eISSN: 2477-7994

Journal of Measurements, Electronics, Communications, and Systems (Vol.9, No. 1, Pg.19-29, June 2022)

Human Action Recognition Using Deep Learning and Nonparametric Model With Some Exchanges in Karl Popper's Viewpoint and Kuhn's Paradigm: A Literature Review From Perspective of Philosophy of Science

Ig. Prasetya Dwi Wibawa*[1], Meta Kallista[2], Ganga Ram Phaijoo[3]

[1]*Electronics Engineering, Telkom University*
[2]*Computer Engineering, Telkom University*
[3]*Department of Mathematics, School of Science, Kathmandu University*
**prasdwibawa@telkomuniversity.ac.id*

Manuscript received January 19, 2022; revised March 10, 2022; accepted May 2, 2022

Abstract

Human skeletal detection and human gesture recognition are interesting subjects that have been investigated during the past three decades. Single-RGB, RGB-D camera, and Initial Measurement Unit (IMU) are some of the sensors for recording human motion data. Numerous methods for gesture recognition and classification have been reviewed in this survey. The classification is divided into nonparametric models and deep learning models, which afterwards will be compared in terms of accuracy and running time, respectively. The feature extractions are separated based on features processed from the sensor data, including skeleton-based features, depth image-based features, and hybrid features. A comparison of accuracy values will be offered based on the model and its attributes. In addition, we present an interchange of perspectives on deep learning and nonparametric models based on Karl Popper's perspective and Kuhn's paradigm in the study of the philosophy of science. By substituting the falsification principle for induction, Popper attempts to refute the traditional empiricist perspective of the scientific method. From the philosophy of science's perspective, the study on human action recognition is in the normal science phase according to Kuhn's paradigm and is corroborated in accordance with Popper's theory.

Keywords: human action recognition; nonparametric model; deep learning model; Karl Popper; Kuhn's Paradigm; philosophy of science
DOI: 10.25124/jmecs.v9i1.2408

1. Introduction

Throughout the history of human understanding, philosophy and science have always related to one another. Philosophy and science are intertwined in their pursuit of truth fragments. The aim of science is to describe, whereas the task of philosophy is to interpret the phenomena of the universe or the truth in mind, whereas the truth of science is derived from experiences and observations. Before doing a survey of human gesture recognition using deep learning and nonparametric models, its ontology, epistemology, and axiology must be understood. The etymology of ontology comes from the Greek language. Ontology derives from the Greek terms "ontos," which means "being," and "logos," which means "science, teachings, or beliefs." Ontology, in terms of terminology, is the branch of science that explores the true nature and essence of things. Epistemology is the branch of philosophy that examines in depth how to get accurate information. Axiology is a branch of research that explores the philosophical nature of values. [1]

Ontology classifications can be defined by their textual definitions, a set of properties, and a logical definition composed of several formulas [2]. We first should establish the ontology for human action recognition in semantic space. According to Ziaeefard [3], human action recognition can be differentiated using semantic space characteristics. As depicted in Fig. 1, the semantic space is separated into body parts (pose and poselets), qualities, linked objects, human-object interactions, and the context of the location. Using a similar method to the human ability to distinguish

The online version of this article can be found at: https://journals.telkomuniversity.ac.id/jmecs/article/view/2408

in semantic space, a machine can learn to recognize human activities from a given image sequences.

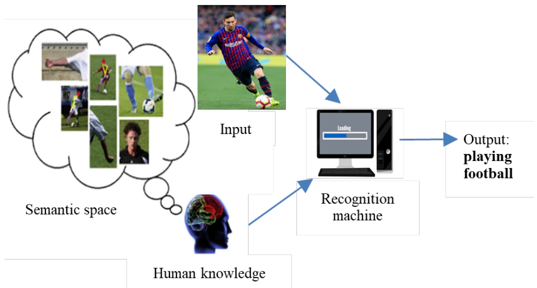


Fig. 1. Semantic space: playing football recognition [3]. Pose (certain parts of body pose of football player), poselet (right hand lifted, left hand straight down), attributes (head looking down), related objects (ball, interaction left foot with ball), context of place (ball field).

In a nutshell, epistemology is the study of how to obtain knowledge. There are some methods to obtain knowledge, i.e., literature review, survey, and interview [4]. For example, in a survey, we can use the Likert scale as a survey instrument and use certain scales in questionnaires. To obtain a good survey, we use reliability and validity tests to measure the acceptance indicator [5]. The knowledge that we want to obtain as well as our study goal, i.e., how to implement a human action recognition algorithm for machine learning by using a vision sensor, We use some parameters as a performance index to compare between nonparametric models and deep learning models, such as accuracy and running time. To get accuracy, we use a confusion matrix as shown in Table 1.

