2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

# Recent Trends in Emotion Analysis: A Big Data Analysis Perspective

Tansel Özyer[1] Duygu Selin Ak[1] Reda Alhajj[2,3,4]
[1]*Department of Computer Engineering*
*TOBB University of Economics and Technology* Ankara, Turkey
[2]*University of Calgary, Calgary Alberta, Canada*
[3]*Istanbul Medipol University, Istanbul, Turkey*
[4]*University of Southern Denmark, Odense, Denmark*
*ozyer@etu.edu.tr, dyakar@etu.edu.tr , alhajj@ucalgary.ca*

***Abstract*—Human action recognition has recently started to find its way into applications in different applications. Accordingly, human action recognition methods are becoming increasingly important in our daily life. They are used for different purposes such as automation, security, surveillance, health, smart home systems, and customer behaviour prediction, among others. Though have more systems with methods provides a rich pool of choices, it is important to well understand the performance of these systems and their success rates in recognizing the right activities in order to decide on the most appropriate system for the current application domain. This survey tackles this issue by analyzing and commenting on the available human action recognition systems and methods.**

***Index Terms*—Human activity recognition, video based recognition, skeleton based recognition,**

## I. Introduction

With the development of machine learning, deep learning, and computer vision methods, people realized the opportunity and benefit of incorporating automated recognition systems in the daily life. The use of human action recognition, which has recently become a trend in recognition systems, is gradually increasing in sectors such as health, safety, automation, robotics, games, etc. The success of the applied method becomes an extremely important issue in human action recognition systems, which are also used in critical issues such as patient care and criminal detection. In this context, it is common to witness the development of new methods which could benefit from the advancement in technology to positively affect the accuracy rate and working performance of the incorporating systems. However, before developing a new method, it is necessary to study well existing methods and investigate whether any of them could satisfactorily achieve the set target or a new method will be needed. To help in this, we examined ten different human action recognition methods which are characterized by satisfying different perspectives. Within the scope of this survey, we obtained a taxonomy by classifying the covered studies. Finally, we also compared the datasets which have been used to train and study the performance of the various models.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

The rest of this survey is organized as follows. Section II briefly covers some of the surveys previously written in 2020. The data sets used in analyzing the performance of these studies are described in Section III. Section IV examines the ten approaches and methods which have been studies in this survey. Section V includes a quantitative analysis of the ten approaches and methods . Future research directions are highlighted in Section VI, and Section VII is closing remarks.

## II. Related Works

The survey conducted by Majumder, Sharmin and Kehtarnavaz [26] compares vision and inertial sensor fusion techniques and multimodality datasets. The methods were examined by distinguishing between single modality and multimodality. In other words, methods with multiple sensors (such as a camera and a Kinect) are emphasized. In addition, the datasets described in the survey have been mostly utilized on multimodality. The employed performance metric was taken into consideration while analyzing the datasets. Only action types, subject and sample number are included for the covered datasets. Minh, et al. [27] examined methods developed for handling sensor based and vision-based datasets. The main purpose of Minh, et al. [27] is to compare the methods using data from different input types. Content information about the datasets has been reported in a table. Many articles are covered in the survey of Beddiar, et al. [28]. Each article has been examined under the titles of activity type, body parts, data input type, input viewpoint and validation mean. Summary information of the datasets is included, and comparison procedures are made according to the activity type. The research of Shah [29] haandles solutions for human action recognition with two different methods, namely representation based and deep network-based solutions. Representation-based solutions are discussed in three different situations for representation type, namely holistic, local, and fusion of features. Along with the solution methods, current approaches to human action recognition and difficulties in the domain are also highlighted. The survey conducted by Karthickkumar and Kumar [30] examined different methods and datasets for human action recognition. The aim of the survey is to show that changes in the

methods and the datasets significantly affect the recognition accuracy. In this context, 7 different human action recognition techniques were mentioned. Furthermore, six different datasets covering single and multi-point of view were examined, and performance comparison of three methods was reported.

## III. The Datasets

All the datasets mentioned and used in the articles covered in this survey are given in detail in Table I. The Web sites, articles and references of the listed datasets have been examined. The information obtained for each dataset in line with the corresponding sources which analyzed the datasets is reported in Table I.

