Anomaly Detection from Crowded Video by Convolutional Neural Network and Descriptors Algorithm: Survey

https://doi.org/10.3991/ijoe.v19i07.38871

Ali Abid Hussan Altalbi(✉), Shaimaa Hameed Shaker, Akbas Ezaldeen Ali
Computer Science Department, University of Technology, Baghdad, Iraq
CS.20.02@grad.uotechnology.edu.iq

Abstract—Depending on the context of interest, an anomaly is defined differently. In the case when a video event isn't expected to take place in the video, it is seen as anomaly. It can be difficult to describe uncommon events in complicated scenes, but this problem is frequently resolved by using high-dimensional features as well as descriptors. There is a difficulty in creating reliable model to be trained with these descriptors because it needs a huge number of training samples and is computationally complex. Spatiotemporal changes or trajectories are typically represented by features that are extracted. The presented work presents numerous investigations to address the issue of abnormal video detection from crowded video and its methodology. Through the use of low-level features, like global features, local features, and feature features. For the most accurate detection and identification of anomalous behavior in videos, and attempting to compare the various techniques, this work uses a more crowded and difficult dataset and require light weight for diagnosing anomalies in objects through recording and tracking movements as well as extracting features; thus, these features should be strong and differentiate objects. After reviewing previous works, this work noticed that there is more need for accuracy in video modeling and decreased time, and since attempted to work on real-time and outdoor scenes.

Keywords—anomaly detection, deep learning, CNN, feature representation, global descriptor, local descriptor

1 Introduction

With regard to real-world applications, activity recognition from videos is a hard area that had attracted much attention recently. Explicit action detection as well as recognition continue to be difficult tasks because of considerable fluctuations within high dimension of video data, a cluttered background [1], changing motion speed, and partial occlusion. Effective solutions to such complex issue could pave the way for several beneficial applications, which include the visual surveillance, monitoring, medical monitoring systems, and human-robot collaboration [2]. This is illustrated in Figure 1.

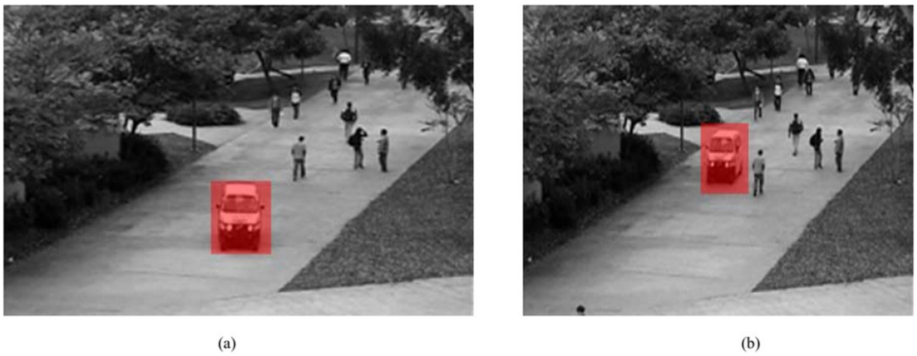


Fig. 1. Vehicle crossing the pedestrian pathway [3]

The concept of action recognition is applicable to many different kinds of human activity. According to how complicated they are, such activities are divided into 4 levels: actions, gestures, interactions, and group activities [4]. Gestures are the fundamental movements regarding body parts of a person and the building blocks of any meaningful movement. An illustration of gesture would be "stretching an arm" and "raising hands." Individual activities or single-person activities which might include many gestures structured in time are called actions. Examples include "bowling" and "walking," "cleaning a sofa" and "brushing teeth." Human activities that include at least two objects or people are called interactions [5]. A pair of objects may interact, for instance, when they shake hands. The definition of group activities is completed by saying that they are actions taken by conceptual groups made up of several people, like "a group conducting a meeting" and "2 groups fighting." [6]. A video can be described as non-stop image sequence with specific semantics and a wealth of data. A video offers 2 key benefits when compared to still images, namely the capacity to retain a constantly varied set of viewpoints on a scene and capacity to record temporal (that is, dynamic) evolution regarding the event. There is no denying that recent advancements in distance sensors have had an impact on machine as well as computer vision applications and research [7]. When looking for RGB images or videos, sensor devices offer detailed information related to the scene view and objects [8]. This is illustrated in Figure 2.

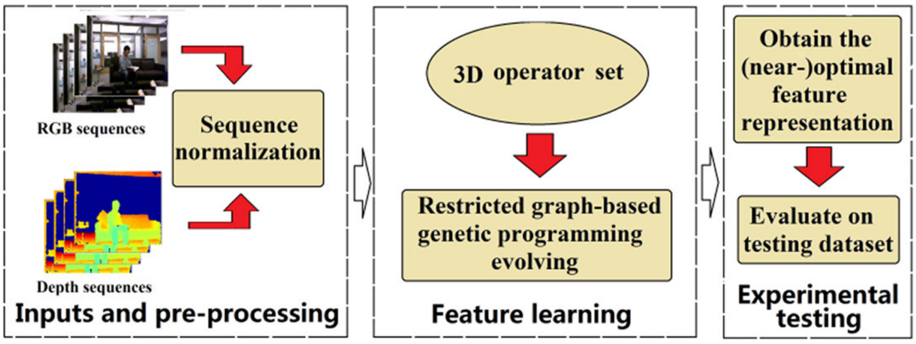

RGB sequences
Sequence normalization
Depth sequences
Inputs and pre-processing
3D operator set
Restricted graph-based genetic programming evolving
Feature learning
Obtain the (near-)optimal feature representation
Evaluate on testing dataset
Experimental testing

