International Journal of Information Management Data Insights 3 (2023) 100142

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

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

Review

Transfer Learning Enhanced Vision-based Human Activity Recognition: A Decade-long Analysis

Abhisek Ray[a], Maheshkumar H. Kolekar[a], R. Balasubramanian[b], Adel Hafiane[c,*]

[a] Indian Institute of Technology Patna, Bihta, 801103, India
[b] Indian Institute of Technology Roorkee, Uttarakhand, 247667, India
[c] INSA CVL, University Of Orléans, PRISME, EA4229, Bourges 18022, France

ARTICLE INFO

Keywords:
Deep learning
Machine learning
Transfer learning
Human Activity Recognition

ABSTRACT

The discovery of several machine learning and deep learning techniques has paved the way to extend the reach of humans in various real-world applications. Classical machine learning algorithms assume that training, validation, and testing data come from the same domain, with similar input feature spaces and data distribution characteristics. In some real-world exercises, where data collection has become difficult, the above assumption does not hold true. Even, if possible, the scarcity of rightful data prevents the model from being successfully trained. Compensating for outdated data, reducing the need and hardship of recollecting the training data, avoiding many expensive data labeling efforts, and improving the foreseen accuracy of testing data are some significant contributions of transfer learning in the real-world application. The most cited transfer learning application includes classification, regression, and clustering problems in activity recognition, image and video classification, wi-fi localization, detection and tracking, sentiment analysis and classification, and web-document classification. Human activity recognition plays a cardinal role in human- to-human and human-to-object interaction and interpersonal relations. Pairing with robust deep learning algorithms and improved hardware technologies, automatic recognition of human activity has opened the door in the direction of constructing a smart society. To the best of our knowledge, our survey is the first to link machine learning, transfer learning, and vision sensor-based activity recognition under one roof.. However, this survey exploits the above connection by reviewing around 350 related research articles from 2011 to 2021. Findings indicate an approximate 15% increment in research publications connected to our topic every year. Among these reviewed articles, we have selected around 150 significant ones that give insights into various activity levels, classification techniques, performance measures, challenges, and future directions related to transfer learning enhanced vision sensor-based HAR.

1. Introduction

Humans have evolved into an essential resource capable of handling cognitive tasks, even in many malicious applications. Human intervention is still inevitable in many industrial practices, even in this machinery-driven world of the twenty- first century. Recognition of human action Gupta (2021); Imran and Raman (2020) has become essential for individual performance appraisal. Manual bookkeeping of such activities can be an untidy and error-prone task. As a result, automatic recognition tools have become popular and an area of interest among the research fraternity. Automatic detection of any suspicious or unexpected human behavior will trigger the alarm for either self-correction or manual intervention. Auto- recognition of human activities is nowadays essential for smooth and error-free industrial and institutional operation.

Human Activity Recognition (HAR) datasets are manufactured by taking the knowledge of three fundamental domain-specific aspects: (i) Data related to the sensor device, 2(i) (ii) Data related to the subject/actor, and (iii) Data related to the sensing background. However, the mutable nature of the above three defies the conventional machine learning assumption that source and target data must belong to the same domain. Knowledge transfer came to the rescue by eliminating this conventional machine learning hypothesis. Apart from this, the older training data are sometimes unsuitable for real-time recognition due to the mutable nature of the sensor and environment. Through the help of transfer learning, we can easily exploit the older samples and utilize the valuable information to enhance the classification, regression, and recognition tasks. It is even more difficult and expensive to collect an

* Corresponding author: Dr. Adel Hafiane, INSA center Val de Loire: Institut National des Sciences Appliquees center Val de Loire, 88, Boulevard Lahitolle, 18022 Bourges, France.
E-mail address: adel.hafiane@insa-cvl.fr (A. Hafiane).

https://doi.org/10.1016/j.jjimei.2022.100142
Received 9 May 2022; Received in revised form 19 November 2022; Accepted 26 November 2022

adequate number of training data samples and label it. Transfer learning significantly contributes to many real-world applications by compensating for old data, lowering the requirement and difficulty of recollecting training data, avoiding many expensive data labeling efforts, and boosting the accuracy of testing data.

"Until now, most review articles from activity recognition backgrounds have summarized the context related to either transfer learning or classifier-based machine/deep learning. The activities addressed in those surveys are vision-based or sensor-based. However, our survey enlists only those activity recognition articles where the machine/deep learning classifiers take advantage of the transfer learning techniques to enhance the recognition performance. In this work, we perform a data-centric and classifier-specific extensive survey on vision-based activity recognition. To the best of our knowledge, our survey is the first to review vision sensor-based human activity recognition using transfer learning enhanced machine/deep learning algorithm. Our paper gives insights into various activity levels, classification techniques, performance measures, challenges, and future directions related to transfer learning enhanced vision sensor-based HAR. Our paper gives insights into various action recognition datasets with specifications and levels of activity associated with them after inferring the context of the data. We also address different classification techniques, performance measures, challenges, and future directions related to transfer learning enhanced vision sensor-based HAR. Our survey guides fresh researchers to become familiar with the information and management of existing datasets and learning methods that help to analyze the gaps and opportunities for future research work.

Many studies have reviewed transfer learning and activity recognition separately. However, a few have reviewed activity recognition in transfer learning platforms, and the number has become scarce while talking about sensor-based HAR in transfer learning enhanced platforms. To the best of our knowledge, Cook, Feuz and Krishnan (2013) is the last published review article that addresses HAR in the transfer learning domain. This paper enlists similar research work from 2011-2021 but more oriented to HAR dataset and classification techiniques.. Deng, Zheng and Wang (2014) investigates sensor-based and vision-based HAR by extensively classifying different HAR methodologies based on their pros and cons. Our paper carries similar content but in the transfer learning platform. Unlike previous surveys, this survey paper does not abide by constraints like sensor- based modeling, architecture-based modeling, classifier-based modeling, or dataset-based modeling, as seen in other surveys. This article combines all these models to forge a complete superficial package that will boost creativity for beginner and intermediate-level researchers. The scope of our paper can be further extended to wearable sensor-based and ambient sensor- based HAR Gupta (2021) in the transfer learning domain. This paper precisely explains the transfer learning technique, various steps, and datasets used in vision sensor-based HAR. This survey also introduces a novel classification hierarchical model related to this research domain.

We can find many studies depicting transfer learning and activity recognition separately. To the best of our knowledge, a few analyzed the HAR based on the transfer learning technique, and the number becomes scarce in our research domain. The contributions of our paper are summarized below.

1. To the best of our knowledge, we are the first to divide classification techniques for vision-based HAR in the form of three modular representations. We categorically discuss these classes in detail for future reference.
2. The frequently used visual datasets (source and target datasets) used in HAR are organized based on their year of evolution, mode of representation, frames per second, resolution, classes, subjects, and the number of videos compared.
3. We chronologically summarize the related research articles by comparing their underneath architecture, source/target datasets, the number of detected classes, and their respective accuracy.
4. We have tried to identify potential research gaps and future directions concerning vision-based HAR. We believe it will pilot new researchers in the right direction after saving their investigation time.

The rest of the paper is organized as follows. The research mehodology is discussed in section II. The overview of transfer learning, including its definition and significance, and architecture related to HAR, is demonstrated in Section III. Section IV introduces various HAR datasets, their classification, and hierarchical tabular representation with the specification. Section V outlines the classification techniques used in vision-based HAR with a three-modular representation format. Performaces of various significant articles are summarized in Section VI . The challenges, and various aspects of future directions are briefly expressed in section VII . In section VIII, the contributions and practical implications are briefly discussed. Finally, section IX concludes the paper along with the improvement that can be considered further.

2. Research methodology

We followed Preferred Reporting Items for Systematic review and Meta-Analysis Protocols (PRISMA-P) Tricco et al. (2018) to single out relevant and significant articles related to our research domain. We accomplished this review by adopting three protocols: searching protocol, inclusion, and exclusion protocol, and scoping review protocol.

2.1. Searching protocol

First, we set the search platforms, i.e., search sites, libraries, or digital databases. Most articles included in this review were taken from Web of Science, IEEE Xplore, and Google Scholar digital libraries. We reach out to the relevant articles by putting the exact or relevant keywords or a combination of them. Some of the keywords are "human activity recognition," "video action classification," "transfer learning," "deep learning," "machine learning," "CNN," or the name of different activity recognition databases. Some of the searched sentences are the combination of more than one keyword with effective meanings. We downloaded around 350 articles during initial consideration for further processing.

2.2. Inclusion and exclusion protocol

We only included those vision-based activity recognition articles that adopt machine learning and transfer learning techniques for model designing. Non-English papers were excluded. We considered the date and type of publication (journal or conference), publishing house, and cite score during preliminary screening. Furthermore, we extended this screening procedure to the abstract composition level, where we validated searched articles' themes to our survey theme. Publications with appropriate matches were included. Finally, we filtered out the 150 most significant articles for further review.

2.3. Scoping review protocol

In this last step of methodology, we systematically reviewed the selected papers after thoroughly apprehending many contextual factors in detail. First, we structured the summary observing the background, objective, source of evidence, eligibility criterion, databases, model algorithms, results, and conclusion from the abstract section. Aftermath, we stepped into the detailed sketch of the paper considering aforementioned factors along with some finer details. For example, computational complexity, real-time deployment possibility, limitations, research gaps and opportunities.

3. Overview

3.1. What is transfer learning?

The transfer learning definition can be well illustrated after debating the following four terms.

Source and Target. As we know, the knowledge from abundantly available data is utilized to exploit infrequent sparse data in transfer learning. The abundantly available and easily obtainable data is the source data, and the other infrequent alien data is the target data. The principal goal is to reduce the target data collection and labeling effort.

Domain. Transfer learning paves the way to anticipate future data intelligently after acquiring prior knowledge about source and target data distribution and feature space. Domain (D) is the complete set of knowledge that describes feature characteristics and their distribution probability. Feature space describes the number and type of characters present in the source or target data domain. The marginal probability of a particular feature among all features is known as the marginal probability distribution. The feature space and their marginal probability distribution of source data constitute the source domain, and the same in target data constitutes the target domain. So a domain D is characterized by a set, $\{\chi, P(X)\}$, where χ is the feature space and $P(X)$ Marginal probability distribution; $X = \{x_1, x_2, x_3,\ldots, x_n\} \in \chi$. Two different domains can vary either in terms of feature space (χ_S, χ_T) or in terms of marginal probability distribution ($P(X_S), P(X_T)$). χ_S is the source domain feature space, χ_T is the target domain feature space, $P(X_S)$ is the source domain marginal probability, and $P(X_T)$ is the target domain marginal probability.

Task. The label information of the source and target samples and the conditional probability of that label for a given specific domain feature constitute the task. Label space (γ) is nothing, but the collection of class labels associated with a set of source or target class objects, and the probability distribution of a class object at a given specific domain feature space is called a conditional probability distribution. So, a task T is characterized by a set, $\{\gamma, P(Y/X)\}$, where (γ) is the label space and (P(Y/X)) is the conditional probability distribution for a give domain feature; $Y = \{y_1, y_2, y_3, \ldots, y_n\} \in \gamma$, $X = \{x_1, x_2, x_3, \ldots, x_n\} \in \chi$. Two different tasks can vary either in terms of label space,i.e., γ_s, γ_T, or in terms of the conditional probability distribution,i.e., $P(Y_S/X_S)$, $P(Y_T/X_T)$. γ_S is the source label space, γ_T is the target label space, $P(Y_S/X_S)$ is the source conditional probability, and $P(Y_T/X_T)$ is the target conditional probability.

The above discussion can be concluded by stating that transfer learning has freed the traditional machine learning hypothesis barrier to solving the rising concern on performance and dataset-related cost.