Table 1: Confusion Matrix

		Prediction	
		Class 1	Class 2
Observation	Class 1	TP	FN
	Class 2	FP	TN

where True positive (TP) is positive observations and positive predicting results, False negative (FN) is positive observations and negative predicting results, True negative (TN) is negative observation results and negative predicting results, and False positive (FP) is negative observations and positive predicting results. The accuracy can be calculated as shown in Eq. (1).

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

Running time depends on the specifications of the hardware used (GPU/processor), the algorithm in the model is built, and the framework of model used. There is a trade-off between accuracy and the amount of time it takes to compute. For example, if you want more accuracy, it will take longer to compute, but if you want less accuracy, it will take less time. Vision sensors, such as the RGB-D sensor, are used for the recognition of human movements, facial recognition, the introduction of human interactions [3], and 3D reconstruction. The RGB-D sensors for example are Kinect, Asus Xtion, and Intel RealSense. These sensors have the capability to capture imagery and are subsequently processed for detection and recognition purposes. There are open-source benchmark datasets for human pose estimation using RGB-D sensor to carry out performance tests of learning models such as MSR Action3D[1] [6], UTKinect-Action3D[2] [7], MSRDaily-Activity3D[3] [8], UTD-MHAD[4] [9], SBU Kinect Interaction[5] [10], NTU RGB+D[6] [11], and PKU-MMD[7] [12].

Axiology is about the value of research that can be used to solve real problems in our society. The implementation of human action recognition has a wide range of applications, such as surveillance cameras (or video surveillance), elderly care, virtual reality, and human-machine interactions [3]. Our approach in terms of axiology is to build a learning and control system for human motion recognition in general. Examples of human motion are Indonesian traditional dancing and traditional martial arts such as pencak silat.

Online recognition, occlusion, variations in camera capture angle, computational time, and biometric changes present a major difficulty in human action recognition. Online recognition is the ability to recognize changes and classify movements instantaneously (in limited time intervals) of video sequences continuously. Occlusion, where an affected part of the body also causes the detection process to become more difficult [13]. Variations in camera capture angles and biometric changes, such as variations in body size, appearance, shape, and sensor-to-subject distance, will impact the algorithm's performance. Time computation is also a factor in influencing an algorithm's performance.

A human action recognition model will be divided into two models, i.e., a nonparametric model [14] and a deep learning model. In nonparametric models, a mathematical model is used to classify a set of statistical data where the data of variables tested in the hypothesis model did not follow a certain probability distribution. Furthermore, the feature extraction results of the next feature are processed and modeled mathematically with certain classification methods in order to obtain the desired human action recognition. While on a deep learning model, feature extraction can be built automatically from deep learning architecture design learning to be subsequently used in motion recognition processes. Examples of classifier methods on nonparametric models are random forest (RF), k-Nearest Neighbor (kNN), Support Vector Machine

[1] research.microsoft.com/en-us/um/people/zliu/actionrecorsrc
[2] cvrc.ece.utexas.edu/KinectDatasets/HOJ3D.html
[3] research.microsoft.com/en-us/um/people/zliu/actionrecorsrc
[4] personal.utdallas.edu/~kehtar/UTD-MHAD.html
[5] github.com/xrenaa/SBU_Kinect_dataset_process
[6] rose1.ntu.edu.sg/datasets/actionrecognition.asp
[7] www.icst.pku.edu.cn/struct/Projects/PKUMMD.html

(SVM), Extreme Learning Machine (ELM), Hidden Markov model (HMM), graph, and template matching. Meanwhile, the examples for the classifier method on deep learning models are the convolution neural network (CNN), the recurrent neural network (RNN), and CNN + LSTM (long short-term memory).

Karl Popper established one of the most popular falsification methods in the philosophy of science. The viewpoint of Karl Popper is a useful beginning point for falsifying suggested theories or hypotheses. Popper produced a comprehensive critique of historicism, holism, and their associated ideas [15]. For the research to be corroborated, every observation, experiment, and method employed must be falsified by others (methods, experiments, or observations). If the proposed theory or hypothesis can withstand a process of falsification, then the theory or observation is supported or strengthened. The idea or hypothesis is provisionally accepted so long as no other theory or scientific observation refutes it [16]. To find a novelty in every field of study, one might begin by examining historical science and its paradigm. As demonstrated in Fig. 2, Kuhn divides the structure of scientific revolutions into four paradigms: pre-science, normal science, anomaly and the development of scientific discoveries, and crisis and the emergence of scientific ideas [17].

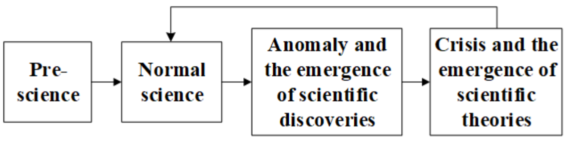


Fig. 2. The structure of scientific revolutions by Thomas Kuhn [17].

