



Review on recent Computer Vision Methods for Human Action Recognition

Azhee Wria Muhamada^a, Aree A. Mohammed^b

^a Computer Science Department, College of Basic Education, University of Sulaimani, Iraq

^b Faculty of Science, Computer Department, University of Sulaimani, Kurdistan Region, Iraq
azhee.muhamad@univsul.edu.iq, aree.ali@univsul.edu.iq

KEYWORDS

Action Recognition techniques;
Deep Learning;
Video datasets;
Image datasets;
Accuracy;
Prediction;
Object;
RNN; CNN;
Classification.

ABSTRACT

Human action recognition has been an important goal of computer vision ever since its starting and has developed considerably within the last years. The recognition of human activities is sometimes thought of to be a straightforward method. Issues occur in advanced scenes involving high velocities. Activity prediction mistreatment of artificial intelligence (AI) by numerical analysis has attracted the eye of many academics. To modify the comparison of these ways, several datasets concerning tagged act created, having nice variation in content and methodology Human activities are a significant challenge in varied fields. There are several friendly applications during this space, as well as sensible homes, valuable artificial intelligence, human-computer interactions, and enhancements in protection in many areas like security, transport, education, and medication through the management of falling or aiding in medication consumption for older people. The advanced improvement and success of deep learning techniques in various pc vision applications encourage the utilization of those ways in the video process. The human presence is a fundamental challenge within the analysis of human behavior through activity. An individual in a more than video sequence may be represented by their motion, skeleton, and abstraction characteristics. This work aims to boost human presentation by gathering various options and, therefore, exploiting the new RNN structure for activities. Throughout this review, the paper focuses on recent advancements within the field of action recognition supported Machin learning. We have compared some of the triumphant human action recognition methodologies to accuracy and prediction along within the review paper.



1. Introduction

In video processing and action classification, human activity recognition can be challenging but crucial. It is being established as part of continuous monitoring of human activity. It has also been proposed for a variety of uses, including health care and older police work, sports injury prevention, position estimation, and residential surveillance. Although significant improvements have been witnessed in human action recognition from video sequences, it is still a weak flaw for several causes, including shifts in perspective and closeness to a camera, nature of the context, and speed diversity. The most challenging portion of the procedure is finding good options. It does affect the algorithmic program's efficiency by decreasing the time and quality of calculations. Nevertheless, handcrafted native options from RGB video captured by second cameras that are incapable of handling complicated actions are provided by the most common techniques of human action recognition (Rodríguez-Moreno et al., 2019). Background extraction is used in several methods to detect a moving individual. Multiple strategies, along with Gaussian distribution (Ijjina et al., 2016), include relating human behavior through various techniques that use motion tracking. To monitor its movement and create trajectories in all sequences, human monitoring is performed. For humans, following is usually an easy process. Fast-moving objects make the situation more complicated. This inspires researchers to use computer vision techniques like Optical Flow, Scale Invariant Feature Transformation (SIFT), a bar chart of familiarized Gradient (HOG), and Mean Shift to address motion following problems. These methods recognize behavior, with the help of the object's appearance and activity of chassis components from RGB video (Jaouedi et al., 2020). Gao et al., 2018 created an associate degree application that promoted act observation for attention. This software keeps track of how patients or the elderly are doing remotely. Medical scientists can advise patients on diet, exercise, and medication adherence by analyzing alterations in everyday human activity such as walking, working, exercising, preparing food, or sleep activity. This is especially important for senior citizens, as such systems allow them to measure reception for a more extended, healthier, and safer approach. The beneficial robotic field is an equally essential part of the act (Gao et al., 2018).

2. Scope and organization

In the last twenty years, many new methods developed in human action recognition (Pham et al., 2015). This was possible due to the proliferation of new algorithms that utilized machine learning. Therefore, it is encouraging to examine the developments in this area, and as such, this paper focuses on researches that use deep learning and human behavior's popularity. The primary aim here is to review the studies in this field. Also, comparing the effectiveness of the approaches that rely on deep learning to other similar works is conducted to identify the pros and cons of each. We use taxonomy, as shown in figure 1, to make the study more available. We also work to distinguishing the methods used for deep learning with regards to action recognition, based on architecture, for instance. Convolutional Neural Networks (CNNs), continuous Neural Networks with Long Short-term Memory (RNN-LSTMs), and other significant models are included. There will also be a review of specific combination architectures.



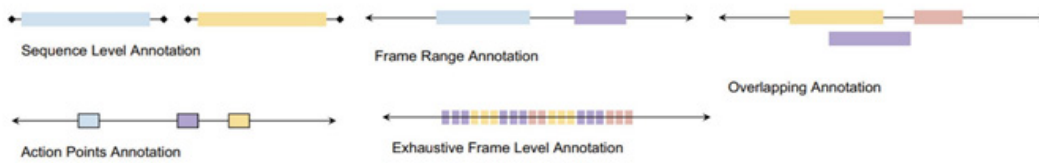


Figure 1: The Different kinds of annotations (Singh et al., 2019)

Image searching, image auto-annotation, understanding of scenes, and object tracking are only a few applications that use it. A prominent subject in computer vision was the tracking of moving objects in video image sequences. Object recognition can be considered a broad term that refers to related tasks in computer vision that include identifying objects in digital images. Image classification, on the other hand, entails guessing the single object's type inside pictures. Object localization is the method of defining the position of an object or more and setting a large box around them inside an image. Object detection integrates these two tasks by localizing and classifying objects inside an image as shown in figure 2.