## IV. Human Action Recognition Approaches and Methods

It is possible to evaluate human action recognition methods from many different angles. The methods discussed in the studies described in the literature were developed by focusing on solutions for some target problems. In addition to the target problem of each method, the various datasets and types used also affect the success of the method. Here, the success and performance of the method are determined by scales such as the complexity of the model, how well it is trained, and the quality of the test data. Within the scope of this survey, some new methods have been examined and compared. In addition, each study was classified according to the utilized methods. Classification taxonomy is shown in Figure 1.

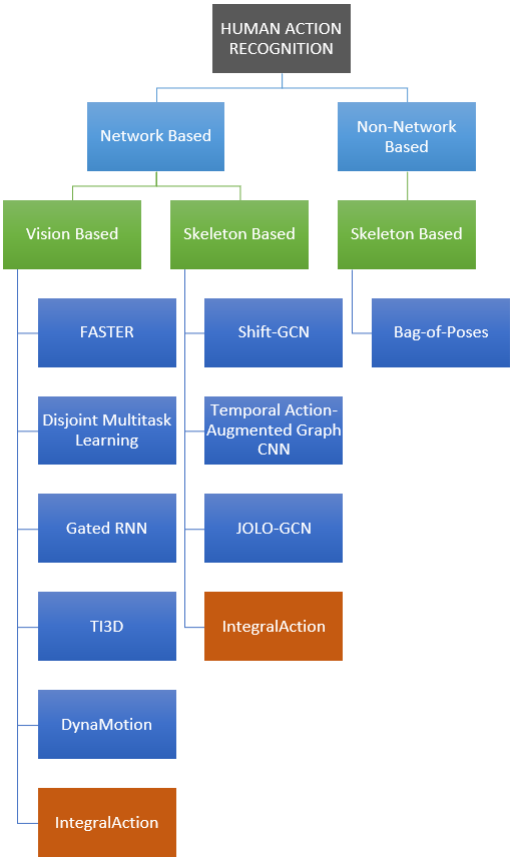


Fig. 1. Taxonomy of the human activity recognition methods covered in this survey.

### A. Network Based Approaches

Due to the increased success of network-based approaches, it has become very common to see the network model highly preferred in complex methods such as human action recognition. With this in mind, a variety of network-based methods have been encountered in the articles reviewed within the scope of this survey.

*1) Skeleton Based Methods:*

*a) Shift-GCN:* Skeleton data for pose estimation methods has recently received considerable interest in the literature, and accordingly has also gained importance in action recognition. In this context, graph convolutional networks (GCN) used in the recognition process are performed with skeleton data. However, Shift-GCN [1] has been developed because GCN methods offer inflexible and complex solutions. Shift-GCN consists of two parts, spatial shift graph convolution and temporal shift graph convolution. Spatial shift graph convolution includes a shift graph operation and a point-wise convolution.

*b) Temporal Attention-Augmented Graph Convolutional Network:* The use of whole-body skeleton in most of the skeleton-based methods incorporated in human action recognition affects the performance of the method. In the context of this study, it is mentioned that not all skeletons obtained from a video are equally important for action recognition. The most informative pose information about the actions is taken, and this information is sufficient for action recognition. Thus, recognition is made before all skeleton data is processed, and this increases computational efficiency. In line with this information, a GCN-based model has been proposed to obtain a subset from the skeletons. In addition, a trainable Temporal Attention Module (TAM) [8] has been developed to extract the most informative skeletal information. It is used in the GCN-based spatio-temporal model to increase the efficiency of the TAM model.

*c) JOLO-GCN:* Skeleton-based action recognition approaches have come up with different methods lately. However, one disadvantage is that sparse skeletal information alone cannot fully characterize human movements in the developed methods. This situation leads to failure to classify some transactions in action recognition processes. To overcome this disadvantage, the JOLO-GCN [9] method has been developed to use human pose skeleton and joint-centered information together. In addition, local movements around each joint are detected using Joint-aligned Optical Flow Patches (JFP). When this hybrid method was compared with other methods using only skeleton-based data, it was observed that the performance and accuracy rates increased.