3.2. Why transfer learning?

Researchers have to analyze the nature of both training and testing data before passing through a machine or deep learning architecture. Traditional machine learning methods work on the assumption that the training, validation, and testing data all come from the same domain, with similar data distribution and feature space. However, the above proposition fails to prove its significance due to the existence of real-world data heterogeneity. Therefore, we need to build and train a separate model for different but related tasks. However, an isolated training approach will make the whole process burdensome, expensive, and time-consuming. Even if we go for separate training, the availability of rightful training data makes its path difficult. Transfer learning came to the rescue. This technique helps boost the performance of test/validation data trained on a dataset belonging to either a different feature space or distinct distribution. Compensating for outdated data, reducing the need and hardship of recollecting the training data, avoiding many expensive data labeling efforts, and improving the foreseen accuracy of testing data are significant contributions of transfer learning in a real-world application.

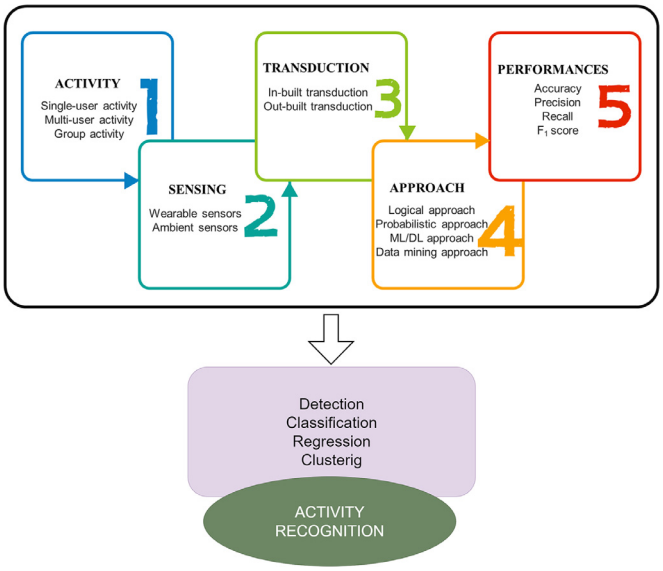

Fig. 1. Steps involved in human activity recognition.

Nowadays, transfer learning takes the reign over similar machine learning knowledge transfer techniques i.e., multi- task learning, domain adaptation, and covariate shift in classification, regression, clustering, and reinforcement learning tasks. Sports video classification, web document classification, image classification of various fields, text classification, sentiment classification Luo and Mu (2022), anomaly detection Al-Sulaiman (2022), emotion recognition Gonegandla and Kolekar (2022) and location estimation based on Wi-Fi signal strength are among the most significant area where practices of different transfer learning techniques are highly beneficial.

3.3. Architecture for human activity recognition

HAR is a prominent area of research while talking about indoor Activity of Daily Living (ADL), outdoor ADL, gesture recognition Anand, Urolagin and Mishra (2021); Chatterjee, Bhandari and Kolekar (2016), sports activity recognition Kolekar and Sengupta (2015), human-object interaction, and human-human interaction. To understand the daily routine activity, we need a systematic tool that can differentiate and understand different activities. Under the diverse situation, the activity recognition algorithm should perform more par than average to establish a well-performing recognition setup.

Five different modules are fused together to form the framework of HAR, as shown in Fig. 1. These five modules are performed activity, sensing type, transduction module, HAR approach, and performance measure.

3.3.1. Activity

People engage in activities throughout the day, from off- bed to on-bed, and even when sleeping. Some human actions are carried out alone, while others are carried out with the assistance of others. These activities can be divided into three categories based on their manner of execution: (i) single- person activity Zhang et al. (2020a) (ii) multi-person activity, Ji, Liu, Pang and Li (2020) and (iii) group activity Tran, Bourdev, Fergus, Torresani and Paluri (2015). The basic issues we normally confront during a group activity are quantifying human roles and integrating these role descriptions into inference techniques. Indoor ADL, outdoor ADL, gesture, and cooking activities can all be classified as single-person or multi-person activities. Sports can be classified as either a single person or a group activity.

3.3.2. Sensing

Advancement in material science and engineering opens the gate for developing large-scale semiconductor units. These units help the researcher to sense physical signals and transduce these into easily processable electrical outputs. The device that performs both sensing and transduction operation is called a transducer. While developing transduction devices, the lowest cost, greater fidelity, and enhanced reliability are major concerns. The occurrence of any physical event within a dynamic environment can be recognized through a sensor before conveying it to the control unit. Over the years, many advanced sensors have evolved, sensing a wide range of various physical, chemical, and biological activities. Most of them are involved in measuring physical properties such as vision, hearing, touch, sense, light, sound waves, pressure, and temperature. Chemical sensors deal with the compositional properties, concentration level, and chemical properties such as the taste and order of the substances. As a subset of chemical sensors, biological sensors or, in short, biosensors, are analytical devices that help in revealing biological states and properties of bio substances. These smart sensors are often used alongside IoT devices to perform intelligence tasks.

Here, we differentiate these sensors from the activity recognition point of view into two classes. First, on-body or wearable sensors comprising accelerometer computes acceleration of action, magnetometer reckons direction and magnetic field of motion, heart rate monitoring device, electrocardiogram displays heart function, electroencephalogram measures brain activity, electromyogram computes muscle tremor, temperature sensors calibrate on- body temperature and pressure sensor which indicates on-body pressure in the course of action. Second, off-body or ambient sensors include the image sensor, video sensor, radio frequency sensor for motion and displacement, the IR sensor for infrared images and videos, the Wi-Fi signal sensor for detection, the GPS track sensor for localization, temperature sensor, and pressure sensor, measure various environmental parameters. Together, these two can update the status of both agent and environment for the contiguous processing unit.

Visual sensing technologies, such as CCTV and camera sensors, are greatly beneficial in keeping an eye on human activities. The quality of the recorded data is decided by assessing image-related attributes and these attributes include viewpoint, lighting environments, illumination changes, occlusion, and image resolution. The HAR data is either in an image sequence or an audio data format, subsequently processed through computer vision and audio signal processing technology Ghosal and Kolekar (2018) to find meaningful information. Different steps may include pre-processing, feature engineering, data modeling, and activity reckoning. Recognizing activity may encompass segmentation, detection, classification, or tracking of the object or subject in interest. These activities are continuously tracked through the sensor attached to the actor/actors or object/objects or environment. Here, we cluster HAR data and environment-related parameters by considering the visual sensor as the primary.

3.3.3. Transduction

Sensor acts as the nervous system, whereas the algorithms and processing units serve as the brain of activity recognition. After signal conversion, the transduction unit connects the sensor output to the processing unit input for further processing. Translating a perceived signal into an appropriate and easily processable form is known as transduction. The processing unit next to it greatly influences the condition of outgoing signals. The sensing and transduction units fuse together to form a transducer most of the time. However, the manufacturer sometimes finds it much easier to position the transduction device outside the sensor, known as out-built transduction.

3.3.4. Approach

The desire to improve a variety of decision variables, such as compatibility, mobility, cost, efficiency, and accuracy, has pushed researchers to pursue new ideas. Every day, the high-tech world takes on a new face by incorporating advanced technical characteristics that result in a more efficient and resilient product than its predecessor. For more efficient and accurate activity recognition, a variety of techniques and algorithms are being developed.Some are no longer in use, while others have advanced to become state-of-the- art techniques. In HAR, the probabilistic and statistical modeling technique expresses uncertainty in reasoning about the performer's activities, plans, and goals. Naïve Bayes Classifier (NBC), Decision Trees (DT), and HMM are some examples of probabilistic approaches. Machine learning approaches take over the probabilistic approach as it depends on other independent inferences that complicate the recognition process in a real-world environment. K-Nearest Neighbors (KNN), Bayes classifier Minimal Learning Machine (MLM), Support Vector Machine (SVM), Extreme Learning Machine (ELM), and Multilayer Perceptron (MLP) are some popular ML techniques practiced in inactivity recognition platforms.

Nowadays, deep learning approaches have become popular in the HAR domain because they can achieve the highest success rate in real-world scenarios. The deep learning technique sets a platform for automatic feature selection and learning compared to machine learning handcrafted feature extraction. Deep Neural Network (DNN), Convolutional Neural Network (CNN), autoencoder Aslam and Kolekar (2022); Aslam, Rai and Kolekar (2022), Restricted Boltzmann Machine (RBM), Recurrent Neural Network (RNN), and generative Adversarial Network (GAN) are some of frequently used DL techniques. These approaches can be used for human activity detection or recognition, classification, regression, and clustering. Different classification and detection techniques are discussed in Section 5 thoroughly.

3.3.5. Performances

Every user should be aware of the acceptable degree of predicting results in a predictive activity. To do so, we must first determine the correct prediction probability. It can be thought of as the ratio of all the correct guesses to the entire number of forecasts. Intuitively, users of a prediction model want to know how much they can trust the forecast results. On the other hand, the researcher devised a set of statistical performance metrics for the quantitative description of predicted performance in numerous aspects under a variety of conditions. When discussing class, we come across two types of labels: truth labels and prediction labels. Prediction labels are model-predicted tags following evaluation, either during validation or testing, while ground truth labels are the actual class label to which that sample belongs. According to these definitions, a true positive (TP) sample is one in which both the ground truth and predictive labels are positive, while a true negative (TN) sample is one in which both the ground truth and predictive labels are negative. If the ground truth label is negative, but the prediction label is positive, the prediction label is false positive (FP); if the ground truth label is positive but the prediction label is negative, the prediction label is a false negative (FN). The most often used performance measures for activity recognition are listed below.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\%$$

$$Precision = \frac{TP}{TP+FP} \times 100\%$$

$$Recall = \frac{TP}{TP+FN} \times 100\%$$

$$F_1\,score = \frac{Precision \times Recall}{Precision + Recall}$$

Apart from these, two more measurements, the Mean Levenshtein Distance (MLD) score and mean average precision, are referenced in several articles. We sort transfer learning enhanced vision-based AR articles from 2011 to 2021 in Table 2 where we only take accuracy as the performance parameter.

4. Visual HAR datasets

The evolution in HAR dataset with time occur due to the content variation and technological mutation. Some of these datasets are old,

Table 1
Popular HAR dataset with specification.

Dataset	FPS/Resolution	Classes/Subjects/Videos	Activity Level
MP-II Cooking Rohrbach, Amin, Andriluka and Schiele (2012)	29.4/1624 × 1224	65/12/44	H-O Level
UCF-101 Soomro, Zamir and Shah (2012)	25/320 × 240	101/-/13,320	H-O/Group Level
DML Smart Action Mohsen Amiri et al. (2013)	30/2HD+1VGA	12/16/932	Atomic/H–O Level
Hollywood 3D Hadfield and Bowden (2013)	24/1920 × 1080	14/-/650	H-O/H–H Level
YouTube Sports 1 M Karpathy et al. (2014a)	-/227 × 227	487/-/11,33,158	H-O/Group Level
Thumos' 14 Thumos14	-/-	101/-/18,000	Atomic/H-O/Group Level
Northwestern-UCLA Wang, Nie, Xia, Wu and Zhu (2014)	30/640 × 480	10/10/1475	Atomic/H-O Level
UTD_MHAD Chen, Jafari and Kehtarnavaz (2015)	30/640 × 480, 320 × 240	27/8/861	Atomic/H-O Level
ActivityNet Caba Heilbron, Escorcia, Ghanem and Niebles (2015)	30/1280 × 720	203/-/27,801	H-O Level
THUMOS'15 Gorban et al. (2015)	-/-	102/-/23,500	Atomic/H-O/Group Level
NTU RGB+*D* 60 Shahroudy, Liu, Ng and Wang (2016)	30/1920 × 1080, 512 × 424	60/40/56,880	Atomic/H-O/H-H Level
YouTube 8 M Abu-El-Haija et al. (2016)	1/-	480/-/82,64,650	H-O/Group Level
Kinetics400 Kay et al. (2017a)	-/658 × 1022	400/-/3,06,245	H-H/H-O Level
PKU-MMD Liu, Hu, Li, Song and Liu (2017)	30/1920 × 1080, 512 × 424	51/66/20,000	H-O/H-H Level
Something-SomethingV2 Goyal et al. (2017)	12/96 × 96	174/1133/2,20,847	H-O Level
AVA Gu et al. (2018a)	1/451 × 808	80/-/230K	Atomic/H-O Level
MLB-YouTube Piergiovanni and Ryoo (2018)	60/-	20/-/4290	H-O/Group Level
Kinetics600 Carreira, Noland, Banki-Horvath, Hillier and Zisserman (2018)	-/658 × 1022	600/-/4,95,547	H-H/H-O Level
SoccerNet Zhou, Xu and Corso (2018)	25/1280 × 720	3/-/6637	H-O/Group Level
YouCook2 YouCook2	-/-	89/-/2000	H-O Level
NTU RGB+*D* 120 Liu et al. (2019)	30/1920 × 1080, 512 × 424	120/106/1,14,480	Atomic/H-O/H-H Level
Kinetics-700 Carreira, Noland, Hillier and Zisserman (2019)	-/658 × 1022	700/-/650K	H-H/H-O Level
MOD20 Perera, Law, Ogunwa and Chahl (2020)	29.97/720 × 720	20/-/2324	H-O/Group Level
HAA-500 Chung, Wuu, Yang, Tai and Tang (2021)	-/1080 × 720	500/-/10,000	Atomic/H-O/Group Level
EduNet Sharma, Gupta, Kumar and Mishra (2021)	30/1280 × 720	20/-/7851	H-O Level
TAD-08 Gang et al. (2021)	-/720 × 576	8/-/2048	H-O Level
Win-Fail Parmar and Morris (2022)	-/1080 × 720	4/-/1634	Atomic/H-O Level