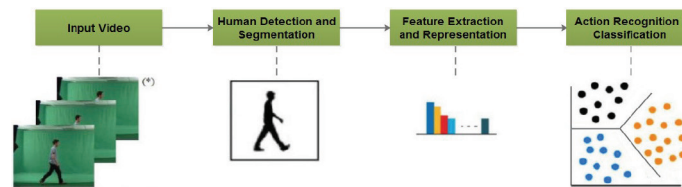


Figure 2: Flow diagram for a typical action recognition system (Ohnishi et al., 2016)

3. Methodology

Intelligent strategies for identifying human activities victimization are often assisted in two main stages: feature extraction and action classification. The first has significant pieces of information for explaining human behaviour; visual, for example, pixel strength or texture, or temporal, such as motion path or flight route. Several scholars have used spatial or visual options to obtain the associated efficient vector of human action data. AlbuSlava 2016 (Albu et al., 2016) and Majed Latah 2017 (Latah et al., 2017) had a technique for predicting client behavior and human actions that relied exclusively on spatial and discourse features. They did, in fact, use the implementation of Convolutional Neural Network (CNN) and 3D-out-put CNN's for extracting spatial information in order to classify both human emotion and actions. According to different scholars, human acts are often more outlined in victimization motion and speed. Worn sensors and Long Short-Term Memory (LSTM) perennial Neural Networks were utilized by Murad and Ryun 2017 (Murad & Ryun, 2017) and Zhen Qina, Yibo Zhang (Qin, Z., Zhang et al., 2020) to describe human motion.

The last is made up of a rotating technique and a measuring instrument. Ning Zhang & Zeyuan Hu (Zhang et al., 2017) relied entirely on temporal options to help the world human silhouette's native optical flow for human activity recognition. Achieving high performance, these strategies remain

effective in complex scenes with varying context, scale, and texture. Such factors result in making the popularity of human activities challenging. A combination of temporal and spatial choices for enhancing the results of classifying human activities were designed. Nicolas Ballas, Li Yao, Chris Pal (Ballas et al., 2015) used a particular model to learn about Spatio-temporal victimization options of Perennial Convolution Networks with Gated Perennial Units (GRU) (RCN). The model is based on the VGG-16 on the associate degree ImageNet transfer learning Model. The recurrent Convolution Neural Networks (RCNN) model for human activity recognition victimization GoogleLeNet architecture was designed by Zhenqi Xu, Jiani Hu (Xu et al., 2016) and Baldominos (Baldominos et al., 2018).

The authors used CNN and RNN for extracting temporal and spatial options in this study. They then mixed options and measured video classification accuracy using the GoogleLeNet design. Zhang, Feng (Zhang et al., 2016) discussed deep-learned Spatio-temporal options by discovering motion de- sripters using a vector of aggregated descriptors and predicting motion descriptors using SIFT geo- metric information.

In order to extract discourse features, an associate degree freelance mathematical space Analysis (ISA) was utilized. Other methods were used by Rui Zhao; Haider Ali (Zhao et al., 2017) to derive human options by using a 3D-CNN architecture and combined the RNN with GRU hidden units. These strategies have been influential in determining the Spatio-temporal options of specific actions that do not necessitate the presenting human body in full. Diego, Cristiano (Faria et al., 2014) used a Bayesian model with skeleton features to predict human actions. By using probabilities weights to offset as posterior probabilities, this model aims to have several classifier changes into one. Using Hidden Markov Models (HMM), Hema Swetha Koppula, Rudhir Gupta (Koppula et al., 2013) sculpt the skeleton of a person. The objects and activities are represented by nodes, while the associations between an object and its activities are represented by edges. The study used the Support Vector Machines (SVM) technique to classify the latter. Bingbing Ni; Yong Pei (Ni et al., 2013) invented an original extraction approach that merged grayscale and depth frames for extracting 3D abstraction and temporal descriptors of human activities. In order to eliminate inaccurate grayscale detection of humans, a depth filter is utilized. Deploying a 3D skeleton and (LOM) native Occupation Model, Jiang, Zicheng (Wang et al., 2013) planned a new attribute for learning human actions, thus eliminating the dependency on 3D joints.

The actions detailed in this study were determined by moving human joints, e.g., for an individual drink, only the hand joint would be extracted. Additionally, Shan and Akella (Shan et al., 2014) and Samuele, Ennio (Cippitelli et al., 2016) both worked towards detecting human skeleton and key poses through RGB-D and K.E respectively that details actions in an extremely large area to achieve the same goal. The SVM technique uses the 3D presentation of human joints for predicting and verifying human actions.

To anticipate human activities, Gaglio and Morana (Gaglio et al., 2014) merged three machine learning methods; the K-means method for detecting a person's skeleton pose, the SVM method for classifying, and the Secret Andre Markoff Models (HMM) method for modeling behavior. Alessandro (Manzi et al., 2017) used RGB-D sensors to determine human skeleton options, the K-means technique for determining posture, and sequential lowest optimization for coaching knowledge. The design goal was to show that a limited number of critical poses are adequate to clarify and acknowledge a person's behaviour. Srijanet (Das et al., 2018), Cruz (Cruz-Silva et al., 2019) and Khaire (Khaire et al., 2018) designed and created a technique that supported skeleton and discourse function extraction in tandem. The RGB-D device, as well as the CNN and LSTM models were used to extract skeleton options. Using the CNN model, the discourse options were discovered.

Yanli (Ji, Y., et al., 2018) proposed a View-guided Skeleton-CNN (VS-CNN) for whimsical read and recognizing human behaviour that continues from weak read variations by envisioning the sequences of the skeleton while also covering a more comprehensive array of reading angles. Wang and Ogunbona (Wang et al., 2018) used the DenseNet CNN model for classifying actions and used an associate degree action recognition model to facilitate the transition of the skeleton to an abstraction presentation by converting the distance values of two joints to plot points. We finalize this section by presenting the main findings from the two surveys conducted by Suriya and Chetan (Singh et al., 2017) and Sigurdsson (Sigurdsson et al., 2018). According to these researchers, strategies focused on human cause estimating and extracting skeleton features will realize more excellent classification rates. The advanced methods are shown in table 1.