*2) Vision Based Methods:*

*a) FASTER:* In standard video classification processes, videos are divided into small parts and each clip is considered independently. However, processing similar clips regardless of the temporal structure is one of the factors that increase computational cost. As a solution to this situation, the Feature Aggregation for Spatio Temporal Redundancy (FASTER) [3]

TABLE I
DETAILED COMPARISON OF SOME DATASETS

| DATASETS / FEATURES | Source | Creator | # of Videos | Video Resolution | FPS | Categories | # of Categories | # of Subjects | Video Duration |
|---|---|---|---|---|---|---|---|---|---|
| KTH [11] | Indoor and outdoor recorded videos | Christian Schuldt, Ivan Laptev and Barbara Caputo | 600 | 120x160 | 25 | Walking, Jogging, Running, Boxing, Hand Waving and Hand Clapping | 6 | 25 | 4s |
| UCF - ARG [12] | Camera mounted on Kingfisher Aerostat hellium baloon, ground camera and a rooftop camera | University of Central Florida | 1440 | 1920x1080 | 60 | Boxing, Carrying, Clapping, Digging, Jogging, Open-Close Trunk, Running, Throwing, Walking and Waving | 10 | 12 | 1s - 104s |
| Youtube - Aerial [13] | Drone videos available on Youtube | Waqas Sultani and Mubarak Shah | 500 | - | - | Cycling, Cliff-Diving, Golf-Swing, Horse-Riding, Kayaking, Running, Skateboarding, Surfing, Swimming, and Walking | 10 | - | - |
| Weizmann [14] | Outdoor videos with background and contains cycle of the action | Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani and Ronen Basri | 90 | 180x144 | 50 | Running,Walking, Skipping, Jack, Jump, Pjump, Side, Wave2, Wave1, Bending | 10 | 9 | ∼3s |
| HMDB51 [15] | Video from Youtube, Google and Prelinger archive | H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre | 7000 | Variable | 30 | General facial actions, Facial actions with object manipulation, General body movements, Body movements with object interaction, Body movements for human interaction | 51 | - | 2-3s |
| JHMDB [16] | Subset of HMDB dataset | Jhuang, Gall, Zuffi, Schmid and Black | 928 | - | - | Brush Hair, Catch, Clap, Climb Stairs, Golf, Jump, Kick Ball, Pick, Pour, Pull-up, Push, Run, Shoot Ball, Shoot Bow, Shoot Gun, Sit, Stand, Swing Baseball, Throw, Walk, Wave | 21 | - | - |
| UCF-101 [17] | From Youtube realistic videos | Khurram Soomro, Amir Roshan Zamir and Mubarak Shah | 13320 | 320x240 | 25 | Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments,Sports. | 101 | - | 1.06 - 71.04s |
| UCF Sports Action [18] [19] | Various sports which are typically featured on broadcast television channels such as the BBC and ESPN | Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah | 150 | 720x480 | 10 | Diving,Golf Swing, Kicking, Lifting, Ride Horse, Running, Skateboarding, Swing-Bench, Swing-Side, Walking | 10 | - | 2.20-14.40s |
| NTURGB+D [20] | Videos collected using 3 different Microsoft Kinect V2 cameras. | Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang | 56880 | 1920x1080 | 30 | Daily actions; drink water, fold paper, jump up, bounce ball etc. | 60 | 40 | ∼1-10s |
| NTURGB+D 120 [20] | Videos collected using 3 different Microsoft Kinect V2 cameras. | Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang | 114480 | 1920x1080 | 30 | Daily actions; drink water, fold paper, jump up, bounce ball etc. | 120 | 40 | ∼1-10s |
| Northwestern UCLA [21] | RGB,depth and human skeleton data captured from Kinect cameras | Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.C | - | - | - | Pick up with one hand, pick up with two hands, drop trash, walk around, sit down,stand up, donning, doffing,throw, carry | 10 | 10 | - |
| Kinetics [22] | Human-object and human-human interaction videos from YouTube | Deepmind | 650000 | Variable | Variable | 400/600/700 different actions according to selected dataset version | 400/600/700 | - | ∼10s |
| Mimetics [23] | Subset of Kinetics 400 dataset | Deepmind | 713 | Variable | Variable | Human-Object Interaction videos | 50 | - | ∼10s |
| Kinetics-Skeleton [24] | Contains skeleton information extracted from videos in the Kinetics dataset with OpenPose | Sijie Yan, Yuanjun Xiong, and Dahua Lin | 300000 | 340x256 | 30 | Different human actions collected from YouTube | 400 | - | - |
| AVA [25] | Movie videos | Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, Jitendra Malik | 410 | 320x400 | - | Atomic visual actions | 80 | - | - |

method has been developed. In this context, a network designed to collect the mixture of different representations called FAST-GRU is proposed. The situation defended in the study is that, considering the similarity of frames close to each other and processing each of these frames causes redundancy.