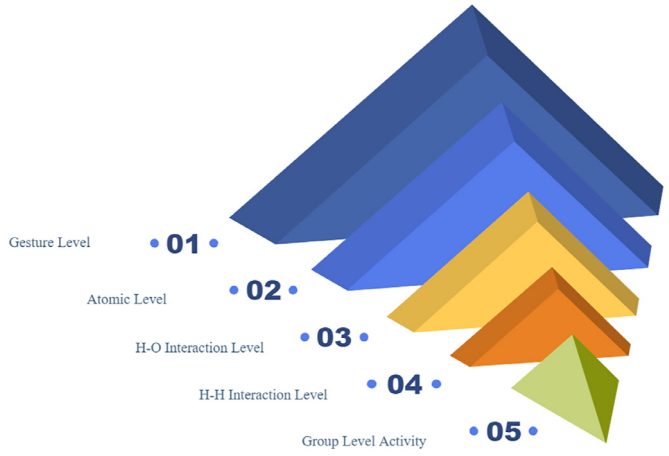

Fig. 2. Level of Activity in Various HAR Dataset.

some are modern, and a few of them have become the benchmark based on their speed, accuracy, and adaptability. Modern datasets have a more detailed description compared to older ones. These descriptions or specifications can be physically perceptible or imperceptible. Variation in frame per second (fps), action count, actor count, videos count per action, modality, resolution, annotation mode, and viewpoint are physically perceptible parameters. Light illumination and occlusion are examples of the physically imperceptible parameter. These factors mentioned above decide the quality of a dataset. Better quality dataset helps to open the gate for exploring more composite models, whereas a challenging dataset helps in scrutinizing the generalizability and the robustness of a model.

Several factors help analyze a dataset and decide the group from where that dataset should belong. We can cluster vision- based HAR datasets into several groups after reckoning the type of action, sensing modality, viewpoints, and the nature of the data. As shown in Table 1, we cite the most widely used datasets related to this domain and classify them by activity type. As shown in Fig. 2, the collected datasets can be categorized into five levels of activity: gesture level activity, atomic level activity, Human-Object (H–O) interaction level activity, Human-Human (H–H) interaction level activity, and group level activity.

(a) *Gesture level activity*: The gesture is defined as a purposive human body movement to convey some meaning or idea. A gesture is usually performed within a short period of time and can be considered the elemental activity among the five groups. Hand-waving, facial expressions, eye movements, and head shakes are examples of gestures.
(b) *Atomic level activity*: An activity can be a single atomic action comprising a sequence of interrelated gestures. This level of action is performed only by an actor without the involvement of any subject or object. Hence, it is sometimes referred to as a solo activity. Knocking, swimming, walking, jogging, and running are examples of atomic-level activity.
(c) *H–O interaction level activity*: Some activities are the results of interaction established between two agents. As we are talking about human activity, one of the agents must be a human, and the other may be a human or an object. If we take the object as a second agent, the action may be termed a Human-Object interaction. The treatment done to the object by the human will decide the type of H–O interaction. All the cooking activities, kicking, hammering, eating, drinking, throwing, and sports activities like weightlifting, batting, and bowling can be laced under the H–O interaction category.
(d) *H–H interaction level activity*: Many activities are taken place with two humans as the agent. This kind of activity results from the interaction between two people; it is called two-person activity or Human-Human interaction. Hugging, wrestling, and shaking hands are some examples of human-human interaction.
(e) *Group level activity*: A group activity is a composite kind of activity that may require interaction between more than one person or one or more than one objects. It involves many sequences of gestures, actions, and interactions. A group study, a cricket match, a group discussion, and a presentation are some examples of group-level activity.

(a) Recognized activity in the UCF101 dataset Soomro et al. (2012b)

(b) Various instances of activities of SPORTS1M dataset Karpathy et al. (2014a)

(c) Various activities in the AVA dataset Gu et al. (2018b)

(d) Extracted frames of MLB-YouTUBE video dataset Piergiovanni and Ryoo (2018b)

Fig. 3. Samples of extracted frames and recognized activity of state-of-the-art datasets.

The datasets related to these activities level is summarized in Table 1 with many related specifications, and the frames of various state-of-the-art datasets are shown in Fig. 3.

5. Classification technique used in vision-based har

The advancement in sensor and hardware technology has added new feathers in the background of the machine and deep learning algorithms. A new algorithm always replaces its previous version while accomplishing a robust model with superior performances. Many deciding factors are responsible for performance enhancement. The intent to improvise these factors, such as compatibility, portability, cost, efficiency, and accuracy, has compelled researchers to adopt numerous efficient and accurate algorithms. Some are obsolete, some have restricted use, and some have become state-of-the-art methods over time. In this paper, we group them into three learning-based approaches; (i) generative-based approach, (ii) discriminative-based approach, and (iii) graph-based approach, as shown in Fig. 4.

5.1. Generative-based approach

Generative models are semi-supervised probabilistic approaches that determine conditional probability distribution $P(X/Y)$ of training data samples (X) considering their corresponding labels (Y). These posterior distributions are obtained by applying the Bayes rule to predict the class probability of a test input. These robust models are less inclined to overfitting issues that give good performance in a limited data environment.

5.1.1. Hidden markov model

Hidden Markov Model (HMM) is a probabilistic and discrete-time framework that advances through a series of hidden states to give a final output observation sequence. Each level of the hidden states is associated with three types of probabilities: starting probability, transition probability, and emission probability. The occurrence probability of an unobserved state is termed as starting probability of the hidden states. The probability at which one hidden state makes the transition to another hidden state of the same level or to the state itself is called transition probability. The state transition probability between two different states is called emission probability. These observations are first diverged out and subsequently converged to give the final output probability. This stochastic model is run through the Markov process, hence, called so. Baum-Welch algorithm trains an HMM framework with five states in a left-to-right composition to yield promising results on the MSRC-12 Kinect gesture dataset Cabrera, Sanchez-Tamayo, Voyles and Wachs (2017). A fast simplex HMM (Fast- SHMM) is introduced by computing MAP adaptation of the Universal Background Model (UBM) for each training sample Rodriguez, Orrite, Medrano and Makris (2017a). The computational cost is reduced by minimizing the number of Gaussians in UBM and estimating the Expectation-Maximization optimum quickly. Like Cabrera et al. (2017), Rodriguez et al. (2017a) employ the HMM-based one-shot learning (OSL) approach but on the Weizmann dataset taking HMDB51, Olympic Sports, and Virat Release 2.0 as source domain datasets. Apart from the above two, Rodriguez, Orrite, Medrano and Makris (2017b); Wen and Zhong (2015) are also HMM-based articles that employ instance transfer learning for knowledge transfer. Arif Ul Alam et al. (2021) proposed Adaptive Order HMM (AO–HMM) and Crossover Path Disambiguation Algorithm (CPDA) to address the issues like fast and multiuser target tracking and re-cluster voxelized Point Cloud Data (PCD) for target activity recognition. The changing order of HMM is decided by the number of active states and their neighbors. Arif Ul Alam et al. (2021) use transductive transfer

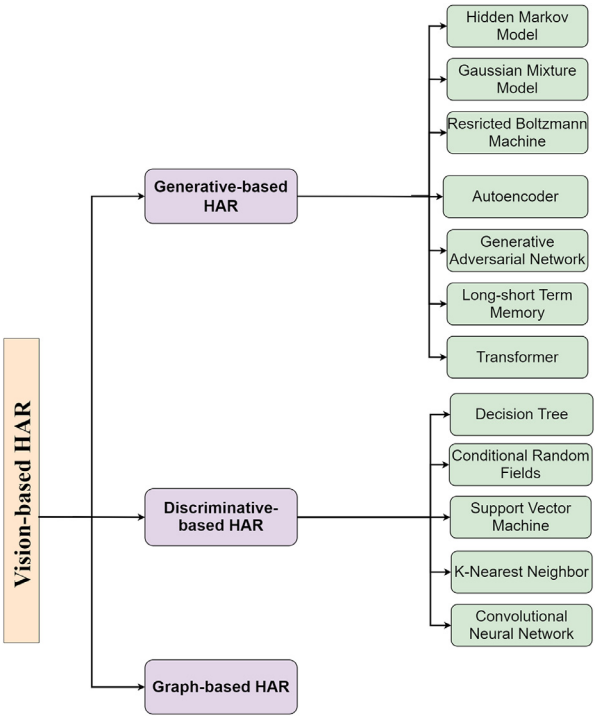

Fig. 4. Classification techniques used in Vision-based HAR.

learning to exercise on the PALMAR and Benedek dataset for recognizing human- object interaction activity.

5.1.2. Gaussian mixture model

K-means clustering is considered a hard clustering method or distance-based clustering method. Hence, it cannot express its significance in undistinguishable or multi-label data environments. So, we shifted to a soft clustering model called the Gaussian Mixture Model (GMM), where a distribution- based clustering technique is adopted instead of a distance- based. In GMM, a dataset of D features can have a mixture of k Gaussian distributions. Each distribution represents a cluster head defined by the D length mean and $D \times D$ co- variance matrix. The expectation-maximization technique determines these variables (means and co-variances) and sets model parameters accordingly.

Xing et al. (2019) use the GMM algorithm to segment the raw RGB images of small and unseen target datasets and send the segmented data to a CNN (AlexNet, GoogleNet, and ResNet-50) model for activity recognition. Xing et al. (2018) use GMM-based segmentation and only pre-trained AlexNet model to implement the same inductive transfer learning through fine-tuning. Ntalampiras and Potamitis (2018) use temporal, spectral, and wavelet features to identify statistically-closely located classes using GMM and KL divergence algorithm. Class-specific HMM and universal HMM use these distance- based class features for class prediction. ESN-based transfer learning technique is adopted to categorize seven human-object interaction level activities. Variational Bayesian Inference (VI) is the generalization of the expectation-maximization approach, which maximizes the likelihood iteratively Jänicke, Tomforde and Sick (2016a). VI is used to determine the latent features of GMM, responsible for reducing the model complexity, and nullify the need for a specific number of components a priori. Transductive transfer learning is used for self-improvisation, i.e., new node insertion.

