scenes, it is difficult to build a universal dataset to evaluate the HPE. Therefore, researchers have created their own datasets to evaluate their proposed techniques for the specific task, which makes the fair comparison on the different algorithms even harder. We summarize the most used publicly available HPE datasets in Tab.1.

B. HAR Datasets

In Tab.2, we list the most relevant datasets according to action/activity recognition. For each dataset, we specify:

- Year of creation
- Problems for which the dataset was defined: Action Classification (AC), Temporal Localization (TL), Spatio-Temporal Localization (STL)
- Involved body parts: U for Upper body, L for Lower body, F for Full body, and H for Hands
- Data modalities available: Depth (D), Skeleton (S), Audio (A), gray scale Intensity (I) and InfraRed (IR)
- Number of classes
- Performance: Acc (Accuracy) and mean Average Precision (mAP)

For more details of different datasets, Chaquet et al. [54] proposed a survey of most important video databases for human action and activity recognition.

VII. Conclusion

In this work, we presented a comprehensive overview of pose-based human action/activity recognition approaches. We defined a general taxonomy covering most of basic and crucial information about human activity analysis and then we reviewed state-of-the-art as welle as recent methods. In this review, we divided HAR into three levels including core technology, HAR systems and HAR applications. We have summarized the classic and representative approaches to human-object segmentation, activity representation and classification. For representation approaches, we roughly sorted out the research trajectory from global to local representations. As the next step, we categorized classification approaches into template-based methods, discriminative models and generative models. Different types of method from the classic DTW to the newest deep learning were summarized. Finally, the most known HAR and HPE datasets were introduced, ranging from classic datasets to recent benchmarks.
Though recent HAR works have achieved great success up to now, applying current approaches in real-world systems or applications still remains nontrivial. In particular, we may conclude that despite the tremendous increase of human activity recognition methods, many problems remain open including modeling of human poses, handling occlusions and annotating data. Moreover, training deep networks on videos still tough. Thus, benefiting from deep models pre-trained on images or other sources would be a better solution to explore since image models have done a good job on capturing spatial relationships of objects, which could also be exploited in action understanding. It is interesting to explore how to transfer knowledge from image models to video models using the idea of inflation or domain adaptation [79,80].

References

[1] Chen, Cheng, et al. "3D human pose recovery from image by efficient visual feature selection." Computer Vision and Image Understanding 115.3 (2011): 290-299.
[2] Qiang, L. I. U., et al. "Hybrid human detection and recognition in surveillance." Neurocomputing 194 (2016): 10-23.
[3] Song, Yale, David Demirdjian, and Randall Davis. "Continuous body and hand gesture recognition for natural human-computer interaction." ACM Transactions on Interactive Intelligent Systems (TiiS) 2.1 (2012): 5.
[4] Unzueta, Luis, et al. "dependent 3D human body posing for sports legacy recovery from images and video." 2014 22nd European Signal Processing Conference (EUSIPCO). IEEE, 2014.
[5] Ke, Shian-Ru, et al. "A review on video-based human activity recognition." Computers 2.2 (2013): 88-131.
[6] Yang, Linjie, et al. "Efficient video object segmentation via network modulation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

TABLE I
PUBLICLY AVAILABLE 2D AND 3D HPE DATABASES

Dataset	Number of frames	Resolution	Body Parts	Dimension	Link
Buffy	Training: 472 Testibg: 276	720x405	Upper body	2D	http://www.robots.ox.ac.uk/ vgg/data/stickmen
Parse	Training: 100 Testing: 205	Different sizes	Full body	2D	http://www.ics.uci.edu/ dramanan/papers/parse
LSP	Training: 1000 Testing: 1000	Different sizes	Full body	2D	http://vision.grasp.upenn.edu/cgi-bin/index.php
FLIC	Training: 3987 Testing: 1016	720x480	Full body	2D	http://vision.grasp.upenn.edu/video/FLIC.zip
PASCAL VOC	47186 images 110008 objects	Different sizes	Upper body	2D	http://pascallin.ecs.soton.ac.uk/challenges/VOC
MPII Human pose	410 activities 25K images	Different sizes	Full body	2D	http://human-pose.mpi-inf.mpg.de
Poses in the wild	30 sequences 900 frames	Different sizes	Upper body	2D	https://lear.inrialpes.fr/research/posesinthewild/dataset
LIP	Single Person: Training: 30462 Testing: 10000 Multi Person: Training: 28280 Testing: 5000, Validation: 5000	Different sizes	Full body	2D	http://www.sysu-hcp.net/lip/overview.php
HumanEva	Training: 50600 Testing: 26400	659x494(color) 644x448(gray)	Full body	3D	http://vision.cs.brown.edu/humaneva
PDT	40 sequences 6 performers	Different sizes	Full body	3D	http://gvvperfcapeva.mpi-inf.mpg.de/public/PersonalizedDepthTracker/index.php
SMMC-10	28 sequences	176x144	Full body	3D	http://ai.stanford.edu/ varung/
EVAL	3 subjects 24 sequences	Different sizes	Full body	3D	http://ai.stanford.edu/ varung/eccv12/