Instead of processing each given frame, FASTER consists of a combination of a model that includes the details of the action and a model that captures the changing scene over time. It aims to cover the entire video at low cost, and hence avoiding duplication. In addition to the FASTER framework, an RNN architecture design called FAST-GRU was prepared. This network is responsible for putting together patterns of different clips. Further, it is stated that FAST-GRU performs a longer learning process than other popular RNN structures.

*b) Disjoint Multitask Learning:* With the developing image technologies, drones have started to take part in the daily life. In this respect, it has become important to capture images from drones used in most areas. The study described in [2] gives a new perspective for action recognition. It is an approach has been developed to work on the videos recorded from the drones. Since the selected video source is very new in recognition, it is extremely difficult to find a ready data set. In this study, an activity recognition process was performed using a limited number of drone videos. The authors followed some methods to increase the number of limited videos. Using Generative Adversarial Network (GAN) [31], realistic looking fake videos can be produced. However, the quality of these videos is not considered sufficient for recognition. Despite that, recent studies show that fake features can be obtained with GAN which consists of 2 different networks, namely generator and discriminator.

*c) Gated-RNN:* In terms of human action recognition, the recognition part and correctly extracting and classifying features to be predicted are equally important. The method described in [6] aims to perform human tracking operations by extracting spatial features from a video. Gaussian Mixture Model (GMM) and Kalman Filter (KF) were used for the classification of feature vectors, while the Gated Recurrent Neural Network (Gated RNN) was used to detect human actions.

*d) Triplet Inflated 3D Convolutional Neural Network (TI3D):* Many studies in the field of human action recognition have been handled with a closed-set perspective. In a closed set classification, all classes are assumed as a priori known. In other words, in closed-set classification, new classes emerging during the test phase are classified according to known classes. Human action recognition is an open-set problem. In open-set classification, the boundaries of known classes are determined. The area outside the boundaries is defined as the open area to which the unknown classes belong. The Triplet Inflated 3D

Convolutional Neural Network (TI3D) [7] model has been developed within the scope of the work described in [7]. It aims to provide high quality feature representation. It has been developed for open-set human action recognition. In addition to feature extraction with TI3D and the open-set recognition framework, it includes classification processes with extreme value machine, and the comparison of the results obtained from these models.

*e) DynaMotion:* It is often preferred to perform recognition from RGB videos in the field of human action recognition. In addition to the data obtained from RGB videos, it is stated that adding extra features from each frame and dynamically linking these features will increase the success in activity recognition. In this study, a new dynamic encoding model was created by extracting temporal information from the movements of the human body. In addition, action recognition processes were carried out with a dynamic motion representation named DynaMotion [10] which feeds CNN.

*3) Skeleton and Vision Based Methods:*

*a) Integral Action:* Both pose-based and vision-based methods are frequently used to find a solution to the human action recognition problem. However, there are some advantages and disadvantages in both methods. Instead of using these two methods alone, the IntegralAction [5] method, which is a more robust and effective method, has been developed. IntegralAction dynamically combines appearance and pose. In this way, it is stated that unnecessary contextual information is filtered out and when the pose information is sufficient for action recognition, it helps people to focus on the motion information.

### *B. Non-Network Based Approaches*

*1) Skeleton Based Methods:*

*a) Bag-of-Poses:* Building on the success of 2D pose-based human action recognition methods, researchers have found a new method to encode 2D poses into the parameter space and compute trajectory features using 2D. They developed a new method for poses encoded into the parameter space. The developed method performs feature extraction from 2D human poses using the OpenPose framework [32]. The Bag-of-Poses method was used to encode low-level spatio-temporal features calculated from 2D poses.