2. Human Action Recognition Features

The extraction of features from the sensor readings of RGB or RGB-D cameras can be categorized as skeleton-based, depth image-based, or hybrid [18]. Table 2 represents the accuracy of human action recognition, while Table 3 depicts the survey-related processing or computation time for skeletal detection. The study of human action recognition is in the normal science phase according to Kuhn's paradigm, including the scientific practice of reasoning, observing, and experimenting within a well-established paradigm or explanatory framework. Recent research in human action recognition employs skeletal estimation, depth-image estimation, and hybrid features as shown in Table 2.

2.1. Depth Image-based Features

The features of the depth image can be extracted using the depth motion map (DMM). For depth sequences with a number of N-frames, DMM can be obtained through Eq. (2) as follows.

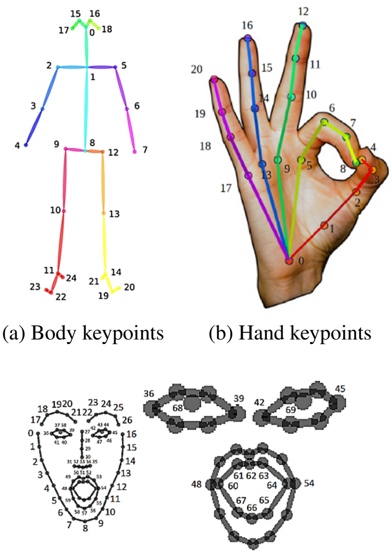
$$DMM_{\{f,s,t\}} = \sum_{i=1}^{N-1} \left| \mathrm{map}^{i+1}_{\{f,s,t\}} - \mathrm{map}^{i}_{\{f,s,t\}} \right| \qquad (2)$$

Fig. 3. Openpose for skeleton readings in single subjects and two subjects, for example, pencak silat motions.

where i represents the frame index, f, s, and t represent orthogonal projection 2D-mapping to the front, side, and top sides, respectively. From the DMM computing results, the next step is to implement human action recognition using histograms of oriented gradients (HoG) [19]. Additionally, the approach of principal component analysis (PCA) can be used to minimize the dimensionality of these features.

2.2. Skeleton-based Features

The skeleton-based method employs CNN or RNN based on the adopted deep learning structure to determine the coordinate position of the joint skeleton using a single-RGB sensor. VNect (Mehta et al. [20]) and OpenPose are open source pretrained models for human pose estimation (Ze Chao et al. [21]). Fig. 3 illustrates a joint skeleton reading application utilizing OpenPose, with a single subject and many subjects. OpenPose has the ability to read Part Affinity Field (PAF) skeletons and a number of human objects. OpenPose divides the body's posture into 25 joints, and it can be used for face and hand readings, as one can see in Fig. 4.



(a) Body keypoints (b) Hand keypoints

(c) Face keypoints

Fig. 4. OpenPose for reading the joint skeleton position for the body (top left), hand (top right), and face (bottom) [21]

Table 2: Survey on Human Action Recognition Literature Using the RGB-D Sensor With Corresponding Datasets

Dataset: MSR Action3D

Features	Nonparametric Model				Deep Learning Model			
	Ref.	Year	Classifier	Acc. (%)	Ref.	Year	Classifier	Acc. (%)
	Wang et al. [22]	2014	Actionlet Ensemble	86	Veeriah et al. [23]	2015	RNN	92.03
	Chaaraoui et al. [24]	2014	SVM	92.46	Du et al. [25]	2015	RNN	94.49
	Theodorakopoulos et al. [26]	2014	kNN	93.61	Núñez et al. [27]	2018	CNN+LSTM	96
	Koniusz et al. [28]	2016	SVM	93.96	Lee et al. [29]	2017	LSTM	97.22
Skeleton-based	Liu et al. [30]	2016	Matching	94.4				
	Guo et al. [31]	2018	SVM	95.24				
	Liu et al. [32]	2018	SVM	95.6				
	Qiao et al. [33]	2017	SVM	95.9				
	Chen et al. [34]	2016	Graph	96.1				
	Jia et al. [35]	2013	SVM	89.3	Wang et al. [36]	2015	CNN	94.58
	Devanne et al. [37]	2015	kNN	92.1	Wang et al. [38]	2016	CNN	100
Depth image-based	Yang et al. [39]	2014	SVM	93.9				
	Liu et al. [40]	2016	SVM	94.28				
	Chen et al. [41]	2017	ELM	96.7				
	Liu et al. [42]	2018	SVM	97.64				
	Wang et al. [22]	2014	SVM	88.2	Liu et al. [43]	2016	CNN	84.07
	Ji et al. [44]	2018	SVM	90.8	Kamel et al. [45]	2018	CNN	94.51
	Jalal et al. [46]	2017	HMM	93.3	Shi et al. [47]	2017	RNN	94.9
Hybrid features	Kong et al. [48]	2016	SVM	93.99				
	Zhu et al. [8]	2013	RF	94.3				
	Ohn-Bar et al. [49]	2013	SVM	94.84				
	Shahroudy et al. [50]	2016	SVM	98.2				