Table 1: Modern methods and their explanation (Jaouedi et al., 2020)

Authors	Years	Methods	Interpretation
Gaglio et al., 2015	2015	Kmeans + HMM + SVM	Skeleton features
Nicolas et al., 2016	2016	GRU + RCN	Spatio-temporal features
Xu et al., 2016 and Baldominos et al.	2016	RCNN	Spatio-temporal features
Zhang et al., 2016	2016	Vectors of locally aggregated descriptors, SIFT and ISA	Spatio-temporal features
Shan and Akella 2016 and Enea et al., 2016	2016	Pose Kinetic Energy + SVM	Skeleton features
AlbuSlava 2016	2016	3D CNN	Spatial features
and Majed Latah 2017	2017	3D CNN	Spatial features
Murad and Ryun 2017 and Qin et al.	2017	Deep recurrent neural networks and multimodal sensors	Motion features
Ning et al., 2017	2017	Local optical flow of a global human silhouette	Motion features
Manzi et al., 2017	2017	Kmeans + Sequential Minimal Optimization	Skeleton features
Srijan et al., 2018 , Cruz et al. and Khaire et al.	2018	RGB-D + CNN + LSTM model	Skeleton and contextual features
Yanli et al., 2018	2018	VS-CNN	Skeleton and contextual features
Hug et al., 2019	2019	The conversion of the distance value of two joints to colors points + CNN	Skeleton and contextual features

Extracting related options from human appearance in video sequences is one of the most current methods for human activities. This is a two-part system: the main part is concerned with the development of a replacement classification model, which is supported by the second human skeleton. The in-home activities from the CAD-60 data have been evaluated by this model's results. The second

part is the recognition of human activity over time in a continuous manner. This part uses three types of approaches to show human activities: visual, temporal, and a second human skeleton. as shown in figure 3, an outline of the combination of two different systems for classifying activities and action recognition (Jaouedi et al., 2020).

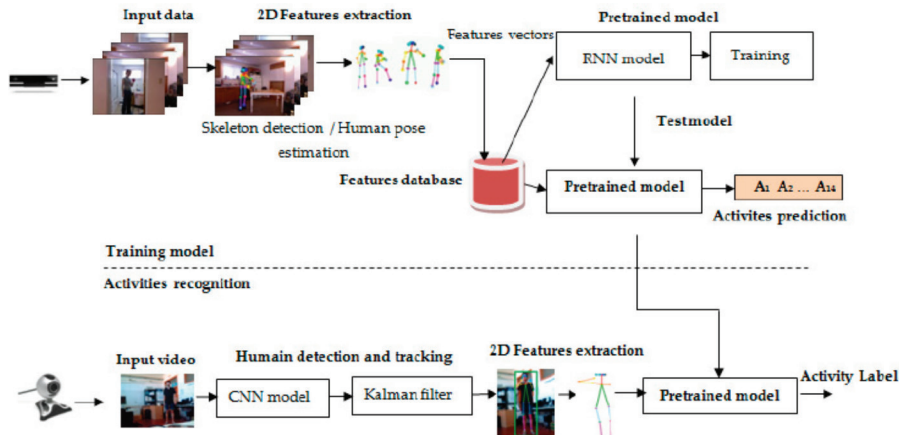


Figure 3: The mixed model for human action recognition (Jaouedi et al., 2020)

4. Evaluation

In universal, a Human Action Recognition algorithm can have two elementary tasks with different evaluation procedures. In the action distribution task, the action is executed in a particular scene is recognized. A Human Action Recognition thus becomes a multiclass classification problematic. If the sequence level ground truth is available, utmost datasets use multiclass accuracy as the metric. Although realistic datasets are unbalanced and long-tailed, some alternative measurements such as precision, recall, and F score are used instead. Specific datasets too permit for covering activity labels both in space and in time. Essentially, the presence of an empty class in activity recognition in datasets like (Idrees et al., 2017). Transforms the problem to a label binary-level compilation task in contrast to the forced-choice multiclass task. Within the activity expectation task. For action segment annotation, mean Average Precision (mAP) calculated over the Intersection over Union (IoU) is used as prescribed in (Liu et al., 2017) (Bojanowski et al., 2014).

A commonly of the measured and issues in this task can be detected in (Minnen et al., 2006). Overall tasks include Spatio-temporal localization, which involves predicting the extent of a task in both bounding boxes and time (Szegedy et al., 2015). The additional task is providing localized event labeling (Krishna et al., 2017). ASLAN (Klipper-Gross et al., 2011) defines an action similarity-labeling task. Generally, benchmark datasets generally have been found to have a robust built-in bias (Torralba et al., 2011), which can cause the study to become constrained. Better evaluation protocols such as cross-dataset testing (Cao et al., 2010) can be used to ensure that the algorithms truly generalize on unseen data.

AlbuSlava 2016 (Albu et al., 2016) and Majed Latah 2017 (Latah et al., 2017) suggested a new technique built only on spatial and contextual features to predict customer behaviour and human actions recognition. They have downtrodden the performance of the Convolutional Neural Network (CNN) and 3D-CNN to extract spatial info for emotions and action classification. Other researchers illustrated that human actions recognition could be best explained using the motion and speed of the object.