## V. Quntitative Analysis of The Approaches and Methods

After ten different articles were explained in detail under Section IV, a comparison was conducted to examine the differences, similarities, performance rates and methods among these articles more clearly. This section covers the quantitative analysis . In this context, analysis results are shared under two subtitles. The recognition results for all the articles in the survey are detailed in Table II. Success rates in configurations that give the best results from the methods are taken into consideration. Only the success metrics in [7] are given in terms of the $F_1$ Score and the Youdens index. Accuracy values have been reported in other articles.

TABLE II
Recognition Results of The Approaches

| DATASET | STUDY | ACCURACY (%) | F1 SCORE | YOUDENS INDEX |
|---|---|---|---|---|
| UCF - ARG Dataset [12] | [2] | 32.5 | - | - |
| YouTube Aerial Dataset [13] | [2] | 68.3 | - | - |
| UCF101 Dataset [17] | [3] | 96.9 | - | - |
| | [6] | 89.3 | - | - |
| | [7] | - | 0.87 | 0.87 |
| | [10] | 98.4 | - | - |
| HMDB51 Dataset [15] | [3] | 75.7 | - | - |
| | [10] | 84.2 | - | - |
| Weizmann Dataset [14] | [4] | 97.85 | - | - |
| NTURGB+D Dataset [20] | [5] | 91.7 | - | - |
| | [8] | 95.8 | - | - |
| | [9] | 98.1 | - | - |
| | [1] | 96.5 | - | - |
| KTH Dataset [11] | [**?**] | 97.16 | - | - |
| | [6] | 96.3 | - | - |
| NTU-120 RGB+D Dataset [20] | [1] | 85.9 | - | - |
| | [9] | 89.7 | - | - |
| Kinetics Dataset [22] | [3] | 75,3 | - | - |
| | [5] | 73.3 | - | - |
| Northwestern Ucla [2] | [1] | 94,6 | - | - |
| UCF Sports Action [18], [19] | [6] | 89.1 | - | - |
| Kinetics Skeleton Dataset [24] | [8] | 59.77 | - | - |
| | [9] | 62.3 | - | - |
| JHMDB [16] | [10] | 87.3 | - | - |
| Mimetics [23] | [5] | 12.8 | - | - |

## VI. Future Direction

Human action recognition methods have been constantly improved, and new methods are occasionally produced. Increasingly higher accuracy rates are achieved on video data sets. In the future, real-time action recognition processes on live video streams will become noticeable. This will be driven by the success to be achieved on existing data sets. In addition, combining real-time detections from Kinect-like sensors that enable skeleton drawing with the information obtained from live video streams and using hybrid methods will help in making action recognition preferable in critical systems.

## VII. Closing Remarks

Human action recognition systems, have been recently adapted into a variety of practical application domains with direct social and scientific benefit. These range from healthcare to homeland security where more precise automated recognition with high accuracy is the target.

In this survey, we discussed ten different human action recognition studies which were conducted in 2020. We analyzed the architectural structures, data processing styles, feature extraction methods and recognition methods of these studies. We shared the quantitative results obtained. Our main objective from this survey is to provide an easily accessible resource for future methods by highlighting the key attractive features of existing methods and their shortcomings which could be covered in any future attempt to develop new methods with more advanced characteristics. We also examining the major datasets which have been frequently preferred in recent studies. Combining the outcome from the study of the existing methods and the associated datasets used in the testing, we

anticipate the researchers in this field will be able to formulate a better understanding of the state of the art and will have a more sharp vision for future research plans and expectations in this domain which is expected to receive more attention in the future based on the rapidly increased interest in the field and the expanding scope of its applications.

## REFERENCES


[1] Cheng, Ke, et al. "Skeleton-Based Action Recognition With Shift Graph Convolutional Network." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[2] Sultani, Waqas, and Mubarak Shah. "Human Action Recognition in Drone Videos using a Few Aerial Training Examples." arXiv preprint arXiv:1910.10027 (2019).

[3] Zhu, Linchao, et al. "FASTER Recurrent Networks for Efficient Video Classification." AAAI. 2020.