Dataset: UTKinect-Action3D

Features	Nonparametric Model				Deep Learning Model			
	Ref.	Year	Classifier	Acc. (%)	Ref.	Year	Classifier	Acc. (%)
	Theodorakopoulos et al. [26]	2014	kNN	90.95	Rahmani et al. [51]	2017	LSTM	95.96
	Wang et al. [52]	2016	Matching	93.47	Lee et al. [29]	2017	LSTM	96.67
	Chen et al. [34]	2016	Graph	95.96	Liu et al. [53]	2016	LSTM	97
Skeleton-based	Vemulapalli et al. [54]	2014	SVM	97.08	Núñez et al. [27]	2018	CNN+LSTM	99
	Guo et al. [31]	2018	SVM	97.85	Liu et al. [55]	2018	LSTM	99
	Koniusz et al. [28]	2016	SVM	98.2				
	Liu et al. [42]	2018	SVM	86	Liu et al. [43]	2016	CNN	82

Depth image-based	Slama et al. [56]	2014	PDF	95.25	Wang et al. [38]	2016	CNN	90.91
					Wang et al. [36]	2015	CNN	91.92
Hybrid features	Raman et al. [57]	2016	HMM	87.9	Liu et al. [43]	2016	CNN	96
	Zhu et al. [8]	2013	RF	91.9				
	Liu et al. [58]	2015	HC-RF	92				
	Zhang et al. [59]	2016	SVM	94.9				
Dataset: MSRDailyActivity3D								
Features	Nonparametric Model				Deep Learning Model			
	Ref.	Year	Classifier	Acc. (%)	Ref.	Year	Classifier	Acc. (%)
Skeleton-based	Zanfir et al. [60]	2013	kNN	73.8	Núñez et al. [27]	2018	CNN+LSTM	63.1
	Qiao et al. [33]	2017	SVM	75				
	Cai et al. [61]	2016	MIL	78.52				
	Liu et al. [42]	2018	SVM	91				
Depth image-based	Oreifej et al. [62]	2013	SVM	80	Wang et al. [36]	2015	CNN	78.12
	Yang et al. [39]	2014	SVM	86.25	Wang et al. [38]	2016	CNN	85
	Jia et al. [35]	2016	SVM	80.63	Luo et al. [63]	2017	CNN+LSTM	86.9
	Chen et al. [41]	2017	ELM	89	Shinde et al. [64]	2018	YOLO	88.358
Hybrid features	Kong et al. [48]	2016	SVM	73.21				
	Ji et al. [44]	2018	SVM	81.3				
	Zhang et al. [59]	2016	SVM	86				
	Kong et al. [65]	2016	DRRL	87.5				
	Shahroudy et al. [50]	2016	SVM	91.25				
	Althloothi et al. [66]	2014	SVM	93.1				
	Jalal et al. [46]	2017	HMM	94.1				

Table 3: Time Computation Parameter (Note: mAP (%) is mean Average Precision)

Vision Sensor	Ref.	Year	Methods	Model	mAP (%)	Time Computation	Framework	Hardware	Output
single-RGB camera	Newell et al. [67]	2016	stacked hourglass	Deep learning: CNN	87.4	75 ms	-	Nvidia TITAN X	body joint (single-person)
single-RGB camera	Mehta et al. [20]	2017	VNect	Deep learning: CNN	76.6	CNN 18 ms, skeleton fitting 7–10 ms, pre-processing and filtering 5 ms (total 33 ms)	Caffe	6-core Xeon CPU 3.8 GHz, Titan GPU	Body joint (multi-person)
single-RGB camera	Zhe Cao et al. [21]	2017	OpenPose: PAF	Deep learning: multistage-CNN	85.6	22 fps - 36 ms (body + foot)	Cuda 8	Nvidia GTX 1080 Ti	Body, fingers, and face (multi-person)
Kinect v2 + FIR camera	Nishi et al. [68]	2017	VICON	Fully convolutional network (FCN)	87.5	50 fps	-	1 GForce GTX Titan X; 2 Nvidia Titan X	body joint
Kinect v2	Vasileiadis et al. [69]	2019	PAF	3D-CNN	87.3	360 ms (0.36 s per frame or 2.8 fps)	Chainer	NVIDIA GTX 970 GPU	body joint
single-RGB camera	Luvizon et al. [70]	2019	Soft-argmax	Deep learning: CNN	90.8	29.3 fps	Tensorflow	NVIDIA GPU K20	body joint