5.1.3. Restricted boltzmann machine

Restricted Boltzmann Machine (RBM) is an unsupervised generative network with fully connected nodes across layers (bi-partite node configuration and hence the term 'restricted') that is capable of learning probability distribution from the seen data to make inferences about unseen data. It has a visible or input layer(v) associated with the seen data and one or multiple hidden layers (h) pointing out the unseen inference data having no output layer. RBM is an energy- based model used in classification, regression, dimensionality reduction, feature learning, collaborative filtering, and topic modeling. Boltzmann distribution (Gibbs Distribution), derived from statistical mechanics in thermodynamics, is implemented to explain the impact of entropy on different states in RBM. It is associated with two biases; (i) hidden bias that helps to produce activation on the forward pass (ii) input bias that helps to produce activation on the backward pass. The gradient- based contrastive divergence algorithm is implemented to carry out learning during training. Multiple RBMs are stacked together to form Deep Belief Network (DBN) Kolekar (2011) to perform layer-wise training. Roder et al. (2021) first, introduce the spectral DBN on HMBD-51 and UCF-101 HAR datasets using the domain adaptation technique. Gradient- DBN and Aggregative-DBN are proposed to employ image gradient and frame fusion in video-based HAR. Gradient- DBN and Aggregative-DBN are proposed to employ image gradient and frame fusion in video-based HAR. A Binary- binary RBM and a gaussian-binary RBM are stacked together to optimize the weights and learn the informative features of triaxial accelerometer HAR data Alsheikh et al. (2016). To train and fit the model parameters, the underlying model should go through the pre-training stage (unsupervised and generative) and fine-tuning stage (supervised and discriminative).

5.1.4. Autoencoder

Autoencoder (AE) is an unsupervised generative ANN model that is embodied with an encoder layer, code layer, and decoder layer (mirror to the encoder layer). The encoder layer only considers the informative data representative of the input to generate low dimensional code and stores it in the code layer, which is the latent space representation of the input data. The decoder layer later collects these codes and reconstructs them back to generate output containing only valuable features. These generated outputs are identical and equidimensional to the input. Regularized (sparse, denoising, and contractive), concrete and variational AE are the most common types used in many machine learning tasks like facial recognition, activity recognition, dimensionality reduction, anomaly detection, machine translation, drug discovery, and popularity prediction. Khan and Roy (2018) use a pre-trained transfer learning framework called UnTran that transfers the first two layers of the source trained Deep Sparse Autoencoder (DSAE) to incorporate with SVM classifier for recognizing human activity on Opportunity, WISDM, and Daily and Sports datasets. This multi-layered classification model helps generalize the model to overcome user-related, sensor-related, and environmental- related diversities. A combined model performs domain adaptation for re-annotation in the cross-dataset platform Sanabria and Ye (2020). The combined model fuse two learning techniques for human activity recognition; (i) knowledge and data-driven learning technique, and (ii)Unsupervised Domain Adaptation technique. The Variational Auto-encoder (VAE) in UDAR has achieved encouraging outcomes while learning latent space representation in minimizing the distance across Aruba and Twor datasets. The discussed framework is effective and robust for adapting the divergence in training data count and sensor noise settings. The semi-supervised Inverse Autoregressive Flow (IAF) based VAE is associated with Bi-Directional GAN (Bi-GAN) classifier to implement Zero- Shot Learning (ZSL) for HAR using synthesized features on UCF101, HMDB51, and Olympic datasets Mishra, Pandey and Murthy (2020). The above model adopts a decoder with skip connections to stabilize the training and prevent overfitting. Khan and Roy (2018) employ the inductive transfer learning method, whereas Mishra et al. (2020); Sanabria and Ye (2020) uses the transductive setting for transferring knowledge across datasets. Autoencoders are very suitable in unsupervised applications like anomaly activity recognition, where we define the data under either normal or abnormal categories.

5.1.5. Generative adversarial network

GAN Aggarwal, Mittal and Battineni (2021) is a synchronous generative model that comprises two sub-models (generator and discriminator). A generator generates a random sample of target dimensions by taking a fixed-length vector as input and sending it over to a discriminator for binary classification (real or fake) along with an actual target domain sample. The generator tries to mislead the discriminator by generating random output close to the real input. Moreover, the discriminator tries to protect from being fooled by updating its weight. This process of "making fools" and "being fooled" is performed iteratively to accomplish recognition and generation tasks that come under unsupervised, semi-supervised, fully supervised, and reinforcement settings. Vondrick, Pirsiavash and Torralba (2016) use Spatio-temporal convolutional GAN for unsupervised HAR in videos from Flicker. Spatio-temporal convolutional architecture helps untangle a scene's foreground from its background. GAN is employed to generate and classify video samples by utilizing scene dynamics. The proposed conditional GAN framework is fed with a class prototype vector to implement Generalized FSL (GFSL) on UCF-101, HMDB- 51, and Olympic-Sports datasets. The GFSL sub-module addresses the inadequate data and seen-data biasing problems. Class prototype Transfer Network (CPTN) generated class prototype vectors with random noise are fed to the generator module to produce synthetic features. The generator has gone through an iterative update based on the discriminatory loss to make random synthetic features close to the real features. A classifier is trained with both real and GAN-generated synthetic features to efficiently address novel data classification problems. Common latent semantic representation can be an excellent asset to generalizing a model in the zero-shot learning setting.

Zhang, Li and Ogunbona (2017a) take connotative and extensional relations for solving poor generalization problems on UCF-101 and HMDB-51 datasets. GAN-based model synthesizes action features and word vectors of unseen classes by exploiting this representation from seen examples. A knowledge-based graph is prepared by relating the word vectors to their corresponding object. Finally, an attention-based Graph Convolutional Network (GCN) is employed to classify the novel samples with better accuracy and enhanced generalizability. Standard and generalized settings of transductive ZSL are realized in Ji et al. (2020) through Bi-directional adversarial GAN and Inverse Auto-regressive flow-based VAE on UCF-101, HMDB-51, and Olympic-Sports datasets. Skip connection of decoder in VAE not only results in more stable training but also additionally prevents overfitting.

5.1.6. Long-Short term memory

Long-Short Term Memory (LSTM) is an RNN variant where multiple layers stacks together to perform time-series signal processing to preserve long-term dependencies between information sequences. The presence of more complex interactive layers in LSTM helps to realize the significance of previous sequential knowledge in manipulating future ones. The whole internal processing is carried out by passing the earlier information through three gates; (i) Forget gate that decides whether completely forget or complete keep the past information, (ii) Input gate that allows only the relevant input information by discarding the others, and (iii) output gate that replace the old cell state with the new one after concatenating the concerning forget gate signal and input gate signals.

LSTM is used as a controller in Ma, Zhang, Wang, Qi and Chen (2020) that controls the gateway (read and write heads) between the received input signal and the external memory module. Memory encoding and retrieval are the primary goal of the read and write heads. Kay et al. (2017b) incorporate the LSTM layer and batch normalization layer that receives spatial features from the CNN module to perform state encoding, temporal order capturing, and long dependency exploring. Read attention- based bi-directional LSTM is used in Shi, Zhang, Xu and Cheng (2020). The discriminative features from CNN are fed to bi-LSTM that contains a forward LSTM module and a backward LSTM module. Similarly, two stacked bi-directional LSTMs look forward and backward in time to garner fine-grained sequence information in Fu, Damer, Kirchbuchner and Kuijper (2021). Time-dependent video-level representation is generated by feeding aggregated fixed-length spatial features from a combined model of ResNet, and AlexNet Careaga, Hutchinson, Hodas and Phillips (2019). A transformer architecture comprising LSTM and class-wise attention module helps re-weight the cross-domain data by assigning the higher weight to more informative data. All of these references adopt inductive learning platforms for transferring knowledge. LSTM is an effective classification tool commonly used alongside CNN for Spatio-temporal exploration while interpreting the video data in HAR.

5.1.7. Transformer

A transformer is an encoder-decoder structure-like model that adopts a self-attention mechanism without counting recurrence and convolution. The transformer processes the input data sequence in a random order that helps allow more parallelization compared to different RNN models. In order to manipulate sequential input data in many NLP and computer vision tasks, the transformer runs through a differential weighting mechanism that measures the contribution of each part by assigning some weights in constructing the whole body. In Haresamudram et al. (2020), the transformer processes the 1D wearable sensor data from the accelerometer and gyroscope through knowledge transfer in two steps for self-supervised HAR learning. The unlabeled continuous data is used to learn the weights of the encoder by self-supervised learning and subsequently use these weights to extract the features further. Cheng et al. (2021) uses a hierarchical transformer to learn skeleton-based action features in an unsupervised manner on NTU RGB+*D* and NW-UCLA datasets. It uses motion predictive pre-training methods that come under transfer learning to learn efficient and distinctive representation. Like Cheng et al. (2021); Haresamudram et al. (2020), Zaher Md Faridee et al. (2022) also uses the transfer learning technique on the proposed STranGAN (spatial transformer-based GAN) model for inertial sensor-based HAR applications. This paper uses domain adaptation via feature alignment to transfer the knowledge between source and target without any labeled training data requirement. The transformer is a highly effective and popular video data interpretation technique despite its highly complex and data-hungry nature.

5.2. Discriminative-based approach

Discriminative models, also called conditional models, are a class of logistical models used for classification or regression. They distinguish decision boundaries through observed data, such as pass/fail, win/lose, alive/dead, or healthy/sick. Logistic regression, conditional random fields, and DT can be categorized under discriminative classifiers. The naive Bayes model, Gaussian mixture model, variational autoencoder, and GAN can be categorized under generative classifiers.

5.2.1. Decision tree

A DT is a flowchart-like tree structure comprising nodes and branches for classification and regression tasks. We can visualize these nodes and branches in three segments: internal connecting nodes, interconnecting branches, and leaf nodes. Each connecting node evaluates an attribute of a given classification or regression task. The branch corresponding to that particular node epitomizes the evaluation result of that attribute, and the terminal node (leaf node) holds a class label for that task. The input feed to the DT may be a discrete set of values or a continuous variable. Based on this, we can specify DT as a classification tree or regression tree, respectively. The superior clustering technique in DT promotes it as a good regressor or a well-performed classifier in a restricted data environment. Integration of new unseen sensors leads to an extent in input space. DT needs to be reformed to adopt this change by replacing specific leaf nodes of the original tree with a subtree Jänicke, Tomforde and Sick (2016b). An iterative semi-supervised training approach is endorsed in Bhattacharya, Nurmi, Hammerla and Plötz (2014) called En-Co-Training. A pool of randomly sampled data generated from an unlabeled opportunity and challenge dataset is generated using this algorithm which is later trained with DT. A DT classifier is deployed on the features extracted from the last layer excluded ResNet-50 network Loey, Manogaran, Taha and Khalifa (2021). The DT classification model computes the output label based on information gain and entropy function. Jänicke et al. (2016b) adopt transductive transfer learning, whereas the Bhattacharya et al. (2014) and Loey et al. (2021) follow an inductive learning platform.

5.2.2. Conditional random fields

As opposed to the problem that arises in image and video recognition using Bag-of-Words (BoW) techniques which stress the presence and absence of word(s) over the sequence of words, Conditional Random Fields (CRF) instead emphasizes word sequences over just word(s). This probabilistic classifier exploits the contextual information of all neighbor classes in a class predicting task and utilizes that knowledge to model the decision boundary between them. Weights are estimated using a maximum likelihood estimator for each user-defined feature. It is then followed by a constant normalization term that represents the sum of all possible state sequences to calculate the required conditional probability distribution. After completing the training with the CRF algorithm on IMD, CDG13, and MSRC-12 gesture recognition datasets, a likelihood metric is activated for prediction Cabrera and Wachs (2017). Wang, Chen, Hu, Peng and Philip (2018a) employ Stratified Transfer Learning (STL) as a knowledge transfer technique and CRF with 30 trees as the classifier for six different feature engineering methods. Here, STL

is meant for transferring knowledge between domains. The same architecture is used in Chen, Wang, Huang and Yu (2019) to evaluate the transfer learning performance between various positions. The one-shot inductive transfer learning is used in Cabrera and Wachs (2017), whereas Wang et al. (2018a) and Chen et al. (2019) follows the transductive platform on OPPORTUNITY, PAMAP2, and UCI DSADS activity datasets.