Fig. 2. The main flowchart for our proposed method [9]

Digital images made up of pixels, or small picture components grouped in 2D, are known as RGB-D images. Numerical values are used for representing the pixels. Red (R), Green (G), and Blue (B) color respective pixels are each represented by three numerical values in an RGB pixel. Typically, such values are 8-bit numbers with values ranging between 0 and 255, in which larger numbers represent stronger intensities. Those RGB pixels, which consist of 8-bit values might be representing 2563 ≈ 16:7 · 106 colors [5]. The majority of traditionally used features for activity and action recognition tasks could be divided into three categories: global, local, and motion representation techniques [10].

The techniques for extracting local features involve two stages: detection and description, using the Histogram of Oriented Gradients (HOG), Spatio-Temporal Interest Points (STIP) detector, Improved Dense Trajectories (IDT), and Histogram of Optical Flow (HOF). The human action recognition task frequently makes use of such techniques as a local feature [11].

2 Literature survey

Because of the changes in illumination, viewpoint, intra-class variation, and partial occlusions, it remains tough and impossible to identify a certain object from an image dataset. Numerous effective approaches were put forth, such as the ones that efficiently expand beyond image domain into the domains of the video and action recognitions. Yet, there is still room for improvement in the present approaches, particularly for real-world movies and videos [12], where there are many variations in posture and attire of the subjects, a partial occlusions, and a dynamic background. Various studies concentrate on part-based techniques, which just examine the 'interesting' parts of videos instead of the entire video, in order to overcome such shortcomings. Flow vectors or trajectories of the corners and the spatiotemporal points of interest could represent these "parts." Even though part-based techniques show promise, they are nevertheless plagued by erroneous background clutter and motion-induced background detection and tracking regarding interesting parts, which hinders a clear and useful representation [13]. Here are some of the most intriguing works and approaches that were proposed for anomaly detection:

2.1 Anomaly detection and localization in crowded scenes by motion-field shape description and similarity-based statistical learning (2018)

In [3], presented Anomaly detection employed a KNN Similarity-based statistical model, which is considered as unsupervised one-class learning technique that doesn't require clustering or prior assumptions, for detecting anomalies throughout space and time. Initially, the study obtained samples of K-NN from training set that correspond to testing sample, and after that built a Gaussian model using similarities between each K-NN sample pair. Then, a joint probability is produced to represent the probabilities regarding similarities between testing sample and K-NN samples under Gaussian model. Through determining if joint probability falls below predefined thresholds regarding space and time, separately, abnormal events could be identified. Due to the fact that the probability determined in this manner is unaffected by motion distortions brought on by perspective distortion, such an approach could be applied to the entire scene. We carry out tests on the real-world surveillance videos, and outcomes show that the suggested approach could accurately find anomalous events in the video sequences and perform better than the state-of-the-art methods.

2.2 Real-world anomaly detection in surveillance videos (2018)

Training labels (which include normal as well as anomalous) are at video level rather than clip level as it has been indicated in [14]'s Anomaly detection by utilized the learn anomaly by deep multiple instance ranking system through exploiting weakly labeled training videos. In the approach, the anomalous and normal videos have been treated as bags and multiple video segments as samples in multiple instance learning (MIL), and an automatic deep anomaly ranking model which has the ability to predict high anomaly scores for the anomalous video segments is learned. In order to properly locate anomalies throughout training, add sparsity as well as temporal constraints of smoothness to ranking loss function. The experimental dataset findings demonstrate that the suggested anomaly detection approach outperforms standard approaches by a large margin.

2.3 Detecting abnormal event in traffic scenes using unsupervised deep learning approach (2019)

[15] indicated that a difficulty with anomaly detection in video surveillance is especially serious when it comes to identifying abnormal events in traffic scene. The suggested method, which makes use of spatio-temporal features, is employed for detecting abnormalities including wrong-side driving, persons crossing the street against the law, and vehicles traveling via pedestrian pathways. Even though a variety of machine learning (ML) techniques are discussed in literature, most of them concentrate on pixel-wise differences and perform poorly when it comes to spotting abnormalities in video traffic scenes. A new framework for unsupervised deep learning (DL) method for abnormal detection. The approach, which combines clustering loss and reconstruction, is joint based and employs ConvLSTM with kmeans. Using reference to prior frame data, convolutional neural networks (CNN) with the LSTM has been utilized for detecting and identifying traffic scene anomalies. The suggested study includes testing and training phases in order to distinguish between abnormal and normal events. The suggested model is put into practice in real time, and it has been discovered that its accuracy rate to detect abnormalities is 93.02% higher than that of the state-of-the-art methodology.