Table 4: Dimension Complexity of Computing Process

Methods	Computation
BMLD[1]-GMM[2]	$\mathcal{O}(J \times K_h D^2)$
LDA[3] + HMM	$\mathcal{O}(K_h MP + P^3) + \mathcal{O}(N_h H^2)$
PCA[4] + NBNN[5]	$\mathcal{O}(m^3 + m^2 r) +$ $\mathcal{O}(r \times n_c \times n_d + log(n_c + n_d))$
SVM[6]	$\mathcal{O}(r^3)$
PCA + STOP[7]	$\mathcal{O}(m^3 + m^2 r) + \mathcal{O}(n_c \times r)$
PCA + CRC[8]	$\mathcal{O}(m^3 + m^2 r) + \mathcal{O}(n_c \times r)$

Notes:
[1]BMLD: bi-gram maximum likelihood decoding
[2]GMM: Gaussian mixture model
[3]LDA: linier discriminant analysis
[4]PCA: principal component analysis
[5]NBNN: Naive Bayes nearest neighbour
[6]SVM: support vector machine
[7]STOP: space-time occupancy patterns
[8]CRC: collaborative representation based classification

2.3. *Hybrid Features*

Using the fusion principle, one can utilize some features as hybrid features. The collaborative representation classifier (CRC) is one application approach for fusion principles [71]. In addition to the RGB-D sensor for the DMM depth image and skeleton feature, inertial characteristics are also incorporated. Chen et al. [19] combine the DMM, skeleton, and inertia parameters of fusion sensors in online movement recognition. Dimensional complexity using L_2-regularized CRC method is $\mathcal{O}(m^3 + m^2 r) + \mathcal{O}(n_c \times r)$, as shown in Table 4. In detail, the computation time for all steps are (2.0 ± 0.4) ms/frame for projected map generation, (3.3 ± 0.6) ms/frame for DMM process, (2.5 ± 1.2) ms/sequence of motion for PCA process, and (1.8 ± 0.5) ms/sequence of motion for human action recognition process.

3. Classifier Methods for Human Action Recognition

The classifier methods for human movement recognition are generally categorized into nonparametric model models and deep learning models. Some of the classifier methods related to these two models can be seen in Fig. 5. In deep learning models, the CNN-based movement recognition process generally focuses on the position processing or the trajectory of the joint skeleton in an image, which is then processed with CNN for its classification.

Li et al. [72] use a joint distance map (JDM) of one or several joint skeletons converted into color variations to obtain temporal information. Mehta et al. [20] introduce the online method for the 3D skeletal pose by using single-RGB cameras. The 2D pose is taken from a joint position without using the depth image method and converted into a 3D pose with the skeleton fitting process. Pham et al. [73] use a 3D joint

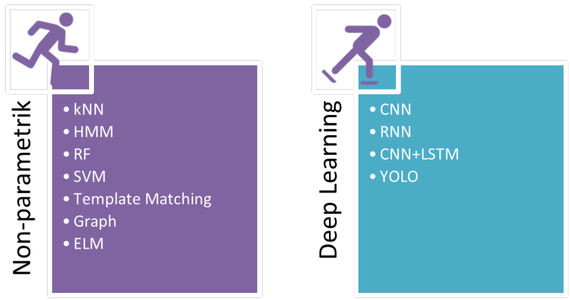


Fig. 5. Some classifier methods for human action recognition.

skeleton coordinate and divide each skeleton into five parts, where each joint is combined in the order of the physical form of the body, which adopts the evolution of 3D spatio-temporal motions. Ze Chao et al. [21] use the CNN + PAF multi-stage architecture to improve the detection performance of multiple subjects with the OpenPose application, as shown in Fig. 6. The network is categorized into two parts: the top predicts confidence maps, while the bottom predicts affinity fields, where **F** denotes feature maps, ρ^t and ϕ^t are the CNNs for inference at Stage t.

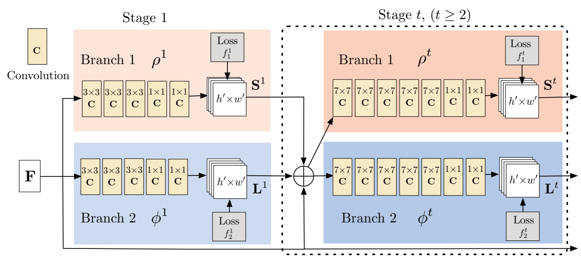
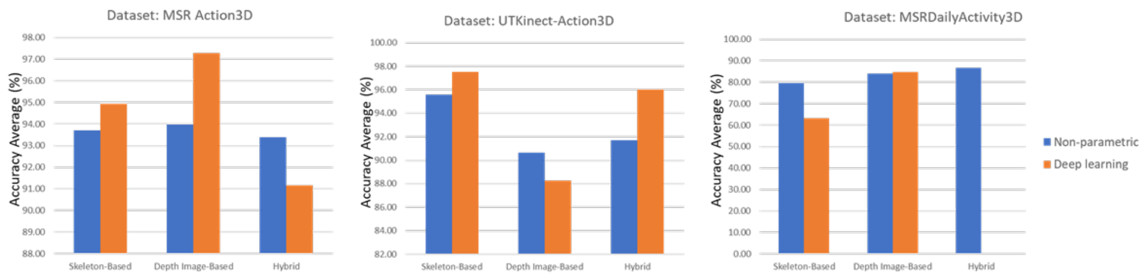


Fig. 6. Two-branch multi-stage CNN+PAF using OpenPose [21]. The first stage is to predict the PAF, $\mathbf{L}^t$, and the second stage is to predict the level of confidence map, $\mathbf{S}^t$.