5.2.3. Support vector machine

Before diving into the deep learning era, the use and popularity of the supervised SVM have become sky-high among classification and regression models. This learning paradigm projects each data sample into a point in *n*- dimensional space. Then, it sets a best-suited hyperplane or decision boundary by maximizing the distance from each category to that boundary. The position of the new sample from that hyperplane in n-dimensional space will decide its class. A binary classification problem adopts a linear SVM, whereas a multi-classification problem uses kernel based non-linear SVM architecture. The extracted Spatio-temporal features of gesture data from a 3D Inception-ResNet model with separable convolution are forwarded to an SVM classifier Li et al. (2021). The above model outperforms many state-of- the-art architectures in performance, computational cost, and efficiency. SVM classifier with radial basis function kernel is used to recognize gesture level activity in Cabrera and Wachs (2017). Inductive transfer learning, or more specifically, OSL, is used in Li et al. (2021) and Cabrera and Wachs (2017). Tran et al. (2015) uses SVM as a classifier in the inductive transfer learning domain. A multi-class Hierarchical SVM (HSVM) is adopted to train on ambient sensor features of synthetic, TU Darmstadt, and RCC datasets which help detect instant semantic attributes of test samples Alam and Roy (2017). The confidence score of the HSVM classifier is measured by Contextual Informativeness (CI). Local dense trajectory video features from UCF101, FCVID, Sports1M, and ActivityNet datasets are aggregated into video-level feature vectors to train a linear SVM classifier Gan, Lin, Yang, De Melo and Hauptmann (2016). The above two models endorse transductive ZSL. Chen et al. (2019) also employs transductive transfer learning techniques on SVM- based classifiers. Rahmani and Mian (2015) trains the SVM classifier on IXMAS and N-UCLA datasets to perform transfer learning in cross-view and cross-dataset scenarios, respectively. In the era of machine learning, SVM has become a very popular and effective classification tool used in the HAR domain. However, its use has been restricted in this large-scale data domain.

5.2.4. K-Nearest neighbor

KNN is another class of supervised learning model proposed for classification and regression practices using the distance matrix. K denotes the number of nearest labeled data points or trained samples considered for evaluating the distance matrix. The respective Euclidean distance of KNN from the test data point is aggregated to compose this matrix. So, a distance matrix illustrates the feature similarity index between the new unlabeled data and its K nearest available labeled data. More congruent the features, the lesser the Euclidean distance, and the test data become more biased toward that class label. KNN learning algorithm is sometimes termed non-parametric learning as no mapping function indulgent, lazy learning as the whole dataset is stored for inference, and instance-based learning as weights are not learned. Depending on the context of use, the output may be a class membership value or an object property value, i.e., KNN classification or KNN regression.

The motion and texture features of the Chalearn gesture challenge and NTU RGB+*D* dataset are extracted using co- variance descriptor after building a Bag of manifold words (BoMW) representation Zhang et al. (2017b). These local features of the distinct category are passed through the KNN classifier to perform the one-shot learning gesture recognition. Key points around motion patterns of the ChaLearn gesture database are detected and tracked by the Shi-Tomasi corner detector and sparse optical flow Karn and Jiang (2016). The Gradient Location and Orientation Histogram feature descriptor is then activated to describe the concerning features of these key interest points. These visual features are clustered and subsequently classified by the k-means algorithm and KNN classifier, respectively, to implement OSL. Apart from Karn and Jiang (2016), Bhattacharya et al. (2014); Zhang et al. (2017b) adopts the same KNN-based architecture for inductive transfer learning implementation. Lang et al. (2018) propose a KNN-based domain adaptation model for micro-doppler data classification after fusing three domain-invariant features, i.e., low-level deep features from CNN, empirical features, and statistical features. Aftermath, a KNN classifier is adapted to classify seven human activities. An Adaptive Spatial- Temporal Transfer Learning (ASTTL) approach is introduced Qin, Chen, Wang and Yu (2019) to deal with negative transfer and domain intensive transfer in cross-domain HAR. The spatial features are exploited by weighting relative importance between the marginal and conditional probability distribution and temporal features by incremental manifold learning. KNN is used as a baseline classifier over UCI DSADS, UCI-HAR, USC–HAD, and PAMAP2 datasets. Along with Lang et al. (2018); Qin et al. (2019), Xu, Hospedales and Gong (2016) also follows the same KNN-based transductive transfer learning mechanism. The use of KNN has been restricted in present days due to its low performance but is still prevalent in restricted data environments and unsupervised learning conditions.

5.2.5. Convolutional neural network

CNN is a deep learning architecture where spatial information of image and video (in vision-based) data is explored through repeated convolution operations for different vision-based applications. Convolutional layers have passed through various levels of transformation to fetch more significant features effectively. The activation, pooling, batch normalization, and dropout layers are other supporting layers to improve the feature's quality and computational efficiency by suppressing the noise and parameter count. Some transfer learning-based CNN architectures are either pre-trained with a definite large dataset or customized by the user according to their dataset and application. AlexNet, GoogleNet, Inception, VGG, ResNet, DenseNet, and EfficientNet are examples of pre- trained transfer learning architecture primarily trained on the ImageNet dataset.

Karpathy et al. (2014b) use the multiresolution CNN model to implement transfer learning on a large (487 classes) Sports- 1 M dataset. A two-steam CNN model is trained on the resolution images of this dataset by various fusion techniques. Five combinations of convolution and max-pooling layers followed by two fully connected (FC) layers are trained on ILSVRC-12 datasets to get a final output of 1000 class distributions Liu, Mei, Zhang, Che and Luo (2015). Eight convolution layers, five pooling layers, and two FC layers are forged together to form a 3D ConvNets model called C3D and trained on a large Sports-1 M dataset for weight initialization Zhu and Newsam (2017). This C3D architecture is later applied to the ActivityNet dataset for classification performances. The last pooling layer of ResNet-50 is fed to an LSTM network with batch normalization to get globally pooled Spatio-temporal features of the Kinetics dataset Kay et al. (2017a). The trained weight of this model is later on various small datasets to validate its performances. The output of a modified pre- trained ResNet-18 comprising 17 convolutions and a pooling layer is fed to another three-layered head model to compute the classification score Du, He and Jin (2018). An avg pooling, an FC, and a softmax layer are stacked together to form the head model. The base model is fine-tuned with the micro-doppler dataset and validated on simulated micro-doppler data. Perrett, Masullo, Burghardt, Mirmehdi and Damen (2021) follow CNN-based architecture in the inductive transfer domain for different vision-based HAR datasets. Akbari and Jafari (2019) follows a similar kind of CNN-based architecture in the transductive transfer learning platform for different vision-based HAR datasets. However, CNN has become a very effective and widespread model nowadays due to its high performance and low complexity in the image and video domain.

5.3. Graph-based approach

Multiple modalities are involved in recognizing a human activity, such as pose, appearance, optical flow, depth, and skeleton. Although the dynamic human skeleton modality is a powerful descriptor in identifying the human action irrespective of illumination change and background dynamics, it has rewarded relatively less attention. Yan et al. (2022) proposed a sensor-based graph model called HAR-ResGCNN that combines graph neural residual structure with transfer learning in a cross-dataset setting to validate the performance in PAMAP-2, mHealth, and TNDA action datasets. The use of deep transfer learning on the data derived from the accelerometer, gyroscope, and magnetometer makes the convergence speed faster and the learning curve better. A Spatio-Temporal GCN trained on NTU-RGB+*D* 60 dataset combined with a zero-shot learning module can able to predict unseen activity on which it never trained Jasani and Mazagonwalla (2019). Although the graph- based HAR algorithm is very effective in view-point variation and background changes, it has received comparatively less attention compared to other classification techniques.

6. Performance Comparison of Transfer Learning enhanced vision-based har

We have already discussed different types of datasets, classification techniques, and various performance parameters used in vision-based HAR in Sections 3, 4, and 5, respectively. This section summarizes transfer learning enhanced vision- based AR articles from 2011 to 2021 in the order of year/link, classification technique, target/source dataset, and accuracy in Table 2. It is clearly observed from Table 2 that the dominance of automatic deep learning approaches was finally turned up over hand-crafted machine learning approaches around the year 2013 after the evolution of the pre-trained CNN framework during the ImageNet competition (ImageNet Large Scale Visual Recognition Challenge). Most spatial feature extraction frameworks in Table 2 are based on the CNN technique. CNN is implemented with sequence-based approaches (RNN, LSTM, attention, and transformer) to exploit features in the temporal dimension along with the spatial one Yue-Hei Alwassel, Heilbron and Ghanem (2018); Mutegeki and Han (2019); Ng et al. (2015); Perrett et al. (2021); Wang et al. (2016b). It can also be observed that some of the articles used deep learning approaches (mostly CNN techniques) for feature extraction and machine learning approaches for classification (mostly SVM and KNN). To take the advantage of transfer learning, most of the HAR framework has taken the advantage of the pre-trained models (ResNet, AlexNet, GoogleNet) whose parameters are optimized on large-scale datasets like ImageNet. In Section 5, we discuss different machine learning and deep learning-related classification framework in detail. The third column of Table 2 bring to notice the target and source dataset of relevant articles. It can be clearly stated that most of the transfer learning- based HAR articles used ImageNet as the source dataset due to large-scale data and class variations. This column also acknowledges the type of operation linked to model training. In some cases, the model is trained on images through a series of 2D operations whereas others pitch about 3D operations on a bunch of sequential frames. The second last column of Table 2 gives the information about the number of classes recognized by different classification models. The last column shows the performances in terms of the accuracy of related HAR articles using transfer learning techniques. By observing the table, we can conclude that the performance is directly proportional to the model's computational efficiency and the relatedness between source and target datasets. But it is inversely proposal to the number of classes or class diversity.

7. Challenges and future research directions

7.1. Transfer learning

Sufficient data availability and outdated data management are two significant issues when dealing with deep learning architecture. Transfer learning becomes very effective for the problems like outdated data compensation, training data recollection, expensive data labeling, and accuracy enhancement. However, the most compelling feature selection criteria and techniques for a successful knowledge transfer are still yet to be explored even in these modern days. Again, powerful transfer learning techniques like ZSL and unsupervised transfer learning need more attention to make transfer learning more effective in the HAR and classification domain. The hidden negotiable relationships between HAR datasets can potentially enhance the performance of the transfer learning-based HAR model. How can a platform be built to promote devalued learning concepts such as relational knowledge transfer?

7.2. Explorable video model

In recent years, there has been much research on interpreting visual models. However, there is a scarcity of research on interpretable video models. Only a few video frames are required for recognizing an action. Furthermore, activities differ in terms of their temporal aspects. The interpretability of complex activities based on keyframes raises some decisive concerns:

1.How are these frames organized in the temporal domain? 2.How do they contribute to the classification task?

3.Whether these frames can be chosen to train the model faster without affecting HAR performance?

7.3. Sensor data modality

The sensor data modality in the HAR dataset can be audio, text, and image. A relation can be established between multimodal data more convincingly than mono- modal data. Recognition of more complex activities has become rich semantic knowledge residing in multimodal data, making the complex activity recognition smoother. When a human remembers something, the former sequence evokes the latter, and so on. Furthermore, understanding the long-term association necessitates an understanding of inter- entity communication. Pre-defined item interactions, for example, occur in a certain activity under specific scene circumstances. As a result, HAR should examine both activities and multimodal data, such as object interpretation, scene interpretation, and activity temporal correlations. The study of multimodal data also supports the prediction of long-duration activities.