2.4 LightAnomalyNet: a lightweight framework for efficient abnormal behavior detection (2021)

A CNN trained utilizing input frames generated by a computationally efficient manner was employed for the suggested anomaly detection in [16] with the use of the lightweight framework (LightAnomalyNet). The suggested framework successfully distinguishes between abnormal and normal events. The study distinguishes specific forms of suspicious conduct, human falls, and violent acts in surveillance videos as abnormal activities. Studies using public datasets demonstrate that LightAnomalyNet performs better than the current approaches with regard to classification accuracy and input frames generation.

2.5 SIMCD: SIMulated crowd data for anomaly detection and prediction (2022)

[17] have proposed the MassMotion crowd simulator for anomaly detection and prediction. High densities and walking in opposite direction of the traffic are two different crowd oddities that are shown in the created datasets. Those datasets include SIMCD-Single Anomaly and SIMCD-Multiple Anomalies for the tasks of the anomaly detection, as well as 2 datasets of the SIMCD-Prediction for the tasks of crowd prediction. This study also illustrates data preprocessing (data preparation) through aggregating data as well as presenting new fundamental features, like the crowd severity level and the level of crowdedness, that are helpful for the creation of anomaly detection models and crowd predictions.

3 Applications of human's activities recognition

Many applications of high-level which depend upon representations that have been obtained from visual input benefit from a capability for detecting, tracking, recognizing, and analyzing human motion. Numerous approaches were put out over the past few years to deal with such issues. Applications that might profit from accurate and effective human action recognition include the following, without being restricted to them [4]. This is illustrated in Figure 3:

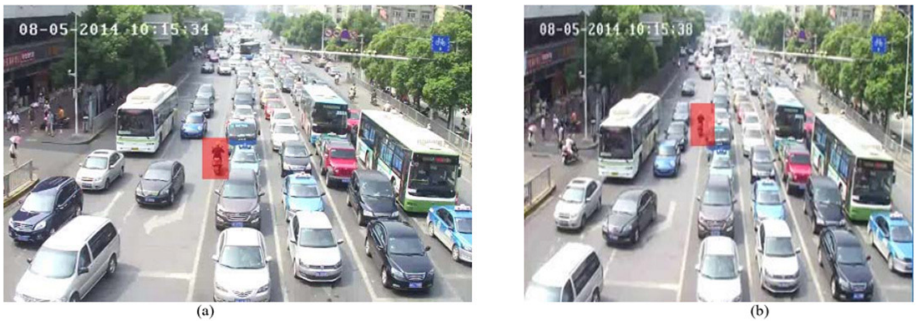


Fig. 3. Scooter moving in the wrong direction [3]

3.1 Intelligent Video Surveillance (IVS)

With the monitoring of video cameras, video surveillance monitors people and interesting objects. Because there is a greater need for a security system, video surveillance had received a lot of attention lately. Surveillance and security systems typically rely on network of the video cameras that have been managed or monitored through a human operator that should be aware of activities taking place in front of cameras. Also, a surveillance system is employed to monitor all public spaces, including government buildings, banks, and airport terminals [18].

3.2 Health Monitoring (HM)

HM and preventative care systems for the patients have applications for the accurate tracking and identification of the individuals in their environs as well as the comprehension and analysis of the activities of patients. Implement, for instance, a surveillance system that can monitor someone's behavior inside their own home without monitoring their privacy. The major goal is to gather data from numerous sensors installed in home and on the mobile phones and deduce the most likely sequence regarding the supervised individual's activities, automatically monitor their behavior and spot any anomalies as they arise. Recognition of human action is thus crucial and frequently employed in medicine, especially in the fields of elderly health surveillance [19].

3.3 Human-Object Interactions (HOI)

The recognition of human activity is challenged in many ways by HOI. It can be difficult to handle interactions between people and objects because of a variety of factors, including an object's size, location, color, and shape [20].

3.4 Human-Computer Interaction (HCI)

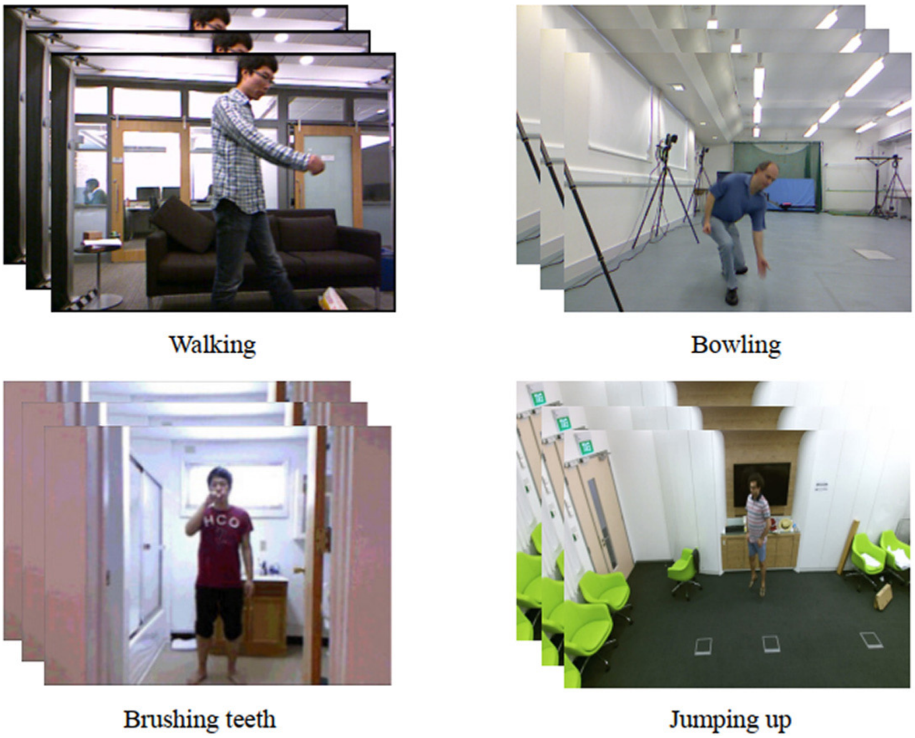
HCI can be defined as a method for examining how humans work with computers. Video cameras are being used for HCI, enabling a natural method of human contact with a device, in addition to being utilized for video surveillance, mobile phones, and laptops. The recognition of gestures and brief activities is therefore one of the crucial needs for the sensor's sides [21].