[7] Zhang, Song-Hai, et al. "Pose2Seg: Detection Free Human Instance Segmentation." arXiv preprint arXiv:1803.10683 (2018).

[8] Maninis, Kevis-Kokitsi, et al. "Deep extreme cut: From extreme points to object segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[9] Cai, Shulian, et al. "Menet: a metric expression network for salient object segmentation." arXiv preprint arXiv:1805.05638 (2018).

[10] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2018): 834-848.

[11] Dollr, Piotr, et al. "Behavior recognition via sparse spatio-temporal features." 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. IEEE, 2005.

[12] Wang, Xijuan. "A Human Body Gait Recognition System Based on Fourier Transform and Quartile Difference Extraction." International Journal of Online Engineering (iJOE) 13.07 (2017): 129-139.

[13] Xia, Lu, Chia-Chih Chen, and Jake K. Aggarwal. "View invariant human action recognition using histograms of 3d joints." 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012.

[14] Yang, Xiaodong, and Ying Li Tian. "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor." 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012.

[15] Li, Wanqing, Zhengyou Zhang, and Zicheng Liu. "Action recognition based on a bag of 3d points." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 2010.

[16] Jalal, Ahmad, Shaharyar Kamal, and Daijin Kim. "A Depth Video-based Human Detection and Activity Recognition using Multi-features and Embedded Hidden Markov Models for Health Care Monitoring Systems." International Journal of Interactive Multimedia and Artificial Intelligence 4.4 (2017).

[17] Seddik, Bassem, Sami Gazzah, and Najoua Essoukri Ben Amara. "Human-action recognition using a multi-layered fusion scheme of Kinect modalities." IET Computer Vision 11.7 (2017): 530-540.

[18] Nakazawa, Atsushi, Hirokazu Kato, and Seiji Inokuchi. "Human tracking using distributed vision systems." Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170). Vol. 1. IEEE, 1998.

[19] Iwasawa, Shoichiro, et al. "Real-time estimation of human body posture from monocular thermal images." Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 1997.

[20] Freifeld, Oren, et al. "Contour people: A parameterized model of 2D articulated human shape." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010.

[21] Chan, C. K., W. P. Loh, and I. Abd Rahim. "Human motion classification using 2D stick-model matching regression coefficients." Applied Mathematics and Computation 283 (2016): 70-89.

[22] Lillo, Ivan, Juan Carlos Niebles, and Alvaro Soto. "Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos." Image and Vision Computing 59 (2017): 63-75.

[23] Rahmani, Hossein, Ajmal Mian, and Mubarak Shah. "Learning a deep model for human action recognition from novel viewpoints." IEEE transactions on pattern analysis and machine intelligence 40.3 (2018): 667-681.

[24] Rabiner, Lawrence R., and Biing-Hwang Juang. "An introduction to hidden Markov models." ieee assp magazine 3.1 (1986): 4-16.

[25] Murphy, Kevin Patrick, and Stuart Russell. "Dynamic bayesian networks: representation, inference and learning." (2002).