7.4. Architecture generalisation

Most of the papers are based on single-mode input data. More than one mode could enhance the performance of the model architecture. How to model the transfer learning architecture so as to perform well on multi-modal source data? How to make the entire architecture robust for online and real-time HAR? It is incredibly challenging to realize the hierarchical framework of complex high-level activities because they encompass deeper semantic and context information. How do we develop an architecture that exploits these high-level activities' co-relation to yield enhanced performance?

7.5. Physical attributes in activities

In deep learning-based HAR, much research has been conducted emphasizing Spatio-temporal features. Only a few of them discuss the depth features present in the image sequences of HAR video data. The

Table 2
Chronological Performance Comparison for human activity recognition.

YEAR/LINK	ARCHITECTURE	TARGET/SOURCE DATASET	ACCURACY(%)
2011/Duan, Xu, Tsang and Luo (2011)	SVM	web video/consumer video	57.9
2011/Liu, Shah, Kuipers and Savarese (2011)	SVM	IXMAS (cross-view)	75.3
2011/Wei and Pal (2011)	RBM	UIUC(cross-action)	82.9
2012/Li, Camps and Sznaier (2012)	SVM	IXMAS (cross-view)	90.57(High)
2013/Rohrbach, Ebert and Schiele (2013)	kNN	Script data/MP-II composite	36.2
2014/Zhu and Shao (2014)	SVM	HMDB51+YouTube/UCF YouTube	91.11
		HMDB51+YouTube/Caltech101	79.02
		HMDB51+YouTube/Caltech256	42.8
		HMDB51+YouTube/Kodak consum	62.6
2014/Yamada, Sigal and Raptis (2013)	kNN	Poser/HUMANEVA-1	–
2014/Bhattacharya et al. (2014)	CNN	Sports1M/UCF-101	65.4
2015/Yue-Hei Ng et al. (2015)	CNN - LSTM + optical flow	Sports 1 M/ImageNet	90.4
	(AlexNet and GoogleNet)	UCF101/ImageNet	88.6
2016/Zhang, Chao, Sha and Grauman (2016)	CNN + key frame + Subshot	Kodak consumer/seed image data	82.3
2016/Wang, Farhadi and Gupta (2016a)	CNN(two stream)	UCF 101/ImageNet	92.4
		HMDB 51/ImageNet	63.4
		ACT/ImageNet	80.6
2016/Wang et al. (2016b)	ConvNet + TSN	HMDB 51/ImageNet	69.4
		UCF101/ImageNet	94.2
2017/Wang, Chen, Hu, Peng and Philip (2018b)	Deep CNN + Stratified TL	OPPORTUNITY/Intra-class	83.96
		PAMAP2/Intra-class	43.47
		UCI DSADS/Intra-class	81.6
2017/Bux Sargano et al. (2017)	AlexNet + SVM/KNN	KTH/ImageNet	98.15
		UCF Sports/ImageNet	91.47
2017/Qiu, Yao and Mei (2017)	Pseudo 3D ResNet	Sports 1 M/ImageNet	87.4(Top5)
		UCF101/ImageNet	93.7(Top3)
		ActivityNet/ImageNet	87.71(Top3)
		ASLAN/ImageNet	80.8
2018/Alwassel et al. (2018)	PCA + LSTM	AVA/ImageNet	91
		THUMOS14/ImageNet	91
2018/Carreira and Zisserman (2017)	I3D(two-stream)	HMDB 51/Kinetics	80.9
		UCF101/Kinetics	98
2018/Wang, Zheng, Chen and Huang (2018c)	USSAR and TNNAR	OPPORTUNITY/Intra-class	87.43
		UCI DSADS/Intra-class	86.76
2018/Ntalampiras and Potamitis (2018)	TL-CHMM	Imbalance audio data/Intra class	94.6
2018/Tran et al. (2018)		kinetics /Seed videos	95.0
		UCF 101/Seed videos	97.3
2019/Ghadiyaram, Tran and Mahajan (2019)	R(2 + 1)D-d	kinetics /(ImageNet1K+IG-Kinetics)	95.3
		something-something	79
		EPIC-Kitchen/(ImageNet1K+IG-Kinetics)	42.7
2019/Korbar, Tran and Torresani (2019)	ir-CSN-152	Sports1M/(ImageNet + AudioNet)	84
2020/An, Bhat, Gumussoy and Ogras (2020)	CNN +TL	UCI HAR/HAPT, UniMiB,WISDM	upto 43%
2021/Coskun et al. (2021)	AMAML	EPIC/EGTEA	60.7(10shots)
2021/Zhu et al. (2021)	PAL(CNN)	ImageNet/Kinetics-100	74.1
		SSV2–100/ImageNet	62.6
		HMDB-51/ImageNet	75.8
		UCF-101/ImageNet	85.3
2021/Sabater et al. (2021)	TCN	NTU RGB+*D*-120/therapy dataset	46.5(1shot)
2021/Ben-Ari, Shpigel Nacson, Azulai, Barzelay and Rotman (2021)	C3D+I3D	Sports1m, ActivityNet V1.2/Kinetics-400	83.12
2021/Perrett et al. (2021)	ResNet-50+Transformer	ImageNet/Kinetics-100	85.9
		SSV2/ImageNet	64.6
		HMDB-51/ImageNet	75.6
		UCF-101/ImageNet	96.1

depth feature carries significant information about activity-related physical factors like distance, movement type, and gait pattern. So, it needs more attention in exploring H–H and H–O interaction level activity. Both dataset and architectural level development are required to successfully comprehend the physical aspect of activities. Apart from these, many other physical aspects, such as acceleration, direction, and movement style, get unnoticed while analyzing complex and interaction-level activity.

7.6. Labeling strategy in har

Some benchmark datasets comprise large numbers of classes with millions of video sequences, such as Kinetics, HAA, and YouTube 8 M. Class labeling is not required for recorded, generated, and crowdsourced dataset as it gets labeled at the time of origination. Accurate and precise labeling followed by thorough verification is imperative for supervised learning schemes. These manually labeling strategies are expensive and sometimes faulty due to human error. Effective crowdsourcing platform comes to the rescue. Amazon Mechanical Turk (AMT) is a widespread annotating strategy for automatic and quick labeling during dataset creation. However, due to lack of generalization, it has its own limitation. So, extra attention should be given to creating more AMT-like annotation strategies. Zero-shot learning is a knowledge transfer approach where activities are classified without any prior training. Besides these, pseudo labeling strategy is sometimes played a pivotal role to annotate large-scale data in one-shot learning, few-shot learning, and semi-supervised learning. Here, we approximate the labels in unannotated data based on the previously annotated data. Pseudo-labeling reduces the overfitting and improves the speed of the model. But this strategy fails to impact when there is not enough labeled data present or the absence of labeled data for a particular class or increment of data does not help the model performance. Robustness and universality are still

debatable research topics for this approach to improve classification accuracy.

7.7. Feature engineering in har

To improve the model performance on unseen data, we need to extract the valuable features from the raw data that better represent the underlying problem to the predictive model. The flexibility, complexity, and performance of a model are profoundly dependent on these extracted and transformed features. Data decomposition and aggregations are two important operations encountered while transforming the raw data. "How do we decompose or aggregate the raw data for a better description of the underlying problem?" is a challenging problem during this extraction process. Apart from this, researchers are still trying to find out the effective solution to some smaller queries like, "How to automate this transformation process?", "How to identify and select the problem-dependent useful feature?", and "what are the manual feature selection criteria?". DT, random forest, regularization variant, principal component analysis, and structural risk minimization are some widely used feature engineering approaches applied in HAR tasks.

7.8. Limited hardware computation

Well-performed HAR models are very hard to implement in real-time due to the constrained computing power (Hardware constrain). As a result, we are forced to compromise either on input data or computationally expensive techniques. And for this, analysts adopt many data reduction techniques like cropping, compression, key-frame extraction, sub-shot, and thresholding. Another method is to embrace those sensing devices that can provide relatively more uncomplicated data forms. Most wearable sensors can be an example of that kind, where we collect mainly the 1-D form of data. Both the compromising techniques lead the model to decline in performance. We conclude with a similar result while adjusting the classification technique. We need a model that comes up with an acceptable trade-off between the computational burden and performance in a constrained computational environment. And to make this viable, our research focus should rest upon a track of informative sensing techniques, efficient descriptor extracting methods, and high-performing model architecture.

7.9. Contextual information gathering

Our model may not be able to recognize high-level behavior or activity properly. For example, our model may fail to recognize "group discussion" activity. Instead, it may be wrongly interpreted as "sitting and talking." Similarly, "running on the road" or "running on the track" can get mixed and misclassified under a more superficial activity, i.e., "running." This shallow activity classification is apparently due to the lack of background knowledge. Relating the semantic features with the logical description between action and behavior through the learning of Natural Language Processing (NLP) may be a possible solution for recognizing these complex activities. This contextual information may provide additional knowledge that helps classify complex activities correctly.

7.10. Negative transfer

In transfer learning, the source domain data representation leverages target domain data for enhancing target domain performance accuracy. However, sometimes, leveraging source domain-specific knowledge reduces the transfer learning performance of the target data. So, we need to keep the knowledge about the origin of negative transfer, factors influencing negative transfer, and tranquilizing algorithms to prevent negative transfer before applying transfer learning to any tasks. Rosenstein, Marx, Kaelbling and Dietterich (2005) introduces negative transfer after finding that the incongruent nature between the source and target data after a certain bound undesirably hamper the performance of the underlying model instead of boosting it. A similar definition is also illustrated in Pan and Yang (2009), which compares the relatedness between the source domain data to the target domain data to introduce the negative transfer while applying transfer learning. The governing factors like domain divergence, transfer learning algorithms, source, and target data quality are needed to be reviewed to rule out the existence of negative transfer in the learning mechanism. Even after the negative transfer is get infused into the transfer model, it can be overcome through many preventive approaches like the secure transfer mechanism, domain similarity estimation algorithm, and distant transfer assessment Zhang, Deng, Zhang and Wu (2020b). So, the negative transfer is a longstanding and formidable concern that needs to be thoroughly reviewed for the vision- based transfer learning model for HAR.

8. Contribution to literature and implication for practice

We have shown an extensive sketch of vision sensor-based HAR using transfer learning. To understand transfer learning, we briefly explain the difference between source and target, domain and task. We also represent five steps followed in HAR: activity types, sensors, transduction, different approaches, and performance measures. We label the vision-based HAR datasets from 2011 to 2021 with detailed specifications. We classify and discuss the learning algorithm in three different categories used for this task. These categories are the generative model, discriminative model, and graph-based model. To the best of our knowledge, we are the first to divide classification techniques for vision-based HAR into three modular representations. We conclude our review by exchanging views on various challenges and future direction. To the best of our knowledge, we are the first to conduct a decade-long review on transfer learning enhanced vision-based HAR, where we discuss related datasets with specifications and three classification formats relevant to our topic. This paper transfer in-depth information about different datasets from 2011 to 2021 and is intended to be managed under various application scenarios. The detailed depiction of classification algorithms under transfer learning scenarios enhances the ideation of researcher for future implementation in this domain.

9. Conclusion

In this extensive survey, we emphasize the idea of using state- of-the-art transfer learning methods that reduce the difficulty and effort behind data collection, data extinction, data labeling, and accuracy enhancement in the action recognition domain. This paper focuses on the vision-based HAR in context- aware applications and empathizes its diversity with transfer learning functionality. This paper's whole-length depiction, investigation, and high points help the researcher achieve in-depth knowledge in vision-based activity recognition using transfer learning techniques.

Apart from transfer learning and all-pervasive applications in vision-based activity recognition, other various orientations still lie down on the floor to investigate and discover for subsequent research such as detection, tracking, design, and classification. This all-inclusive survey is supposed to strengthen further research in activity recognition grassland.