4. Discussions

We obtain a comparison of accuracy for human action recognition using three different datasets as shown in Fig. 7 from data processing on Table 2. Nonparametric models have an average accuracy 90% while deep learning models have an average accuracy about 89.1%. Although deep learning has become popular recently, nonparametric models still have a better performance index in terms of average accuracy for human action recognition. A Deep learning model has been used as a skeleton detection model, which has better time processing and is suitable for online recognition. For example, the OpenPose algorithm approximately has a computation time of 36 ms (22 fps) using Nvidia GTX 1080 Ti which is considered fast enough for online recognition as shown in Fig. 3.

From Popper's viewpoint, to accommodate the falsification process, we propose a flowchart for the falsifying step with scientific observation or experiment as shown in Fig. 8. We put falsification process as well as



(a) MSR Action3D dataset (b) UTKinect-Action 3D dataset (c) MSRDailyActivity3D dataset

Fig. 7. The accuracy comparison for nonparametric models and deep learning models using benchmark datasets: (a) MSR Action3D dataset, (b) UTKinect-Action 3D dataset, and (c) MSRDailyActivity3D dataset.

evaluating process on the same step. Kuhn's paradigm of the scientific revolution, human action recognition is now in the stage of normal science. The method that is used for classification now is the machine learning method that has been developed in 19th of century with minor modifications in the algorithm, particularly in deep learning models. There is still no emergence of new scientific discoveries and theories yet. Human action recognition using nonparametric and deep learning models deserves additional research into more challenging problems such as occlusion, shading, unusual activities, viewpoint variation, camera motion, background clutter, and execution rate [74].

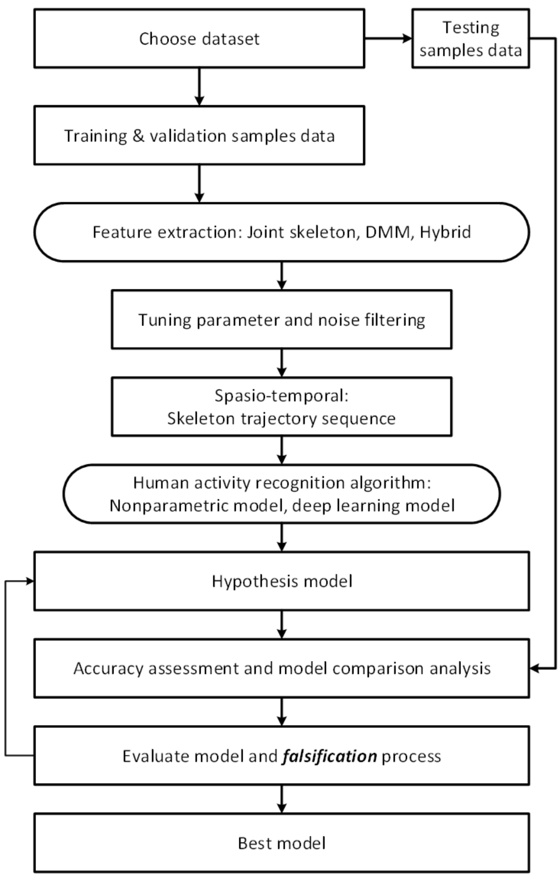


Fig. 8. Searching method for the best learning model of human action recognition with added Popper's falsification.

5. Conclusion

This paper reviews human action recognition, incorporating some philosophy of science approaches. Deep learning and nonparametric approaches have been studied in order to determine the state of the art in human action recognition using different types of features such as depth-image based features, skeleton-based features, and hybrid features. From the philosophy of science's perspective, the study of human action recognition is in the normal science phase according to Kuhn's paradigm, including the scientific practice of reasoning, observing, and experimenting within a well-established paradigm or explanatory framework as shown in the literature study in Table 2. In accordance with Popper's theory, the human action recognition study is corroborated by the usage of a methodology to falsify a method through the performance of evaluation metrics. Although the deep learning is more favorable nowadays, the evaluation performance findings indicate that deep learning and nonparametric methods yield equivalent outcomes.

Acknowledgment

The first author is gratefully acknowledge Dr. Dimitri Mahayana for giving insightful knowledge of philosophy of science in course EL7090.