3.5 Human-Robot Interaction (HRI)

Another crucial use of vision-based activity recognition is in HRI. The capability to perceive interaction activities from robot-centered viewpoint area is provided by HRI, which enables a robot for detecting human behaviors. For determining the purpose of the people nearby, their method aids the robot in completing the next environment tasks [22].

4 Dataset survey

Utilizing suitable human action datasets is a crucial condition for developing a ML system for human action recognition. Those datasets ought to have a wide enough human action variety. Additionally, development of such dataset ought to reflect real-world scenarios. The present section showcases 5 well-known modern action recognition datasets [23]. This is illustrated in Figure 4:



Walking

Bowling

Brushing teeth

Jumping up

Fig. 4. Action and activity video samples, like: Walking [24], Bowling [25], Brushing teeth [26] & Jumping up [8]

4.1 MSR Daily Activity 3D dataset

Microsoft and Northwestern University gathered MSR Daily Activity 3-D (MSR3D) dataset [24] in the year 2012. A Kinect device captured the MSR3D dataset, which concentrated on day-to-day activities in the living room. The camera has been fixed in front of a sofa that is present in the scene in the dataset. There are 10 subjects and 16 actions in the dataset. Every participant performs each activity in two separate positions. Those activities, like "eating, drinking, reading a book, talking on the phone, using a laptop, writing on a paper, cheering up, utilizing vacuum cleaner, still, playing a game, sitting, tossing paper, walking, standing up, lying down on the sofa, and playing guitar," could be utilized for supervised learning. 320 samples make up the total number of activity videos.

4.2 On-line RGBD dataset

On-line RGBD (ORGBD) action dataset [23] aims towards the recognition of the human actions depending upon the RGB-D video data. The ORGBD, which concentrated on HOI, was captured by the Kinect device. 16 participants performed each action twice. ORGBD includes "seven" different categories of living room activity, including "eating, drinking, picking up a phone, utilizing the laptop, reading a book, reading phone (such as an SMS), and using a remote control." ORGBD dataset sample frames.

4.3 Gaming 3D dataset

Kingston University gathered the Gaming 3D (G3D) dataset in the year 2012. Microsoft Kinect recorded the G3D dataset, which concentrated on real-time action recognition in gaming environment. A total of 10 people is featured in the dataset as they perform 20 different gaming actions, including "punch left, punch right, kick right, tennis swing forehand, kick left, tennis serves, golf swing, tennis swing backhand, defend, jump, aim and fire a gun, throw bowling ball, walk, run, climb, flap and clap, wave, steer a car and crouch."

4.4 Cornell Activity dataset

RGB-D video sequences were included in the Cornell Activity (CAD-60) dataset [26] that Cornell University collected in the year 2011. It is inspired by the observation that real daily activities hardly ever take place in organized settings. Thus, actions have been conducted in uncontrolled background through Microsoft Kinect, which included different activities which have been done within 5 indoor environments: Kitchen, office, bathroom, bedroom, and living room.

4.5 NTU RGB+D dataset

Nanyang Technological Univ. gathered the NTU RGB+D dataset [26] in the year 2016. For 3-D action recognition tasks, it has been defined as one of the largest size benchmark datasets that are available. 60 distinct actions were produced in 56; 880 RGB-D video samples. NTU RGB+D dataset's 60 action classes.

5 Discussion and summarizing

In this work, some cutting-edge techniques for identifying human action in actual video data have been presented and assessed. Additionally, it has shown that on various difficult datasets, the suggested solutions outperform state-of-the-art. The techniques for detecting people in a scene and identifying their activities or actions [27]. A total of 4 viewpoints on such techniques were investigated: grayscale, RGB, depth, and RGB-D videos and images. Two distinct parts of methods are employed to increase the recognition regarding human action. For improving human action recognition from 3D sensor data, the 2nd part of the contributions has been characterized through employing one of DNN approaches, namely, CNN [28]. The first part of contributions is dependent upon the hand-crafted features.