[26] Vapnik, Vladimir, Steven E. Golowich, and Alex J. Smola. "Support

TABLE II
A LIST OF POPULAR HUMAN ACTIVITY/ACTION DATABASES

Year	Dataset	Problem	Body Parts	Modality	Nb Classes	Performance (%)
2004	KTH	AC	F	I	6	Acc= 98.67 [55]
2006	XMAS	AC	F	RGB, A	13	Acc= 98.79 [56]
2007	HDM05	AC	F	S	100	Acc= 98.17 [57]
2008	HOHA (Hollywood 1)	AC, TL	F, U, L	RGB	8	Acc= 71.90 [58]
2008	UCF Sports	AC, STL	F	RGB	10	Acc= 95.80 [59]
2009	Hollywood 2	AC	F, U, L	RGB	12	mAP= 78.50 [60]
2009	UCF 11 (YouTube Action)	AC, STL	F	RGB	11	Acc= 93.77 [61]
2010	Highfive	AC, STL	F, U	RGB	4	mAP= 69.40 [62]
2010	MSR Action 3D	AC	F	D, S	20	Acc= 97.30 [63]
2010	MSR Action II	STL	F	RGB	3	mAP= 85.0 [64]
2010	Olympic Sports	AC	F	RGB	16	Acc= 96.60 [65]
2011	Collective Activity (extended)	AC	F	RGB	6	Acc= 90.23 [66]
2011	HMDB51	AC	F, U, L	RGB	51	Acc= 73.60 [67]
2012	MPII Cooking	AC, TL	F, U	RGB	65	mAP= 72.40 [68]
2012	MSRDailyActivity3D	AC	F, U	RGB,D,S	16	Acc= 97.50 [69]
2012	UCF 101	AC, TL	F, U, L	RGB	101	Acc= 94.20 [70]
2012	UCF 50	AC	F, U, L	RGB	50	Acc= 97.90 [71]
2012	UTKinect-Action3D	AC	F	RGB,D,S	10	Acc= 98.80 [72]
2013	J-HMDB	AC, STL	F, U, L	RGB,S	21	Acc= 71.08 [73]
2013	Berkely MHAD	AC	F	RGB,D,S,A	11	Acc= 100 [57]
2014	N-UCLA Multiview Action3D	AC	F	RGB,D,S	10	Acc= 90.80 [72]
2014	Sports 1-Million	AC	F, U, L	RGB	487	Acc= 73.10 [74]
2014	THUMOS-14	AC, TL	F, U, L	RGB	101, 20	mAP= 71.60 [75]
2015	THUMOS-15	AC, TL	F, U, L	RGB	101, 20	mAP= 80.80 [65]
2015	ActivityNet	AC, TL	F, U, L	RGB	200	mAP= 93.23 [76]
2016	NTU RGB+D	AC	F	RGB,D,S,IR	60	Acc= 77.70 [77]
2018	Kinetics	AC	F	RGB	600	Acc= 81.50 [78]

vector method for function approximation, regression estimation and signal processing." Advances in neural information processing systems. 1997.

[27] Tipping, Michael E. "The relevance vector machine." Advances in neural information processing systems. 2000.

[28] Min, Seonwoo, Byunghan Lee, and Sungroh Yoon. "Deep learning in bioinformatics." Briefings in bioinformatics 18.5 (2017): 851-869.

[29] Ferrari, Vittorio, Manuel Marin-Jimenez, and Andrew Zisserman. "Progressive search space reduction for human pose estimation." 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008.

[30] Shotton, Jamie, et al. "Real-time human pose recognition in parts from single depth images." (2011): 1297-1304.

[31] Chen, Xianjie, and Alan L. Yuille. "Articulated pose estimation by a graphical model with image dependent pairwise relations." Advances in neural information processing systems. 2014.

[32] Zuffi, Silvia, et al. "Estimating human pose with flowing puppets." proceedings of the IEEE International Conference on Computer Vision. 2013.

[33] Chen, Quan-Qi, et al. "Saliency-context two-stream convnets for action recognition." 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016.

[34] Brunetti, Antonio, et al. "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey." Neurocomputing 300 (2018): 17-33.

[35] Raza, Mudassar, et al. "Appearance based pedestrians head pose and body orientation estimation using deep learning." Neurocomputing 272 (2018): 647-659.

[36] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[37] Elhayek, Ahmed, et al. "Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[38] Neverova, Natalia, et al. "Multi-scale deep learning for gesture detection and localization." Workshop at the European conference on computer vision. Springer, Cham, 2014.

[39] Gkioxari, Georgia, et al. "R-cnns for pose estimation and action detection." arXiv preprint arXiv:1406.5212 (2014).

[40] Fan, Xiaochuan, et al. "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[41] Pfister, Tomas, et al. "Deep convolutional neural networks for efficient