References

[1] K. Liu, W. Hao, and Y. Qin, "The ontology of virtual geographical environment," in *2010 18th International Conference on Geoinformatics*. IEEE, 2010, pp. 1–6.

[2] C. Roussey, F. Pinet, M. A. Kang, and O. Corcho, "An introduction to ontologies and ontology engineering," in *Ontologies in Urban development projects*. Springer, 2011, pp. 9–38.

[3] M. Ziaeefard and R. Bergevin, "Semantic human activity recognition: A literature review," *Pattern Recognition*, vol. 48, no. 8, pp. 2329–2345, 2015.

[4] J. Zhu, R. Liu, Q. Liu, T. Zheng, and Z. Zhang, "Engineering students' epistemological thinking in the context of project-based learning," *Ieee transactions on education*, vol. 62, no. 3, pp. 188–198, 2019.

[5] S. Bajpai and R. Bajpai, "Goodness of measurement: Reliability and validity," *International Journal of Med-*

ical Science and Public Health, vol. 3, no. 2, pp. 112–115, 2014.

[6] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 9–14.

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

[8] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3D action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 486–491.

[9] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*. IEEE, 2015, pp. 168–172.

[10] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Supplementary material for "two-person interaction detection using body-pose features and multiple instance learning"."

[11] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3D human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[12] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.

[13] Y. Kong and Y. Fu, "Modeling supporting regions for close human interaction recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 29–44.

[14] P. Thanh Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery," *Sensors*, vol. 18, no. 1, p. 18, 2017.

[15] F.-Y. Wang, "From piecemeal engineering to twitter technology: toward computational societies," *IEEE Intelligent Systems*, vol. 27, no. 4, pp. 2–3, 2012.

[16] K. Popper, *The logic of scientific discovery*. Routledge, 2005.

[17] T. S. Kuhn, *The structure of scientific revolutions*. Chicago University of Chicago Press, 1970, vol. 111.

[18] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2712–2716.

[19] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of real-time image processing*, vol. 12, no. 1, pp. 155–163, 2016.

[20] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3D human pose estimation with a single rgb camera," *Acm transactions on graphics (tog)*, vol. 36, no. 4, pp. 1–14, 2017.

[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Real-time multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[22] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 914–927, 2013.

[23] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4041–4049.

[24] A. A. Chaaraoui and F. Florez-Revuelta, "Optimizing human action recognition based on a cooperative co-evolutionary algorithm," *Engineering Applications of Artificial Intelligence*, vol. 31, pp. 116–125, 2014.

[25] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[26] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 12–23, 2014.

[27] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognition*, vol. 76, pp. 80–94, 2018.

[28] P. Koniusz, A. Cherian, and F. Porikli, "Tensor representations via kernel linearization for action recognition from 3D skeletons," in *European conference on computer vision*. Springer, 2016, pp. 37–53.

[29] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1012–1020.

[30] B. Liu, H. Yu, X. Zhou, D. Tang, and H. Liu, "Combining 3D joints moving trend and geometry property for human action recognition," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2016, pp. 000 332–000 337.

[31] Y. Guo, Y. Li, and Z. Shao, "Dsrf: A flexible trajectory descriptor for articulated human action recognition," *Pattern Recognition*, vol. 76, pp. 137–148, 2018.

[32] B. Liu, Z. Ju, and H. Liu, "A structured multi-feature representation for recognizing human action and interaction," *Neurocomputing*, vol. 318, pp. 287–296, 2018.

[33] R. Qiao, L. Liu, C. Shen, and A. van den Hengel, "Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition," *Pattern Recognition*, vol. 66, pp. 202–212, 2017.

[34] H. Chen, G. Wang, J.-H. Xue, and L. He, "A novel hierarchical framework for human action recognition," *Pattern Recognition*, vol. 55, pp. 148–159, 2016.

[35] C. Jia and Y. Fu, "Low-rank tensor subspace learning for rgb-d action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4641–4652, 2016.

[36] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Deep convolutional neural networks for action recognition using depth map sequences," *arXiv preprint arXiv:1501.04686*, 2015.

[37] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on riemannian manifold," *IEEE transactions on cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2014.

[38] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 498–509, 2015.

[39] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 1028–1039, 2016.

[40] M. Liu and H. Liu, "Depth context: a new descriptor for human activity recognition by using sole depth sequences," *Neurocomputing*, vol. 175, pp. 747–758, 2016.

[41] C. Chen, M. Liu, H. Liu, B. Zhang, J. Han, and N. Kehtarnavaz, "Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition," *IEEE Access*, vol. 5, pp. 22 590–22 604, 2017.

[42] M. Liu, H. Liu, and C. Chen, "Robust 3D action recognition through sampling local appearances and global distributions," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1932–1947, 2017.