Murad and Ryun (Murad et al., 2017) and Qin (Qin, Z., Zhang et al., 2020) applied body-worn sensors and Long Short-Term Memory (LSTM) Recurrent Neural Networks for human motion interpretation. The latter involves gyroscope and accelerometer measuring. In the same situation, Ning Zhang and Zeyuan Hu (Zhang et al., 2017) only used temporal features based on the local optical flow of the total human silhouette for HAR. Despite the level of performance found, these behaviors stay sensitive in complex scenes, which present differences in background, scale, and texture. These make the recognition of human activities hard. To improve the results of human action classification, researchers in this field planned a mixture of spatial and temporal features. Here, Nicolas (Ballas et al., 2015) applied a novel ideal to learn Spatio-temporal features using Gated Recurrent Units (GRU) with Recurrent Convolution Networks (RCN). This model is based on the pre-trained VGG-16 on an ImageNet transfer learning Model.

Xu and Deng (Xu et al., 2016) and Baldominos (Baldominos et al., 2018) were proposed RNN and CNN to classify video based on its temporal and spatial features. The authors used the GoogleLeNet architecture to combine these features and perform accurate video classification. Zhang (Zhang et al., 2016) make a similar point. Shan and Akella (Shan et al., 2014) and Enea (Cippitelli et al., 2016) employed RGB-D sensors for human skeleton segmentation and kinetic energy to discover essential poses in a vast environment that display intensive motion positions. To identify human behaviors, Gaglio and Morana (Gaglio et al., 2014) used three machine-learning approaches. To detect a human 3D skeletal position, they employed the K-means technique, the SVM method for classification, and Hidden Markov Models (HMM) to model activity. RGB-D sensors were used to identify human skeleton elements, the K-means approach for posture selection, and Sequential Minimal Optimization for training data by Manzi and Dario (Manzi et al., 2017). The objective of this phase was to locate and illustrate that a few fundamental postures are enough to represent and recognize a human activity.

A technique based on the combination of the skeleton and contextual feature extraction was suggested and developed by Srijanet (Das et al., 2018), Cruz (Cruz-Silva et al., 2019), and Khaire (Khaire et al., 2018). The RGB-D sensor and the CNN and LSTM models, and used to extract skeleton features. The CNN model is used to detect the contextual information. Yanli (Ji, Y., et al., 2018) developed a View-guided Skeleton-CNN (VS-CNN) model for arbitrary human view and human action recognition that keeps view differences weak by displaying skeleton sequences and covers a broader range of view angles. Yanli (Ji, Y., et al., 2018) created a View-guided Skeleton-CNN (VS-CNN) system for human variable view and human action recognition that displays skeleton sequences and covers a wider variety of view angles while keeping view differences mild. In the experiments on several different datasets and models, we get some results that the accuracy of the distinguishing approach started 77.00% to 96.00% end, respectively, as shown in table 2.

Table2: Methods accuracy with different data set

Index No	Authors	Years	Methods	Interpretation	Accuracy %	Dataset
1	Gaglio et al., 2015	2015	Kmeans+HMM+SVM	Skeleton feature	77.3	CAD-60 dataset
2	Nicolas et al., 2016	2016	GRU+RCN	Spatio-temporal feature	78.3	UCF 101 dataset
3	Xu et al., and Baldominos et al. 2016	2016	RCNN	Spatio-temporal feature	86.3	Opportunity dataset
4	Zhang et al., 2016	2016	Vector of locally aggregated descriptors, SIFT and ISA	Spatio-temporal feature	82.10	CAD-60 dataset
5	Shan and Akella, 2014 and Enea et al., 2016	2016	Pose Kinetic Energy+SVM	Skeleton feature	83.6	UCF 101 dataset
6	Albu et al., 2016 and Majed Latah 2017	2017	3D CNN	Spatial feature	92.2	CAD-60 dataset
7	Murad and Ryun 2017 and Qin et al. 2017	2017	Deep recurrent neural networks and multimodal sensors	Motion feature	96.7	UCI-HAD dataset
8	Ning et al., 2017	2017	Local optical flow of a global human an silhouette	Motion feature	95.0	KTH dataset
9	Zhang et al., 2017	2017	Kmeans+Sequential optimization	Skeleton feature	92.0	CAD-60 dataset
10	Das et al., 2018, Cruz-Silva et al., 2018 and Khaire	2018	RGB.D+CNN+LSTM model	Skeleton and contextual feature	96.0	MSRDaily Activity 3D dataset
11	Ji, Y., et al., 2018	2018	VS-CNN	Skeleton and contextual feature	96.0	CAD-60 dataset
12	Wang et al., 2019	2019	The conversion of the distance value of two joint to colors points + CNN	Skeleton and contextual feature	88.0	CAD-60 dataset

5. Video datasets

Being naive and scripted, the previous datasets were recorded in arid conditions. While they were seen as benchmarks, the result of these data sets does not imply being an accurate measure of the ability to simplify over actual data, particularly because present algorithms usually achieve 95th percentile or better accuracy on them. The study in (Ohnishi et al., 2016) provides an excellent overview

of those early datasets. Newer datasets, such as UCF and Hollywood, usually contain unconstrained videos that simulate real-world scenarios. The measurability of advanced and annotation methods, including search query data assortment, pic scripts/ sports statement parsing, AMT, and crowdsourcing, has resulted from the rise in the Internet or YouTube data sets. There has also been a substantial rise in the number of categories considered for datasets, to the point that something less than one hundred categories is regarded as a “small dataset” (Aggarwal et al., 1999). Furthermore, in-depth activity datasets, in which a large number of categories belong to one domain, are a stimulating foundation, as they measure the ability to manage variations at extremely low interclass (Feichtenhofer et al., 2016) (Herath et al., 2017). Although the number of the dataset for human action recognition of Egocentric dataset on video, as shown in table 3.