5.1 Hand-crafted features

A proposal was made for human action recognition using 3D sensor data. Starting with processing, the system aligned the RGB with the depth frames and removed noise from input depth data [29]. It was suggested for extracting global and local feature vectors. The local feature vectors have been created through extracting such features from 3D sensor data utilizing Optical Flow (OF), Speeded-Up Robust Features (SURF) and Motion History Image (MHI) to find motion interest points. For representing motion and appearance features of all actions, HOG descriptor has been utilized to the images and MHI-OF from the RGB and the depth video channels [30]. In addition, global Hu-moments shape MHI descriptor was used for extracting the global features [31]. With regard to each one of the RGB-D video actions, global and local vectors were eventually concatenated into a single vector. With the use of One-vs-All multi-class and k-means clustering classifiers, such feature vector values have been evaluated in relation to BoW pipeline. The suggested method is quite effective and resistant to cluttered background, rotation, illumination changes, scale, and translation [32]. The findings of the experiment demonstrated that the suggested system might distinguish between various actions, even when they appeared to be comparable, like standing and sitting [33]. Various computer vision applications depend on the anomaly action recognition system. According to the techniques used to describe human action [34]. According to how the features are represented or extracted from video data, the human action recognition system is divided into 3 classes in terms of feature representation: global features, local features, and combinations of features, or combination of the global with local features [35]. This is illustrated in Figure 5.

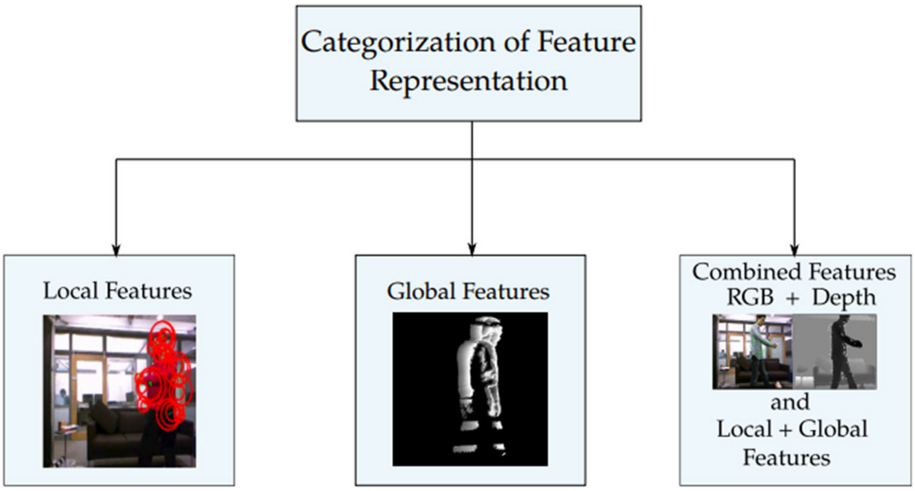

Categorization of Feature Representation
Local Features
Global Features
Combined Features RGB + Depth
and
Local + Global Features

Fig. 5. Categorization of feature representation [36]

A global descriptor gives a description of the entire image. They are typically not particularly robust because a modification to a portion of the image might lead to failure because it will alter the descriptor that is produced. An image patch is described by a local descriptor [37]. An image is matched using a variety of local descriptors, which makes the process more robust because not all of the descriptors must match in order for a comparison to be established. They become more robust to variations in the matched images as a result. SIFT is a major illustration of this. Generally, global features are utilized for lower-level applications like object detection and classification and local features have been utilized for applications of higher level like the object recognition. Combining the local and global features increases recognition accuracy while increasing computing costs [38]. Tables 1, 2, and 3 provide an overview of a few methods that have been previously published and employed (global, local, and combination) features for human action recognition tasks [39].

Table 1. Summary of previous methods that have been utilized, with the local features for the process of action recognition

Reference	Mechanism	Format	Performance
Noguchi et al. [40]	SURF detector was utilized. A codebook has been generated with the use of the k-means clustering. Depending on the codebook, a BoW vector is produced. Used the BoW vectors trained for training SVM.	RGB	the accuracy achieves 86%
Yang et al. [41]	Suggested using the SURF-MHI-HOG approach to separate the temporal and spatial components of RGB videos. BoW vectors are created next, and they are ultimately assessed with the use of a linear SVM classifier.	RGB	achieves top 3 performances in 5 events
Benoit et al. [42]	Utilized Retina model to improve detection potential and perform low-level image processing. A double spatiotemporal filtering process ensures that the video data is properly organized.	RGB	This model has been viewed as image processing kernel, it exhibits interesting characteristics for the image
Sabin et al. [43]	Utilized a Retina model for detecting the BoW from local features like SURF, SIFT, and FREAK by identifying salient areas in the video frames and creating spatio-temporal descriptors. KNN was utilized for the process of the classification.	RGB	Boosts in the performance were viewed on various image signature types from the SIFT to the binary FREAK. T.
Dalal et al. [44]	HOG detectors that mix appearance and motion are suggested, using a linear SVM architecture.	RGB	tested at 8% miss rate on the Test
Uijlings et al. [38]	Utilized a BoW pipeline to calculate trade-off for video classification using HOF, HOG, and MBH descriptors. To investigate the efficiency trade-off, hierarchical kmeans-based and RF-based vocabulary have both been employed.	RGB	Highly powerful for the process of the classification, it's computationally expensive as well.