[43] Z. Liu, C. Zhang, and Y. Tian, "3D-based deep convolutional neural network for action recognition with depth sequences," *Image and Vision Computing*, vol. 55, pp. 93–100, 2016.

[44] X. Ji, J. Cheng, W. Feng, and D. Tao, "Skeleton embedded motion body partition for human action recognition using depth sequences," *Signal Processing*, vol. 143, pp. 56–68, 2018.

[45] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1806–1819, 2018.

[46] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern recognition*, vol. 61, pp. 295–308, 2017.

[47] Z. Shi and T.-K. Kim, "Learning and refining of privileged information-based rnns for action recognition from depth sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3461–3470.

[48] Y. Kong, B. Satarboroujeni, and Y. Fu, "Learning hierarchical 3D kernel descriptors for rgb-d action recognition," *Computer Vision and Image Understanding*, vol. 144, pp. 14–23, 2016.

[49] E. Ohn-Bar and M. Trivedi, "Joint angles similarities and hog2 for action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 465–470.

[50] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2123–2129, 2015.

[51] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1506–1515.

[52] C. Wang, Y. Wang, and A. L. Yuille, "Mining 3D keypose-motifs for action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2639–2647.

[53] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3D human action recognition," in *European conference on computer vision*. Springer, 2016, pp. 816–833.

[54] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[55] 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 Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2017.

[56] R. Slama, H. Wannous, and M. Daoudi, "Grassmannian representation of motion depth for 3D human gesture and action recognition," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 3499–3504.

[57] N. Raman and S. J. Maybank, "Activity recognition using a supervised non-parametric hierarchical hmm," *Neurocomputing*, vol. 199, pp. 163–177, 2016.

[58] A.-A. Liu, W.-Z. Nie, Y.-T. Su, L. Ma, T. Hao, and Z.-X. Yang, "Coupled hidden conditional random fields for rgb-d human action recognition," *Signal Processing*, vol. 112, pp. 74–82, 2015.

[59] H. Zhang and L. E. Parker, "Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 541–555, 2014.

[60] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2752–2759.

[61] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 141–154, 2015.

[62] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 716–723.

[63] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," in *Proceedings of the IEEE conference*

on computer vision and pattern recognition, 2017, pp. 2203–2212.

[64] S. Shinde, A. Kothari, and V. Gupta, "Yolo based human action recognition and localization," *Procedia computer science*, vol. 133, pp. 831–838, 2018.

[65] Y. Kong and Y. Fu, "Discriminative relational representation learning for rgb-d action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2856–2865, 2016.

[66] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles, "Human activity recognition using multi-features and multiple kernel learning," *Pattern recognition*, vol. 47, no. 5, pp. 1800–1812, 2014.

[67] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[68] K. Nishi and J. Miura, "Generation of human depth images with body part labels for complex human pose recognition," *Pattern Recognition*, vol. 71, pp. 402–413, 2017.

[69] M. Vasileiadis, C.-S. Bouganis, and D. Tzovaras, "Multi-person 3D pose estimation from 3D cloud data using 3D convolutional neural networks," *Computer Vision and Image Understanding*, vol. 185, pp. 12–23, 2019.

[70] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," *Computers & Graphics*, vol. 85, pp. 15–22, 2019.

[71] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *2011 International conference on computer vision*. IEEE, 2011, pp. 471–478.

[72] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, 2017.

[73] K. Liu, W. Hao, and Y. Qin, "The ontology of virtual geographical environment," in *2010 18th International Conference on Geoinformatics*. IEEE, 2010, pp. 1–6.

[74] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: research and evaluation challenges," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 5, pp. 650–663, 2014.

Author Information

Ignatius Prasetya Dwi Wibawa earned a Bachelor of Science in Electrical Engineering in 2011 and a Master of Science in control and intelligent system in 2013 from Institut Teknologi Bandung (ITB). Currently, he is pursuing a Doctoral degree at School of Electrical Engineering and Informatics, ITB. Beginning in 2014, he worked as a lecturer at the School of Electrical Engineering at Telkom University. Control systems, machine learning, and artificial intelligence comprise his current area of research.

Meta Kallista earned a Bachelor of Science in Mathematics in 2010 from Brawijaya University, a Master of Science in Mathematics in 2013, and a Doctoral degree in 2019 from Institut Teknologi Bandung. Since 2018, she has been working as a lecturer at the Computer Engineering, School of Electrical Engineering at Telkom University. Her current research field includes mathematical modelling and applied science in tropical diseases.

Ganga Ram Phaijoo is an assistant professor of mathematics at Kathmandu University in Dhulikhel, Nepal. In 2018, he obtained his Ph.D. in Mathematics from the Department of Mathematics at Kathmandu University. His research interests include Mathematical Modeling of Infectious Diseases, Prey-Predator Models, Numerics in ODEs and PDEs, and Applied and Computational Mathematics.

Open Access Policy