Table 3: Evolution of Egocentric datasets (Singh et al., 2019)

Dataset	Year	Focus	Method	Annotation	Classes
GTEA	2011	Head-mounted camera	Manual annotation of unscripted video	Frame range	7
ADL	2012	Chest mounted camera, activities of daily living, person-object actions	Semi-scripted	Frame range + bounding boxes (object)	18
GTEA Gaze	2012	Head-mounted camera	Manual annotation of unscripted video	Frame range + gaze data	25
Dog Centric activity	2014	Dog mounted camera, animal-human actions	Manual annotation of unscripted video	Sequence level	10
GTEA gaze+	2015	Head-mounted camera, cook-ing actions	Semi-scripted (recipe as script)	Frame range + gaze data, scripts	44
MILADL	2015	Paired wrist and head-mounted camera, ADL	Semi-scripted, manual annotation	Frame range	23
IIIT extreme sports	2017	Head extreme sports actions	YouTube	Frame range	18
Charades Ego	2018	Paired agocentric and third person actions	Hollywood in homes, crowd-sourcing	Frame range + object labels	157
Epic Kitchens	2018	Head-mounted camera, kitchen actions	Crowdsourcing, AMT annotation	Frame range + bounding boxes (object)	125
EGTEA	2018	Head-mounted camera, cook-	Semi-scripted	Frame range + gaze data, hand mask, scripts	106
FPVO	2019	Chest mounted camera, office	Manual annotation of actions unscripted video	Frame range	20

5.1 Aslan dataset

The Action Similarity LAbeliNg or (ASLAN) dataset (Krishna et al., 2017) can consist of 1,571 video clips divided into 432 complex action categories. Victimization data was gathered using YouTube search queries and supported the CMU dataset’s categories (Wang et al., 2018) as well as some

new classes. A total of 3,631 actions were extracted as sequences, with several classes allowing for specific series to be created. The dataset determines the metric of similarity of actions, unlike action classification, which is concerned with the likeness of activities in two videos and has been used as a benchmark by (Sharma et al., 2015), as shown in figure 4.



Figure 4: Example classes from the Aslan dataset (Singh et al., 2019)

5.2 UCF11 (2009) UCF50 (2010) and UCF101 (2012)

The UCF datasets are a collection of increasingly valuable datasets acquired from the internet as part of a project by the University of Central Florida’s Department of Electrical Engineering and Computer Science. These datasets are complicated to gather since they use unscripted videos from unscripted sources.

UCF11 (Liu et al., 2009), also known as “Actions in the Wild”, was one of the first datasets to incorporate unconstrained inputs, utilizing raw data from YouTube videos. It includes 11 action lessons relating to sports and daily activities, such as Basketball Shooting, Walking with a Dog, and Juggling. It comprises 1168 videos divided into 25 groups, each with a different cultural background. UCF50 (Reddy et al., 2013) improves on UCF11 by adding many more classes (50 actions) and reducing interclass variance much further, as shown in figure 5.

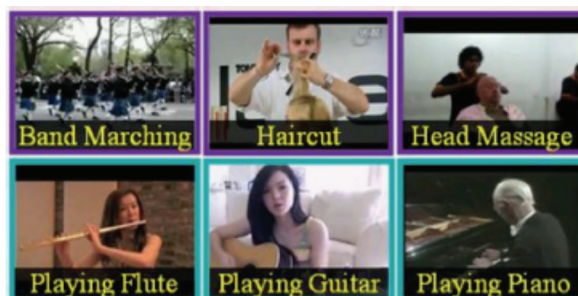


Figure 5: Example classes from the UCF101 dataset (Singh et al., 2019)

6. Image datasets

Specific actions can only be done as a result of static images (Sonwalkar et al., 2015). While despite the fact that the temporal context plays a crucial role in making (Singh et al., 2019) however that has received less attention than the recognition of images, it is still possible to identify video datasets, as shown in table 4.

Table 4: Evaluation of images datasets (Singh et al., 2019)

Dataset	Year	Focus	Method	Class
Willow	2010	Still Images	Flickr search query	7
PPMI	2010	Person-object interactions, Musical instruments	Flickr search query	12
Stanford40	2011	Still Images	Google, Bing, Flickr search query	40
TUHOI	2014	Person-object interactions	Crowdsurceing, crowd flower	2974
HICO	2015	Person-object interactions	Flickr search query	600
BU-Action		UCF101, Action Net class	Google, Bing, Flickr search query	101(BU101-F), 101(BU101-UF), 23(BU203)

6.1. Willow (2010) Dataset

This dataset (Sonwalkar et al., 2015) is a 968-picture still image dataset of typical human behavior culled from Flickr mistreatment search queries after manual filtering. The dataset is divided into seven categories, including Computer Interaction, Horse Riding, and Walking. Every individual within the image is allocated by manually annotated bounding boxes. A plaything exists with seventy pictures in each category, with a reminder to look at the set. Classification accuracy and mAP are the metrics used for this purpose, as shown in figure 6.



Figure 6: A few example classes from the Willow dataset (Delaitre et al., 2010)

6.2. Stanford forty actions (2011)

Stanford forty actions (Sonwalkar et al., 2015) is a picture dataset of forty human actions occurring daily gathered by search queries on Google, Bing, and Flickr. As many as 9,352 pictures exist, each with bounding boxes for the person acting in it. There are approximately 180–300 pictures in each class. A hundred of these are used to make the plaything, and the rest are used to make the check set, as shown in figure 7.



Figure 7: A some example classes from Stanford forty actions dataset

6.3. TUHOI (2014)

The metropolis The Universal Human Object Interaction Dataset (Ballas et al., 2015) includes the interaction between human and object and includes 189 typical objects in ten, 805 images of DET dataset in the ImageNet 2013 contests (Sonwalkar et al., 2015). Mistreatment is annotated for 2974 different activities in the form “verb + object,” which is a crowdsourcing program called Crowd flower. Dog, watercraft, and ball are examples of objects, while Eat, Strike, and Throw are examples of verbs. The images are divided into 50–50 trains and look at sets, every action painted in one of the two sets. Area unit recall, precision, accuracy, as well as the measurement of F1 were used as metrics, as shown in figure 8.