pose estimation in gesture videos." Asian Conference on Computer Vision. Springer, Cham, 2014.
[42] Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
[43] Sun, Xudong, Pengcheng Wu, and Steven CH Hoi. "Face detection using deep learning: An improved faster rcnn approach." Neurocomputing 299 (2018): 42-50.
[44] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
[45] Zhang, Ning, et al. "Panda: Pose aligned networks for deep attribute modeling." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
[46] Pishchulin, Leonid, et al. "Poselet conditioned pictorial structures." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.
[47] Carreira, Joao, et al. "Human pose estimation with iterative error feedback." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[48] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European Conference on Computer Vision. Springer, Cham, 2016.
[49] Belagiannis, Vasileios, and Andrew Zisserman. "Recurrent human pose estimation." 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017). IEEE, 2017.
[50] Lifshitz, Ita, Ethan Fetaya, and Shimon Ullman. "Human pose estimation using deep consensus voting." European Conference on Computer Vision. Springer, Cham, 2016.
[51] Zhou, Xiaowei, et al. "Sparseness meets deepness: 3D human pose estimation from monocular video." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[52] Pfister, Tomas, James Charles, and Andrew Zisserman. "Flowing convnets for human pose estimation in videos." Proceedings of the IEEE International Conference on Computer Vision. 2015.
[53] Hu, Peiyun, and Deva Ramanan. "Bottom-up and top-down reasoning with hierarchical rectified gaussians." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
[54] Chaquet, Jose M., Enrique J. Carmona, and Antonio Fernndez-Caballero. "A survey of video datasets for human action and activity recognition." Computer Vision and Image Understanding 117.6 (2013): 633-659.
[55] Zhou, Tongchi, et al. "Learning semantic context feature-tree for action recognition via nearest neighbor fusion." Neurocomputing 201 (2016): 1-11.
[56] Turaga, Pavan, Ashok Veeraraghavan, and Rama Chellappa. "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision." 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008.
[57] Chaudhry, Rizwan, et al. "Bio-inspired dynamic 3d discriminative skeletal features for human action recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2013.
[58] Saha, Suman, et al. "Deep learning for detecting multiple space-time action tubes in videos." arXiv preprint arXiv:1608.01529 (2016).
[59] Shao, Ling, Li Liu, and Mengyang Yu. "Kernelized multiview projection for robust action recognition." International Journal of Computer Vision 118.2 (2016): 115-129.
[60] Liu, An-An, et al. "Hierarchical clustering multi-task learning for joint human action grouping and recognition." IEEE transactions on pattern analysis and machine intelligence 39.1 (2017): 102-114.
[61] Peng, Xiaojiang, et al. "Action recognition with stacked fisher vectors." European Conference on Computer Vision. Springer, Cham, 2014.
[62] Wang, Heng, et al. "A robust and efficient video representation for action recognition." International Journal of Computer Vision 119.3 (2016): 219-238.
[63] Luo, Jiajia, Wei Wang, and Hairong Qi. "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps." Proceedings of the IEEE international conference on computer vision. 2013.
[64] Chen, Wei, and Jason J. Corso. "Action detection by implicit intentional motion clustering." Proceedings of the IEEE international conference on computer vision. 2015.
[65] Li, Yingwei, et al. "Vlad3: Encoding dynamics of deep features for action recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[66] Amer, Mohamed R., et al. "Monte carlo tree search for scheduling activity recognition." Proceedings of the IEEE International Conference on Computer Vision. 2013.
[67] Wang, Hongsong, Wei Wang, and Liang Wang. "How scenes imply actions in realistic videos?." 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016.
[68] Zhou, Yang, et al. "Interaction part mining: A mid-level approach for fine-grained action recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[69] Shahroudy, Amir, et al. "Deep multimodal feature analysis for action recognition in rgb+ d videos." IEEE transactions on pattern analysis and machine intelligence 40.5 (2018): 1045-1058.
[70] Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." European conference on computer vision. Springer, Cham, 2016.
[71] Duta, Ionut C., et al. "Spatio-temporal vlad encoding for human action recognition in videos." International Conference on Multimedia Modeling. Springer, Cham, 2017.
[72] Kerola, Tommi, Nakamasa Inoue, and Koichi Shinoda. "Cross-view human action recognition from depth maps using spectral graph sequences." Computer Vision and Image Understanding 154 (2017): 108-126.
[73] Peng, Xiaojiang, and Cordelia Schmid. "Multi-region two-stream R-CNN for action detection." European conference on computer vision. Springer, Cham, 2016.
[74] Yue-Hei Ng, Joe, et al. "Beyond short snippets: Deep networks for video classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[75] Jain, Mihir, Jan C. Van Gemert, and Cees GM Snoek. "What do 15,000 object categories tell us about classifying and localizing actions?." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[76] Montes, Alberto, et al. "Temporal activity detection in untrimmed videos with recurrent neural networks." arXiv preprint arXiv:1608.08128 (2016).
[77] Liu, Jun, et al. "Spatio-temporal lstm with trust gates for 3d human action recognition." European Conference on Computer Vision. Springer, Cham, 2016.
[78] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[79] Tsai, Yi-Hsuan, et al. "Learning to adapt structured output space for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[80] Zhao, Shanshan, et al. "Geometry-Aware Symmetric Domain Adaptation for Monocular Depth Estimation." arXiv preprint arXiv:1904.01870 (2019).