(Continued)

Table 1. Summary of previous methods that have been utilized, with the local features for the process of action recognition *(Continued)*

Reference	Mechanism	Format	Performance
Kellokumpu et al. [45]	Features that were taken from the descriptors of the LBP-TOP. HMM has been used in order to enhance the activity of the classification.	RGB	computationally simple. Following specific trends in the computer vision researches
Xia et al. [46]	Produced local DCSF in order to characterize 3-D depth cuboid surrounding extracted DSTIPs. On the computed descriptors, k-means clustering, PCA, and SVM classifier were used.	Depth	Outperformed the state-of-art activity recognition approaches on the depth videos
Wang et al. [47]	The Trajectory shape descriptor was suggested. To detect motions and structural data, HOF, HOG, and MBH have been acquired. A multi-class SVM classifier, BoFs, and k means clustering were used to assess the descriptor's performance.	RGB	shown to be effective for action classification, but it might also be applied to video retrieval and action localization.
Koperski et al. [48]	For extracting a 3D Trajectory, motion and depth information were combined. Enhanced SURF performance for actions with low movement rates. By utilizing the k-means algorithm, BoW is obtained. As a classifier, a non-linear SVM is employed.	RGB-D	enhances performance on actions with low movement rate
Xiao et al. [49]	Suggested shape descriptors for 3D Trajectory. MBH was applied to the depth channel. The representation of video feature was produced using a linear SVMs classifier.	RGB-D	Achieved state-of-art performance and greatly outperforms the baseline approaches (STIP-based).

Table 2. A summary of previous methods that have utilized the global features for the purposes of action recognition

Reference	Mechanism	Format	Performance
Carletti et al. [50]	features that were extracted from combined global features, such as depth variations, the <transform, and Hu-moments. Lastly, a GMM classifier is used.	Depth	tested on MHAD and Mivia datasets, with very encouraging results.
AlAzzo et al. [51]	extracted global features from videos using seven Hu-moment invariants. The classifier is an efficient Euclidean distance classifier (EDC).	RGB	For the UCF101 and KTH datasets, respectively, the maximum classification accuracy in the study is 92.11% and 93.4%.

(Continued)

Table 2. A summary of previous methods that have utilized the global features for the purposes of action recognition *(Continued)*

Reference	Mechanism	Format	Performance
Wang et al. [52]	Used the 2DPCA-2DLDA technique and a global GIST feature to suggest a compact representation. SVM classifier is used for recognizing the actions.	RGB	Makes an excellent trade-off between feature discrimination and dimensions, and the accuracy of action recognition is promising.
Douze et al. [53]	Without the need for segmentation, a low-dimensional representation of the scene was created using global GIST. To learn feature vector and convert descriptors into a BoW format for the purpose of learning the visual vocabulary, k-means clustering is utilized.	RGB	on a single machine, it takes 0.18 seconds for an image to be retrieved from an image database with 110 million images.
Bobick et al. [54]	Utilized MEI and MHI to represent the action. To build a recognition system, a matching algorithm for the temporal template consisting of seven Hu moments was applied.	RGB	The technique works in real-time on common platforms, automatically conducts temporal segmentation, and is invariant to linear changes in speed.
Blank et al. [55]	Viewed the actions as silhouettes of moving body. K-NN and euclidean distance were used for evaluating the classification scheme.	RGB	An effective way for dealing with partial occlusions, non-rigid deformations, considerable viewpoint and scale changes, high levels of irregularity in the execution of an action, and low-quality video
Tsai et al. [56]	for representing local motions regarding body parts in global temporal model, combine OF and an MHI. Action models are trained and tested using an SVM classifier.	RGB	with a quick processing rate of 47 frames per second on 200 × 150 images, both test datasets had 100% recognition rates.
Caetano et al. [57]	created the OFCM, a spatiotemporal feature descriptor. So as to explain and capture the local space-time features regarding motion over OF adjacent region, they also extract a collection of features referred to as the Haralick features. The OFCM technique was assessed using an SVM classifier.	RGB	outperforms various commonly utilized spatiotemporal feature descriptors, including HOG3D, HOF, and MBH, demonstrating its suitability for usage as video representation.
Muchtar et al. [58]	enhanced moving object detection with GLCM and bit-plane representation.	RGB	this enables the system to utilize motion history and get rid of shadows, respectively.
Lloyd et al. [59]	GLCM text feature analysis and Haralick features were used for detecting an approach for detecting the events. For training, an RF classifier is utilized.	RGB	for the data from the real world and violent flows, the technique gets ROC values of 0.98 and 0.91, respectively.