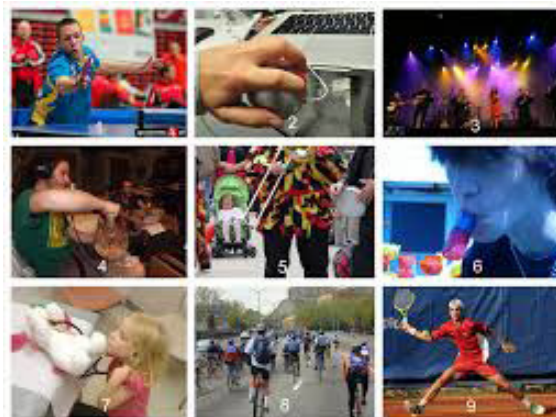


Figure 8: A some example classes from TUHOI dataset

6.4. HICO (2015) and HICO-DET (2018)

Humans Interacting with Common Objects is a dataset that examines a wide range of common senses interactions with identical object. There are 47,774 images and six hundred action groups in the (verb + object) category. There are eighty objects (e.g., bicycle, mobile phone, and apple) and 117 categories (e.g., cut, feed, and ride) that are shared by several objects. The bottom line is that each picture contains several action labels as well as a “No Action” label (for example, “individual is near to but not interacting with bicycle” = “Bike No Action”), as shown in figure 9.

The square measure of the pictures was retrieved from Flickr based on queries about the categories and checked by AMT staff. mAP per picture is the metric for analysis, with a training–test split of 80–20. HICODET (2018) (Singh et al., 2019) adds instance-level annotations to the dataset, which include bounding box-es for the objects and people involved in the acts (while specifically disregarding irrelevant persons) and relation of individual and the pertinent object.

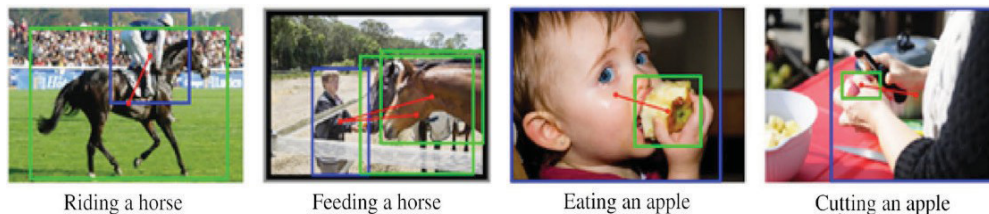


Figure 9: Sample HICO-DET dataset (Singh et al., 2019)

6.5. BU-ACTION (2017)

Massachusetts University Action Datasets (Ma et al., 2017) include three different datasets that were retrieved via queries that corresponded to the categories of benchmark video data- sets, namely UCF101 and ActivityNet. BU101-filtered incorporates UCF101 classes as well as 2769 images from the Stanford40 Dataset. Twenty-three 8 K resolution images are created after manual filtering. The UCF101 categories are also present in BU101- unfiltered, but the photos do not appear to be filtered after the queries, resulting in a dataset of 204 K pictures. BU203- unfiltered, on the other hand, is a set of unfiltered images retrieved from the 203 Activity Net Classes, as shown in figure 10.



Figure 10: Sample picture from BU101. Each row demonstrates images of one action. Top to bottom: Hula Hoop, Jumping Jack, Salsa Spin, Drumming, Frisbee Catch (Ma et al., 2017)

6.6. DATASET CAD-60

Along with CAD-120 data sets; this contains RGB-D video sequences of humans carrying out activities that are recorded using the Microsoft Kinect sensing device. The ability to identify human behaviors is crucial for developing personal assistant robots that can perform useful tasks. Our CAD dataset contains twelve varying activities (comprised of several sub-activities) carried out by four persons in a variety of settings, such as a room, a front room, an office. Robots were used to test the device, which responded to the detected activities, as shown in figure 11.



Figure 11: Sample CAD-60 Action Dataset (Kim et al., 2015)

7. Conclusion

We present efficient techniques for classifying human activity and action recognition in this review paper, which will be accompanied by the subsequent steps: preparation of a new deep learning model utilizing characteristics of human skeleton and activity recognition, CNN model, Kalman filter, Long Short-Term Memory (LSTM). Also, the trained model was substituted by a deep learning model that assisted RNN with Gated Recurrent Unit and used more different approach techniques. The accuracy of act classification was enhanced by greater use of the deep learning model RNN with key point features. We test systems using a continuous video sequence or a series of image datasets (collective action from the CAD-60 dataset).

The system's psychological function capability was enhanced by an arrangement of the learning CNN model, skeleton options, human trailing, and also the deep Learning RNN with Gated Recurrent Unit. As detailed activity detection becomes a serious challenge, having an inventory of datasets in various complexity levels would allow having more categories. This paper will help interested readers choose approximate algorithms and datasets for designing new solutions. Despite substantial progress in recent years, this review paper carries out analysis and comparisons of recognition accuracy between various methods and integrating two or more system-based approaches and other techniques shown in table 4. This paper examines current datasets and offers an analysis of their content and development methods, as well as a thorough description of the most important datasets. The survey also reveals that, while scripted datasets were once standard, current benchmarks focus on annotated internet datasets, particularly YouTube datasets being similar to "Action Recognition in the Wild". The existing datasets illustrate the accuracy of recognition of a range of human actions using multiple approaches for a larger dataset.