Declaration of Competing Interest

No conflict of interest

References

Abu-El-Haija, Sami, Kothari, Nisarg, Lee, Joonseok, Natsev, Paul, Toderici, George, Varadarajan, Balakrishnan et al. (2016).Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Aggarwal, Alankrita, Mittal, Mamta, & Battineni, Gopi (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights, 1*(1), Article 100004 pages 9.

Akbari, Ali, & Jafari, Roozbeh (2019).Transferring activity recognition models for new wearable sensors with deep generative domain adaptation. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, pages 85–96/

Alam, Mohammad Arif Ul, & Roy, Nirmalya (2017). Unseen activity recognitions: A hierarchical active transfer learning approach. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)* (pp. 436–446). IEEE. pages.

Alsheikh, Mohammad Abu, Selim, Ahmed, Niyato, Dusit, Doyle, Linda, Lin, Shaowei, & Tan, Hwee-Pink (2016). Deep activity recognition models with triaxial accelerometers. *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Al-Sulaiman, Talal (2022). Predicting reactions to anomalies in stock movements using a feed-forward deep learning network. *International Journal of Information Management Data Insights, 2*(1), Article 100071 pages 11.

Alwassel, Humam, Heilbron, Fabian Caba, & Ghanem, Bernard (2018).Action search: Spotting actions in videos and its application to temporal action localization. In Proceedings of the European Conference on Computer Vision (ECCV), pages 251–266.

An, Sizhe, Bhat, Ganapati, Gumussoy, Suat, & Ogras, Umit (2020).Transfer learning for human activity recognition using representational analysis of neural networks. arXiv preprint arXiv:2012.04479.

Anand, Kartik, Urolagin, Siddhaling, & Mishra, Ram Krishn (2021). How does hand gestures in videos impact social media engagement-insights based on deep learning? *International Journal of Information Management Data Insights, 1*(2), Article 100036.

Arif Ul Alam, Mohammad, Mahmudur Rahman, Md, & Widberg, Jared Q. (2021). Palmar: Towards adaptive multi-inhabitant activity recognition in point- cloud technology. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications* (pp. 1–10). IEEE. pages.

Aslam, Nazia, & Kolekar, Maheshkumar H. (2022). Unsupervised anomalous event detection in videos using spatio-temporal inter-fused autoencoder. *Multimedia Tools and Applications*, 1–26 pages.

Aslam, Nazia, Rai, Prateek Kumar, & Kolekar, Maheshkumar H. (2022). A3N: Attention-based adversarial autoencoder network for detecting anomalies in video sequence. *Journal of Visual Communication and Image Representation, 87*, Article 103598 pages 15.

Ben-Ari, Rami, Shpigel Nacson, Mor, Azulai, Ophir, Barzelay, Udi, & Rotman, Daniel (2021).Taen: Temporal aware embedding network for few-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2786–2794.

Bhattacharya, Sourav, Nurmi, Petteri, Hammerla, Nils, & Plötz, Thomas (2014). Using unlabeled data in a sparse-coding framework for human activity recognition. *Pervasive and Mobile Computing, 15*, 242–262.

Bux Sargano, Allah, Wang, Xiaofeng, Angelov, Plamen, & Habib, Zulfiqar (2017). Human action recognition using transfer learning with deep representations. In *2017 International joint conference on neural networks (IJCNN)* (pp. 463–469). IEEE. pages.

Cabrera, Maria E., Sanchez-Tamayo, Natalia, Voyles, Richard, & Wachs, Juan P. (2017). One-shot gesture recognition: One step towards adaptive learning. In *2017* 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 784–789). IEEE. pages.

Cabrera, Maria Eugenia, & Wachs, Juan Pablo (2017). A human-centered approach to one-shot gesture learning. *Frontiers in Robotics and AI, 4*(8) pages 18.

Careaga, Chris, Hutchinson, Brian, Hodas, Nathan, & Phillips, Lawrence (2019).Metric-based few-shot learning for video action recognition. *arXiv preprint arXiv:1909.09602*.

Carreira, Joao, Noland, Eric, Banki-Horvath, Andras, Hillier, Chloe, & Zisserman, Andrew (2018).A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Carreira, Joao, Noland, Eric, Hillier, Chloe, & Zisserman, Andrew (2019).A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

Carreira, Joao, & Zisserman, Andrew (2017).Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308.

Chatterjee, Subhamoy, Bhandari, Piyush, & Kolekar, MaheshKumar H. (2016). A novel krawtchouk moment zonal feature descriptor for user-independent static hand gesture recognition. In *2016 IEEE Region 10 Conference (TENCON)* (pp. 387–392). IEEE. pages.

Chen, Chen, Jafari, Roozbeh, & Kehtarnavaz, Nasser (2015). 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)* (pp. 168–172). IEEE. pages.

Chen, Yiqiang, Wang, Jindong, Huang, Meiyu, & Yu, Han (2019). Cross-position activity recognition with stratified transfer learning. *Pervasive and Mobile Computing, 57*, 1–13.

Cheng, Yi-Bin, Chen, Xipeng, Chen, Junhong, Wei, Pengxu, Zhang, Dongyu, & Lin, Liang (2021). Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). IEEE. pages.

Chung, Jihoon, Wuu, Cheng-hsin, Yang, Hsuan-ru, Tai, Yu-Wing, & Tang, Chi-Keung (2021). HAA500: Human-centric atomic action dataset with curated videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13465–13474). pages.

Cook, Diane, Feuz, Kyle D., & Krishnan, Narayanan C. (2013). *Transfer learning for activity recognition: A survey. Knowledge and information systems, 36*(3), 537–556.

Coskun, Huseyin, Zia, M. Z. eeshan, Tekin, Bugra, Bogo, Federica, Navab, Nassir, Tombari, Federico, et al., (2021). Domain-specific priors and meta learning for few-shot first-person action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pages 14.

Deng, Wan-Yu, Zheng, Qing-Hua, & Wang, Zhong-Min (2014). Cross-person activity recognition using reduced kernel extreme learning machine. *Neural Networks, 53*, 1–7.

Du, Hao, He, Yuan, & Jin, Tian (2018). Transfer learning for human activities classification using micro-doppler spectrograms. In *2018 IEEE International Conference on Computational Electromagnetics* (pp. 1–3). IEEE. pages.

Duan, Lixin, Xu, Dong, Tsang, Ivor Wai-Hung, & Luo, Jiebo (2011). Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(9), 1667–1680.

Fu, Biying, Damer, Naser, Kirchbuchner, Florian, & Kuijper, Arjan (2021). Generalization of fitness exercise recognition from doppler measurements by domain-adaption and few-shot learning. In *International Conference on Pattern Recognition* (pp. 203–218). Springer. pages.

Gan, Chuang, Lin, Ming, Yang, Yi, De Melo, Gerard, & Hauptmann, Alexander G. (2016).Concepts not alone: Exploring pairwise relationships for zero- shot video activity recognition. In *Thirtieth AAAI conference on artificial intelligence*.

Gang, Zhao, Wenjuan, Zhu, Biling, Hu, Jie, Chu, Hui, He, & Qing, Xia (2021). A simple teacher behavior recognition method for massive teaching videos based on teacher set. *Applied Intelligence, 51*(12), 8828–8849.

Ghadiyaram, Deepti, Tran, Du, & Mahajan, Dhruv (2019).Large-scale weakly- supervised pre-training for video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12046–12055.

Ghosal, Deepanway, & Kolekar, Maheshkumar H. (2018).Music genre recognition using deep neural networks and transfer learning. In *Interspeech*, pages 2087–2091.

Gonegandla, Pranesh, & Kolekar, Maheshkumar H. (2022). Automatic song indexing by predicting listener's emotion using EEG correlates and multi-neural networks. *Multimedia Tools and Applications, 81*, 1–11 pages.

Gorban, A., Idrees, H., Jiang, Y.-.G., Roshan Zamir, A., Laptev, I., Shah, M. et al. (2015).THUMOS challenge: Action recognition with a large number of classes. http://www.thumos.info/.

Goyal, Raghav, Kahou, Samira Ebrahimi, Michalski, Vincent, Materzynska, Joanna, Westphal, Susanne, Kim, Heuna et al. et al. (2017).The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850.

Gu, Chunhui, Sun, Chen, Ross, David A., Vondrick, Carl, Pantofaru, Caroline, Li, Yeqing et al.,(2018a). et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056.

Gupta, Saurabh (2021). Deep learning based human activity recognition (HAR) using wearable sensor data. *International Journal of Information Management Data Insights, 1*(2), Article 100046 pages 18.

Hadfield, Simon, & Bowden, Richard (2013).Hollywood 3D: Recognizing actions in 3D natural scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3398–3405.

Haresamudram, Harish, Beedu, Apoorva, Agrawal, Varun, Grady, Patrick L., Essa, Irfan, Hoffman, Judy et al. (2020).Masked reconstruction based self- supervision for human activity recognition. In *Proceedings of the 2020 International Symposium on Wearable Computers*, pages 45–49.

Heilbron, Fabian Caba, Escorcia, Victor, Ghanem, Bernard, & Niebles, Juan Carlos (2015).Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Imran, Javed, & Raman, Balasubramanian (2020). Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing, 11*(1), 189–208.

Jänicke, Martin, Tomforde, Sven, & Sick, Bernhard (2016a). Towards self-improving activity recognition systems based on probabilistic, generative models. In *2016 IEEE International Conference on Autonomic Computing (ICAC)* (pp. 285–291). IEEE. pages.

Jänicke, Martin, Tomforde, Sven, & Sick, Bernhard (2016b).Towards self-improving activity recognition systems based on probabilistic, generative models. In *2016 IEEE International Conference on Autonomic Computing*, pages 285–291. IEEE.

Jasani, Bhavan, & Mazagonwalla, Afshaan (2019).Skeleton based zero shot action recognition in joint pose-language semantic space. *arXiv preprint arXiv:1911.11344*.

Ji, Zhong, Liu, Xiyao, Pang, Yanwei, & Li, Xuelong (2020). SGAP-Net: Semantic- guided attentive prototypes network for few-shot human-object interaction recognition. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 11085–11092 pages.

Karn, Nabin Kumar, & Jiang, Feng (2016). Improved gloh approach for one-shot learning human gesture recognition. In *Chinese Conference on Biometric Recognition* (pp. 441–452). Springer. pages.

Karpathy, Andrej, Toderici, George, Shetty, Sanketh, Leung, Thomas, Sukthankar, Rahul, & Fei-Fei, Li (2014a).Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Karpathy, Andrej, Toderici, George, Shetty, Sanketh, Leung, Thomas, Sukthankar, Rahul, & Fei-Fei, Li (2014b).Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Kay, Will, Carreira, Joao, Simonyan, Karen, Zhang, Brian, Hillier, Chloe, Vijayanarasimhan, Sudheendra et al. et al. (2017a).The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kay, Will, Carreira, Joao, Simonyan, Karen, Zhang, Brian, Hillier, Chloe, Vijayanarasimhan, Sudheendra et al. et al. (2017b).The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Khan, Md Abdullah Al Hafiz, & Roy, Nirmalya (2018). Untran: Recognizing unseen activities with unlabeled data using transfer learning. In 2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI) (pp. 37–47). IEEE. pages.

Kolekar, Maheshkumar H. (2011). Bayesian belief network based broadcast sports video indexing. *Multimedia Tools and Applications, 54*(1), 27–54.

Kolekar, Maheshkumar H., & Sengupta, Somnath (2015). Bayesian network-based customized highlight generation for broadcast soccer videos. *IEEE Transactions on Broadcasting, 61*(2), 195–209.

Korbar, Bruno, Tran, Du, & Torresani, Lorenzo (2019).Scsampler: Sampling salient clips from video for efficient action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6232–6242.

Lang, Yue, Wang, Qing, Yang, Yang, Hou, Chunping, Huang, Danyang, & Xiang, Wei (2018). Unsupervised domain adaptation for micro-doppler human motion classification via feature fusion. *IEEE Geoscience and Remote Sensing Letters, 16*(3), 392–396.