Table 3. A summary of previous methods which have utilized the feature combination approaches for the purposes of action recognition

Reference	Mechanism	Format	Performance
Qian et al. [60]	Combined global and local features are characterized by object bounding box of the human blobs and binary MEI, respectively. With the use of binary tree architecture and multi-class SVM, the combined features were categorized.	RGB	show the perfect identification performance and high robustness of the system.
Solmaz et al. [61]	for the purpose of proving the classification task, a spatiotemporal GIST3D feature descriptor which combines with HOF and HOG descriptors was suggested. A SVM with many classes was trained.	RGB	The global descriptor and local descriptor had led to maximum classification accuracy levels on the UCF-50 and HMDB-51 datasets.
Wang et al. [62]	Combined local patch coding and global GIST functionality. The human actions were represented using the BoW approach. The SVM classifier was used to test the recognition performances.	RGB	performed on KTH and UCF sports dataset demonstrate that the proposed representation is effective for human action recognition.
Zhao et al. [63]	SVM classifier and K-means clustering were used to calculate the performance of the approach, which first extracted HOF and HOG from RGB and after that mixed them with LDP from Depth.	RGB-D	Showed that optimal performance has been accomplished in the case of the extraction of the points of interest entirely from the RGB channels, and combine RGB based descriptors and depth map based descriptors.
Fanello et al. [64]	Motions and appearance characteristics which have been characterized by the 3-DHOFs and GHOGs together. After that, simultaneous online video segmentations and action recognition with the use of linear SVMs was developed.	RGB-D	obtain very good results on the ChaLearn Gesture Dataset and with a Kinect sensor.

5.2 Deep neural network methods

Depending on RGB-D data, a deep 3D CNN model was proposed for classifying and recognizing human actions. Through utilizing 3D-CNN, the model extracts features from temporal dimensions. On NTU RGB+D dataset, an action recognition task was trained using a 3D-CNN that used optical flow volumes [65]. It was also tested for 2 additional data splits and tests. 3D-CNN trained and tested using depth data was used as a comparative. Also, the 3D-CNN functioned as an extractor of fixed features. An SVM was employed for classification with the use of features that were retrieved from various layers. Lastly, a 2-Model-SVM classification was performed using the combined feature vectors from the two 3D-CNN models [66].

To understand data such as text, speech, and images, DL, a branch of ML, seeks to learn various representations and abstraction levels [67]. DL techniques can process videos or images in their original state and automatically represent, extract, and classify their characteristics [68]. For representing and recognizing actions, such methods make use of the trainable feature extractors and computational models that have numerous layers for the processing. CNNs [69] are the primary example of a DL method utilized for the process of the action recognition. This is illustrated in Figure 6:

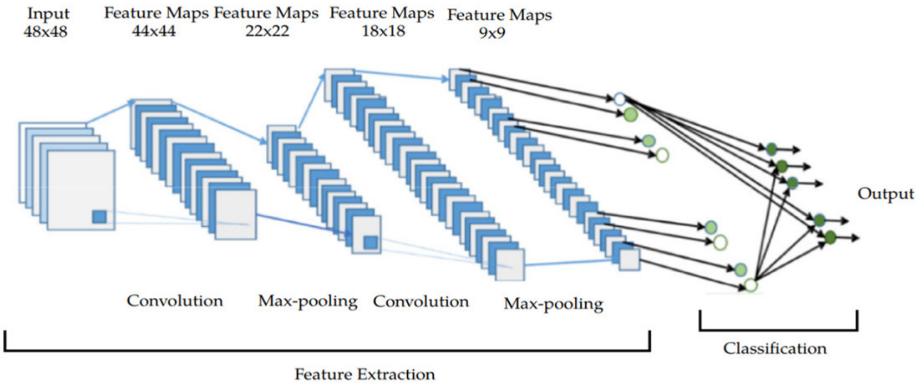


Fig. 6. General CNN structure, which includes input layer, several alternating convolution and max-pooling layers, a single fully-connected layer as well as a single layer of classification [70]

The classification layer, which is a fully connected network, receives its input from the output of the last CNN layer [71]. The classification layer has been implemented using feed-forward NNs. The classification layer has been implemented using feed-forward NNs. The feature that has been extracted from this layer has been used as input for final NN's weight matrix dimension [72]. However, the fully connected layers' network or learning parameters are expensive. In the top classification layer, corresponding class evaluation has been estimated. Depending on highest evaluation, the classifier outputs the relevant classes [73]. Table 4 provides a summary of the cutting-edge techniques for video and image processing that were discussed and implemented using DL, particularly CNN [74]. A summary of certain methods which have been presented before and utilized DNN approaches features for the tasks of the human action recognition.

Table 4. Summary of the previous methods based on CNN for human action recognition