References

- Aggarwal, J.K. & Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, 73, 428–440.
- Albu, V. (2016). Measuring Customer Behavior with Deep Convolutional Neural Networks; BRAIN. *Broad Research in Artificial Intelligence and Neuroscience*, 7(1), pp. 74–79.
- Baldominos, A., Saez, Y. & Isasi, P. (2018). Evolutionary Design of Convolutional Neural Networks for Human Activity Recognition in Sensor-Rich Environments. *Sensors*, 18 (4), 1288.
- Ballas, N., Yao, L., Pal, C. & Courville, A. (2015). Delving Deeper into Convolutional Networks for Learning Video Representations. *arXiv preprint arXiv:1511.06432*.
- Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C. & Sivic, J. (2014, September). Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, Springer, Cham, pp. 628-643.
- Cao, L., Liu, Z. & Huang, T. S. (2010, June). Cross-dataset action detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1998–2005.
- Cippitelli, E., Gasparrini, S., Gambi, E. & Spinsante, S. (2016). A Human Activity Recognition System Using Skeleton Data from RGBD Sensors. *Computational Intelligence and Neuroscience*.
- Cruz-Silva, J. E., Montiel-Pérez, J. Y. & Sossa-Azuela, H. (2019, October). 3-D Human Body Posture Reconstruction by Computer Vision. In *Mexican International Conference on Artificial Intelligence*, Springer, Cham, pp. 579-588.
- Das, S., Koperski, M., Bremond, F. & Francesca, G. (2018). A fusion of appearance based CNNs and temporal evolution of skeleton with LSTM for daily live in action. *Recognition ArXiv e-prints*.
- Delaitre, V., Laptev, I. & Sivic, J. (2010, August). Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC 2010-21st British Machine Vision Conference*.
- Faria, D. R., Premebida, C. & Nunes, U. (2014, August). A probabilistic approach for human everyday activities recognition using body motion from RGB-D images. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, IEEE, pp. 732-737.
- Feichtenhofer, C., Pinz, A. & Zisserman, A. (2016). Convolutional twostream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933-1941.
- Gaglio, S., Re, G. L. & Morana, M. (2014). Human Activity Recognition Process Using 3-D Posture Data. *IEEE Transactions on Human-Machine Systems*, 45(5), 586-597.
- Gao, Y., Xiang, X., Xiong, N., Huang, B., Lee, H. J., Alrifai, R., Jiang, X. & Fang, Z. (2018). Human Action Monitoring for Healthcare based on Deep Learning. *IEEE Access*, 6, 52277–52285.
- Herath, S., Harandi, M. & Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and vision computing*, 60, 4-21.
- Idrees, H., Zamir, A. R., Jiang, Y. G., Gorban, A., Laptev, I., Sukthankar, R. & Shah, M. (2017). The THUMOS challenge on action recognition for videos ‘in the wild’. *Computer Vision and Image Understanding*, 155, 1–23.
- Ijjina, E. P. & Krishna Mohan, C. (2016). Hybrid deep neural network model for human action recognition. *Applied Soft Computing*, 46, 936–952.

- Jaouedi, N., Perales, F. J., Buades, J. M., Boujnah, N. & Bouhlel, M. S. (2020). Prediction of Human Activities Based on a New Structure of Skeleton Features and Deep Learning Model. *Sensors*, 20(17), 4944.
- Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H. T. & Zheng, W. S. (2018, October). A Large-scale RGB-D Database for Arbitrary-view Human Action Recognition. In *Proceedings of the 26th ACM Multimedia Conference on Multimedia*, pp. 1510-1518.
- Khaire, P., Kumar, P. & Imran, J. (2018). Combining CNN Streams of RGB-D and Skeletal Data for Human Activity Recognition. *Pattern Recognition Letters*, 115, 107-116.
- Kim, H. & Kim, I. (2015). Human Activity Recognition as Time-Series Analysis. *Mathematical Problems in Engineering*.
- Kliper-Gross, O., Hassner, T. & Wolf, L. (2011). The action similarity labeling challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3), 615–621.
- Koppula, H. S., Gupta, R. & Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research*, 32(8), 951-970.
- Krishna, R., Hata, K., Ren, F., Fei-Fei, L. & Carlos Niebles, J. (2017). Densecaptioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715.
- Latah, M. (2017). Human action recognition using support vector machines and 3D convolutional neural networks. *Int. J. Adv. Intell. Informatics*, 3(1), 47.
- Liu, C., Hu, Y., Li, Y., Song, S. & Liu, J. (2017). PKU-MMD: a large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*.
- Liu, J., Luo, J. & Shah, M. (2009, June). Recognizing realistic actions from videos in the Wild. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1996–2003.
- Ma, S., Bargal, S. A., Zhang, J., Sigal, L. & Sclaroff, S. (2017). Do less and achieve more: Training CNNs for action recognition utilizing action images from the Web. *Pattern Recognition*, 68, 334–345.
- Manzi, A., Dario, P. & Cavallo, F. (2017). A Human Activity Recognition System Based on Dynamic Clustering of Skeleton Data. *Sensors*, 17(5), 1100.
- Minnen, D., Westeyn, T., Starner, T., Ward, J. A. & Lukowicz, P. (2006). Performance metrics and evaluation issues for continuous activity recognition. *Performance Metrics for Intelligent Systems*, 4, 141–148.
- Murad, A. & Pyun, J. Y. (2017). Deep Recurrent Neural Networks for Human Activity Recognition. *Sensors*, 17(11), 2556.
- Ni, B., Pei, Y., Moulin, P. & Yan, S. (2013). Multilevel Depth and Image Fusion for Human Activity Detection. *IEEE Transactions on Cybernetics*, 43(5), 1383-1394.
- Ohnishi, K., Kanehira, A., Kanazaki, A. & Harada, T. (2016). Recognizing activities of daily living with a wrist-mounted camera. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3103–3111.
- Pham, H. H., Khoudour, L., Cruzil, A., Zegers, P. & Velastin Carroza, S. A. (2015). Video-based human action recognition using deep learning: a review.
- Qin, Z., Zhang, Y., Meng, S., Qin, Z. & Choo, K. R. (2020). Imaging and fusing time series for wearable sensors based human activity recognition. *Information Fusion*, 53, 80–87.
- Reddy, K. K. & Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5), 971–981.