Li, Binlong, Camps, Octavia I., & Sznaier, Mario (2012). Cross-view activity recognition using hankelets. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1362–1369). IEEE. pages.

Li, Lianwei, Qin, Shiyin, Lu, Zhi, Zhang, Dinghao, Xu, Kuanhong, & Hu, Zhongying (2021). Real-time one-shot learning gesture recognition based on lightweight 3D inception-ResNet with separable convolutions. *Pattern Analysis and Applications, 24*, 1–20 pages.

Liu, Chunhui, Hu, Yueyu, Li, Yanghao, Song, Sijie, & Liu, Jiaying (2017).PKU MMD: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*.

Liu, Jingen, Shah, Mubarak, Kuipers, Benjamin, & Savarese, Silvio (2011). Cross-view action recognition via view knowledge transfer. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3209–3216). IEEE. pages.

Liu, Jun, Shahroudy, Amir, Perez, Mauricio, Wang, Gang, Duan, Ling-Yu, & Kot, Alex C. (2019). NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence, 42*(10), 2684–2701.

Liu, Wu, Mei, Tao, Zhang, Yongdong, Che, Cherry, & Luo, Jiebo (2015).Multi- task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3707–3715.

Loey, Mohamed, Manogaran, Gunasekaran, Taha, Mohamed Hamed N., & Khalifa, Nour Eldeen M. (2021). A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic. *Measurement, 167*, Article 108288 pages 11.

Luo, Manman, & Mu, Xiangming (2022). Entity sentiment analysis in the news: A case study based on negative sentiment smoothing model (NSSM). *International Journal of Information Management Data Insights, 2*(1), Article 100060.

Ma, Chunyong, Zhang, Shengsheng, Wang, Anni, Qi, Yongyang, & Chen, Ge (2020). Skeleton-based dynamic hand gesture recognition using an enhanced network with one-shot learning. *Applied Sciences, 10*(11), 3680.

Mishra, Ashish, Pandey, Anubha, & Murthy, Hema A. (2020). Zero-shot learning for action recognition using synthesized features. *Neurocomputing, 390*, 117–130.

Mohsen Amiri, S., Pourazad, Mahsa T., Nasiopoulos, Panos, & Leung, Victor C. M. (2013). Non-intrusive human activity monitoring in a smart home environment. In *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)* (pp. 606–610). IEEE. pages.

Mutegeki, Ronald, & Han, Dong Seog (2019). Feature-representation transfer learning for human activity recognition. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)* (pp. 18–20). IEEE. pages.

Ng, Joe Yue-Hei, Hausknecht, Matthew, Vijayanarasimhan, Sudheendra, Vinyals, Oriol, Monga, Rajat, & Toderici, George (2015).Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702.

Ntalampiras, Stavros, & Potamitis, Ilyas (2018). Transfer learning for improved audio-based human activity recognition. *Biosensors, 8*(3), 60 pages 12.

Pan, Sinno Jialin, & Yang, Qiang (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345–1359.

Parmar, Paritosh, & Morris, Brendan (2022).Win-Fail action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 161–171.

Perera, Asanka G., Law, Yee Wei, Ogunwa, Titilayo T., & Chahl, Javaan (2020). A multiviewpoint outdoor dataset for human action recognition. *IEEE Transactions on Human-Machine Systems, 50*(5), 405–413.

Perrett, Toby, Masullo, Alessandro, Burghardt, Tilo, Mirmehdi, Majid, & Damen, Dima (2021).Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 475–484.

Piergiovanni, A.J., .& Ryoo, Michael S. (2018).Fine-grained activity recognition in baseball videos. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops*, pages 1740–1748.

Qin, Xin, Chen, Yiqiang, Wang, Jindong, & Yu, Chaohui (2019). Cross-dataset activity recognition via adaptive spatial-temporal transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3*(4), 1–25.

Qiu, Zhaofan, Yao, Ting, & Mei, Tao (2017).Learning spatio-temporal representation with pseudo-3d residual networks. In proceedings of the IEEE International Conference on Computer Vision, pages 5533–5541.

Rahmani, Hossein, & Mian, Ajmal (2015).Learning a non-linear knowledge transfer model for cross-view action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2458–2466.

Roder, Mateus, Almeida, Jurandy, Rosa, Gustavo H. D. e, Passos, Leandro A., Rossi, André L. D., & Papa, João P. (2021). From actions to events: A transfer learning approach using improved deep belief networks. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 01–08). IEEE. pages.

Rodriguez, Mario, Orrite, Carlos, Medrano, Carlos, & Makris, Dimitrios (2017a).Fast simplex-hmm for one-shot learning activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 41–48.

Rodriguez, Mario, Orrite, Carlos, Medrano, Carlos, & Makris, Dimitrios (2017b).Fast simplex-hmm for one-shot learning activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 41–48.

Rohrbach, Marcus, Amin, Sikandar, Andriluka, Mykhaylo, & Schiele, Bernt (2012). A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1194–1201). IEEE. pages.

Rohrbach, Marcus, Ebert, Sandra, & Schiele, Bernt (2013). Transfer learning in a transductive setting. *Advances in neural information processing systems, 26*.

Rosenstein, Michael T., Marx, Zvika, Kaelbling, Leslie Pack, & Dietterich, Thomas G. (2005).To transfer or not to transfer. In In NIPS'05 Workshop, Inductive Transfer: 10 Years Later.

Sabater, Alberto, Santos, Laura, Santos-Victor, Jose, Bernardino, Alexandre, Montesano, Luis, & Murillo, Ana C. (2021).One-shot action recognition in challenging therapy scenarios. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2777–2785.

Sanabria, Andrea Rosales, & Ye, Juan (2020). Unsupervised domain adaptation for activity recognition across heterogeneous datasets. *Pervasive and Mobile Computing, 64*, Article 101147.

Shahroudy, Amir, Liu, Jun, Ng, Tian-Tsong, & Wang, Gang (2016).NTU RGB+D: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019.

Sharma, Vijeta, Gupta, Manjari, Kumar, Ajai, & Mishra, Deepti (2021). EduNet: A new video dataset for understanding human activity in the classroom environment. *Sensors, 21*(17), 5699.

Shi, Zhenguo, Zhang, Jian Andrew, Xu, Yi Da Richard, & Cheng, Qingqing (2020). Environment-robust device-free human activity recognition with channel- state-information enhancement and one-shot learning. *IEEE Transactions on Mobile Computing*.

Soomro, Khurram, Zamir, Amir Roshan, & Shah, Mubarak (2012).Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tran, Du, Bourdev, Lubomir, Fergus, Rob, Torresani, Lorenzo, & Paluri, Manohar (2015).Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Tran, Du, Wang, Heng, Torresani, Lorenzo, Ray, Jamie, LeCun, Yann, & Paluri, Manohar (2018).A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 6450–6459.

Tricco, Andrea C., Lillie, Erin, Zarin, Wasifa, O'Brien, Kelly K., Colquhoun, Heather, Levac, Danielle, et al., (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of internal medicine, 169*(7), 467–473.

Vondrick, Carl, Pirsiavash, Hamed, & Torralba, Antonio (2016). Generating videos with scene dynamics. *Advances in neural information processing systems, 29*, 613–621.

Wang, Jiang, Nie, Xiaohan, Xia, Yin, Wu, Ying, & Zhu, Song-Chun (2014).Cross- view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656.

Wang, Jindong, Chen, Yiqiang, Hu, Lisha, Peng, Xiaohui, & Philip, S. Y. u (2018a). Stratified transfer learning for cross-domain activity recognition. In *2018 IEEE international conference on pervasive computing and communications (PerCom)* (pp. 1–10). IEEE. pages.

Wang, Jindong, Chen, Yiqiang, Hu, Lisha, Peng, Xiaohui, & Philip, S. Y. u (2018b). Stratified transfer learning for cross-domain activity recognition. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (pp. 1–10). IEEE. pages.

Wang, Jindong, Zheng, Vincent W., Chen, Yiqiang, & Huang, Meiyu (2018c).Deep transfer learning for cross-domain activity recognition. In proceedings of the 3rd International Conference on Crowd Science and Engineering, pages 1–8.

Wang, Limin, Xiong, Yuanjun, Wang, Zhe, Qiao, Yu, Lin, Dahua, Tang, Xiaoou, et al., (2016b). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20–36). Springer. pages.

Wang, Xiaolong, Farhadi, Ali, & Gupta, Abhinav (2016a).Actions˜ transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2658–2667.

Wei, Bin, & Pal, Christopher (2011). Heterogeneous transfer learning with rbms. In *Proceedings of the AAAI Conference on Artificial Intelligence: 25* (pp. 531–536). pages.

Wen, Jiahui, & Zhong, Mingyang (2015). Activity discovering and modelling with labelled and unlabelled data in smart environments. *Expert Systems with Applications, 42*(14), 5800–5810.

Xing, Yang, Lv, Chen, Wang, Huaji, Cao, Dongpu, Velenis, Efstathios, & Wang, Fei- Yue (2019). Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE transactions on Vehicular Technology, 68*(6), 5379–5390.

Xing, Yang, Tang, Jianlin, Liu, Hong, Lv, Chen, Cao, Dongpu, Velenis, Efstathios, et al., (2018). End-to-end driving activities and secondary tasks recognition using deep convolutional neural network and transfer learning. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1626–1631). IEEE. pages.

Xu, Xun, Hospedales, Timothy M., & Gong, Shaogang (2016). Multi-task zero- shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision* (pp. 343–359). Springer. pages.

Yamada, Makoto, Sigal, Leonid, & Raptis, Michalis (2013). Covariate shift adaptation for discriminative 3d pose estimation. *IEEE transactions on pattern analysis and machine intelligence, 36*(2), 235–247.

Yan, Yan, Liao, Tianzheng, Zhao, Jinjin, Wang, Jiahong, Ma, Liang, Lv, Wei et al. (2022).Deep transfer learning with graph neural network for sensor-based human activity recognition. *arXiv preprint arXiv:2203.07910*.

Zaher Md Faridee, Abu, Chakma, Avijoy, Misra, Archan, & Roy, Nirmalya (2022). *STran-GAN: Adversarially-learnt spatial transformer for scalable human activity recognition*: 23. Smart Health.

Zhang, Hongguang, Zhang, Li, Qi, Xiaojuan, Li, Hongdong, Torr, Philip H. S., & Koniusz, Piotr (2020a). Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference* (pp. 525–542). Glasgow, UK: Springer. *August 23–28, 2020Proceedings, Part V 16*pages.

Zhang, Jing, Li, Wanqing, & Ogunbona, Philip (2017a).Joint geometrical and statistical alignment for visual domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1859–1867.

Zhang, Ke, Chao, Wei-Lun, Sha, Fei, & Grauman, Kristen (2016).Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1059–1067.

Zhang, Lei, Zhang, Shengping, Jiang, Feng, Qi, Yuankai, Zhang, Jun, Guo, Yuliang, et al., (2017b). Bomw: Bag of manifold words for one-shot learning gesture recognition from kinect. *IEEE Transactions on Circuits and Systems for Video Technology, 28*(10), 2562–2573.

Zhang, Wen, Deng, Lingfei, Zhang, Lei, & Wu, Dongrui (2020b).A survey on negative transfer. arXiv preprint arXiv:2009.00909.

Zhou, Luowei, Xu, Chenliang, & Corso, Jason J. (2018).Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhu, Fan, & Shao, Ling (2014). Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision, 109*(1–2), 42–59.

Zhu, Xiatian, Toisoul, Antoine, Perez-Rua, Juan-Manuel, Zhang, Li, Martinez, Brais, & Xiang, Tao (2021).Few-shot action recognition with prototype- centered attentive learning. arXiv preprint arXiv:2101.08085.

Zhu, Yi, & Newsam, Shawn (2017). Efficient action detection in untrimmed videos via multi-task learning. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 197–206). IEEE. pages.