[4] da Silva, Murilo Varges, and Aparecido Nilceu Marana. "Human action recognition in videos based on spatiotemporal features and bag-of-poses." Applied Soft Computing 95 (2020): 106513.

[5] Moon, Gyeongsik, et al. "IntegralAction: Pose-driven Feature Integration for Robust Human Action Recognition in Videos." arXiv preprint arXiv:2007.06317 (2020).

[6] Jaouedi, Neziha, Noureddine Boujnah, and Med Salim Bouhlel. "A new hybrid deep learning model for human action recognition." Journal of King Saud University-Computer and Information Sciences 32.4 (2020): 447-453.

[7] Gutoski, Matheus, André Eugęnio Lazzaretti, and Heitor Silvério Lopes. "Deep metric learning for open-set human action recognition in videos."

[8] Heidari, Negar, and Alexandros Iosifidis. "Temporal Attention-Augmented Graph Convolutional Network for Efficient Skeleton-Based Human Action Recognition." arXiv preprint arXiv:2010.12221 (2020).

[9] Cai, Jinmiao, et al. "JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition." arXiv preprint arXiv:2011.07787 (2020).

[10] Asghari-Esfeden, Sadjad, Mario Sznaier, and Octavia Camps. "Dynamic Motion Representation for Human Action Recognition." The IEEE Winter Conference on Applications of Computer Vision. 2020.

[11] Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Vol. 3. IEEE, 2004.

[12] "Ucf-arg dataset," https://www.crcv.ucf.edu/data/UCF-ARG.php, accessed: 2020-09-8.

[13] Sultani, Waqas, and Mubarak Shah. "Human Action Recognition in Drone Videos using a Few Aerial Training Examples." arXiv preprint arXiv:1910.10027 (2019).

[14] Gorelick, L., Blank, M., Shechtman, E., Irani, M.,Basri, R. (2007). Actions as space-time shapes. IEEE transactions on pattern analysis and machine intelligence, 29(12), 2247-2253.

[15] Kuehne, Hildegard, et al. "HMDB: a large video database for human motion recognition." 2011 International Conference on Computer Vision. IEEE, 2011.

[16] Jhuang, Hueihan, et al. "Towards understanding action recognition." Proceedings of the IEEE international conference on computer vision. 2013.

[17] K.Soomro, R.Zamir,and M.Shah,"Ucf101:A dataset of 101 human actions classes from videos in the wild," in ICCV, 2013.

[18] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah, Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition, Computer Vision and Pattern Recognition, 2008.

[19] Khurram Soomro and Amir R. Zamir, Action Recognition in Realistic Sports Videos, Computer Vision in Sports. Springer International Publishing, 2014.

[20] Shahroudy, Amir, et al. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[21] Wang, Jiang, et al. "Cross-view action modeling, learning and recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[22] Carreira, Joao, et al. "A short note on the kinetics-700 human action dataset." arXiv preprint arXiv:1907.06987 (2019).

[23] Weinzaepfel, Philippe, and Grégory Rogez. "Mimetics: Towards Understanding Human Actions Out of Context." arXiv preprint arXiv:1912.07249 (2019).

[24] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." arXiv preprint arXiv:1801.07455 (2018).

[25] Gu, Chunhui, et al. "Ava: A video dataset of spatio-temporally localized atomic visual actions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[26] Majumder, Sharmin, and Nasser Kehtarnavaz. "Vision and Inertial Sensing Fusion for Human Action Recognition: A Review." IEEE Sensors Journal (2020).

[27] Dang, L. Minh, et al. "Sensor-based and vision-based human activity recognition: A comprehensive survey." Pattern Recognition 108 (2020): 107561.

[28] Beddiar, Djamila Romaissa, et al. "Vision-based human activity recognition: a survey." Multimedia Tools and Applications 79.41 (2020): 30509-30555.

[29] Shaha, Hetal. "A Survey on Representation Based Methods for Human Action Recognition from Video." -: 12.

[30] Karthickkumar, S., and K. Kumar. "A survey on Deep learning techniques for human action recognition." 2020 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2020.

[31] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

[32] Cao, Zhe, et al. "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields." IEEE transactions on pattern analysis and machine intelligence 43.1 (2019): 172-186.