- Rodríguez-Moreno, I., Martínez-Otzeta, J. M., Sierra, B., Rodríguez, I. & Jauregi, E. (2019). Video Activity Recognition: State-of-the-Art. *Sensors*, 19(14), 3160.
- Shan, J. & Akella, S. (2014, September). 3D human action segmentation and recognition using pose kinetic energy. In 2014 IEEE International Workshop on Advanced Robotics and Its Social Impacts, IEEE, pp. 69-75.
- Sharma, S., Kiros, R. & Salakhutdinov, R. (2015). Action recognition using visual attention. arXiv preprint arXiv:1511.04119.
- Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A. & Alahari, K. (2018). Charades-ego: a large-scale dataset of paired third and firstperson videos. arXiv preprint arXiv:1804 - 09626.
- Singh, R., Sonawane, A. & Srivastava, R. (2019). Recent evolution of modern datasets for human activity recognition: a deep survey. *Multimedia Systems*, 26(2), 83-106.
- Singh, S., Arora, C. & Jawahar, C. V. (2017). Trajectory aligned features for first person action recognition. *Pattern Recognit*, 62, 45–55.
- Sonwalkar, P., Sakhare, T., Patil, A. & Kale, S. (2015). Hand gesture recognition for real time human machine interaction system. *International Journal of Engineering Trends and Technology (IJETT)*, 19(5), 262-264.
- Srijan, D., Michal, K., Francois, B. & Gianpiero, F. (2018). A Fusion of Appearance based CNNs and Temporal evolution of Skeleton with LSTM for Daily Living Action Recognition. arXiv 2018, arXiv:1802.00421v1.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D. & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9.
- Torralba, A. & Efros, A. A. (2011, June). Unbiased look at dataset bias. In *CVPR 2011*, IEEE, pp. 1521–1528.
- Wang, J., Liu, Z., Wu, Y. & Yuan, J. (2013). Learning Actionlet Ensemble for 3D Human Action Recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(5), 914–927.
- Wang, P., Li, W., Ogunbona, P., Wan, J. & Escalera, S. (2018). RGB-Dbased human motion recognition with deep learning: A survey. In *Computer Vision and Image Understanding*, 171, 118–139.
- Xu, Z., Hu, J. & Deng, W. (2016, July). Recurrent convolutional neural network for video classification. In 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. IEEE.
- Zhang, L., Feng, Y., Han, J. & Zhen, X. (2016, March). Realistic human action recognition: When deep learning meets VLAD. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1352-1356.
- Zhang, N., Hu, Z., Lee, S. & Lee, E. (2017). Human Action Recognition Based on Global Silhouette and Local Optical Flow. *Adv. Eng. Res.*, 134, 1-5.
- Zhao, R., Ali, H. & Van der Smagt, P. (2017, September). Two-stream RNN/CNN for action recognition in 3D videos. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 4260-4267.

References not cited

- Abebe, G., Catala, A. & Cavallaro, A. (2018, August). A first-person vision dataset of office activities. In IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction, Springer, Cham, pp. 27-37.
- Alfaro, A., Mery, D. & Soto, A. (2016). Action recognition in video using sparse coding and relative features. arXiv preprint arXiv:1605.03222 DOI: 10.3390/app7010110.
- Ali, K. H. & Wang, T. (2014, July). Learning features for action recognition and identity with deep belief networks. In 2014 International Conference on Audio, Language and Image Processing, IEEE, pp. 129-132 (2014).
- Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E. & Wray, M. (2018). Scaling egocentric vision: The epic-kitchens dataset. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 720-736.
- Fathi, A., Li, Y. & Rehg, J. M. (2012, October). Learning to recognize daily actions using gaze. In European Conference on Computer Vision (pp. 314-327). Springer, Berlin, Heidelberg.
- Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y. & Malik, J. (2018). AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2018.00633)
- Iwashita, Y., Takamine, A., Kurazume, R. & Ryoo, M. S. (2014, August). First-person animal activity recognition from egocentric videos. In 2014 22nd International Conference on Pattern Recognition (pp. 4310-4315). IEEE.
- Li, Y., Liu, M. & Rehg, J. M. (2018). In the eye of beholder: joint learning of gaze and actions in first person video. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 619-635.
- Li, Y., Ye, Z. & Rehg, J. M. (2015). Delving into egocentric actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 287-295).
- Pirsiavash, H. & Ramanan, D. (2012, June) Detecting activities of daily living in first-person camera views. In 2012 IEEE conference on computer vision and pattern recognition, pp. 2847-2854. IEEE.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1), 1929-1958.
- Steinkraus, D., Buck, I. & Simard, P. Y. (2005, August). Using GPUs for machine learning algorithms. In Eighth International Conference on Document Analysis and Recognition (ICDAR'05), IEEE, pp. 1115-1120.
- Sung, J., Ponce, C., Selman, B. & Saxena, A. (2011). Human activity detection from rgb-d images. CoRRabs/1107.0169 DOI:10.1109/APSIPA.2014.7041588

Author's Biography

	<p>Azhee W. Muhamad is an assistance Lecturer in Basic Education College / Department of Computer at Sulaymaniyah University. He got my MSc from the Department of Computer Science at the Hamdard University of India at 2011, His area of interest is Machine Learning, Computer Vision.</p>
	<p>Aree A. Mohammed MSc in Atomic Physics 2001, MSc in Computer Science 2003, and PhD in Computer Science 2008. He is Head of Computer Department at College of Science / University of Sulaymaniyah and member of its Scientific Committee. He has many research papers in Image Processing and Multimedia fields either in international proceedings or in International Journals.</p>