Reference	Mechanism	Format	Performance
Razavian et al. [75]	Demonstrated the viability of the process of the feature extraction from the CNN images with the use of OverFeat network. The retrieved features were then classified using SVM.	RGB	in the majority of visual recognition tasks, strongly features acquired through DL with convolutional nets must be the top contender.
Athiwaratkun et al. [76]	better chances of utilizing features taken from a pre-trained network. The feature vectors have been trained using RF and SVM classifiers.	RGB	Lower-layer features might produce better results for classification than top-layer activations, which are typically used for representing image features for other tasks in literature.
Taylor et al. [77]	Learned feature map representations regarding image sequences using a CNN model.	RGB	exhibited competitive performance on KTH and Hollywood2 datasets.
Simonyan et al. [78]	The two-stream CNN design that is suggested creates several optical flow images.	RGB	It has been trained on and tested against the state of the art on the common video actions standards of HMDB-51 and UCF-101. Additionally, it vastly outperforms earlier attempts to classify videos using deep nets.
Karpathy et al. [79]	CNN was improved from large-scale video by increasing its connectivity in the time domain to acquire local spatiotemporal data.	RGB	show significant performance improvements in comparison with strong feature-based baselines (55.3% to 63.9%), yet just a surprisingly modest enhancements in comparison with single-frame models (59.3% to 60.9%).
Carreira et al. [80]	suggested a 2-Stream Inflated 3-D ConvNet from temporal as well as the spatial domains.	RGB	in the case when it comes to action classification, 3D models significantly outperform current technology, scoring 97.9% on the UCF-101 and 80.2% on the HMDB-51.
Baccouche et al. [81]	for capturing the nature of the video data, the 3D-CNN was suggested. The network is trained in order to assign a limited number of succeeding frames a vector of spatio-temporal features, and after that uses the feature vectors for classifying the complete sequences.	RGB	surpasses current deep models and produces outcomes that are equivalent to the finest efforts in the field.
Latah et al. [82]	for extracting spatio-temporal features from neighboring video frames, a 3D-CNN method was utilized. Each case was classified using the extracted features using an SVM classifier.	RGB	on the KTH action recognition dataset, the suggested architecture was trained and assessed, and it performed well.

(Continued)

Table 4. Summary of the previous methods based on CNN for human action recognition *(Continued)*

Reference	Mechanism	Format	Performance
Song et al. [83]	Enhanced RGB-D scene recognition through applying the pre-trained RGB-CNN models as well as fine-tuning regarding RGB model of scene recognition to the target of RGB-D data.	RGB-D	on both depth-only and integrated RGB-D data from SUN and NYU2, the system achieves cutting-edge accuracy.
Wang et al. [84]	To suggest human action recognition, a 3D-CNN was applied to depth map sequences. By creating a weighted DMM, they were able to extract information on body motion and shape at various temporal scales. The configuration of this motion data was designed for CNN input.	Depth	unlike current approaches, which saw a decline in performance as the number of operations increased, this algorithm maintained its performance on large dataset.
Wang et al. [85]	constructed a network over the video segments using max-pooling operators and 3D convolutions. The feature map could be obtained by running a 3D-CNN to illustrate the 3D activity recognition model.	Depth	in the case when rotation invariant texture features are extracted and categorized, the approach could be noticeably better than the other cutting-edge techniques.
Asadi et al. [86]	a MMDT depending on scene flow was suggested to characterize video sequences in order to enhance the human action recognition. This study integrated manually created and learning-based 2D-CNN features.	RGB-D	demonstrates how the accuracy is improved by the new representation. Additionally, the combination of learning-based and handcrafted features improves the final performance and yields cutting-edge outcomes.
Liu et al. [87]	Enhanced action recognition depending on viewpoint-invariant features acquired from the depth data regarding the RGB videos, as well as simultaneously collected view-invariant human pose characteristics depending upon CNN model.	RGB-D	achieves an accuracy improvement of up to 7.2% over the closest rival.

A few of the upcoming research directions include saliency detection for global and local features as well as 3D skeleton joints for CNN-based human action recognition.

6 Conclusion

Using handcrafted features, ML, and a DNN-based technique, this study provided an overview of the most recent technologies for recognizing human actions. Public data sets that are well-known and open to testing will also be provided to offer greater insight into this area. Modeling the crowded video scene and identifying abnormal movements should be done quickly and accurately. This is accomplished by supplying both global and local features that are powerful and effective. The visible and moving (short-term) part of the object has been represented through points in the frame or image, and anything that deviates from these particular natural movements is regarded as abnormal movement. A more difficult and crowded data set will be employed, and video modeling will be more accurate and faster. For improving the recognition of the human actions from 3D sensor data, anomalies in the object should be identified through tracking and recording movements, extract features, and making sure that these features are strong and separate the object from other objects through the use of one of the DNN approaches, namely, CNN.

7 References

[1] A. A. Karim, "Construction of a robust background model for moving object detection in video sequence," *Iraqi J. Sci.*, vol. 59, no. 2, pp. 969–979, 2018. https://doi.org/10.24996/ijs.2018.59.2B.19

[2] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimed. Inf. Retr.*, vol. 2, no. 2, pp. 73–101, 2013. https://doi.org/10.1007/s13735-012-0024-2

[3] X. Zhang, S. Yang, X. Zhang, W. Zhang, and J. Zhang, "Anomaly detection and localization in crowded scenes by motion-field shape description and similarity-based statistical learning," *arXiv Prepr. arXiv1805.10620*, 2018.

[4] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *Acm Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011. https://doi.org/10.1145/1922649.1922653

[5] A. E. Ali and N. F. Hassan, "Proposing a scheme for human interactive proof test sing plasma effect," *Baghdad Sci. J.*, vol. 16, no. 2, 2019. https://doi.org/10.21123/bsj.16.2.0409

[6] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, 2013. https://doi.org/10.1109/TCYB.2013.2265378

[7] A. A. A. Karim and R. A. Sameer, "Static and dynamic video summarization," *Iraqi J. Sci.*, vol. 60, no. 7, pp. 1627–1638, 2019. https://doi.org/10.24996/ijs.2019.60.7.23

[8] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019. https://doi.org/10.1109/CVPR.2016.115

[9] L. Liu and L. Shao, "Learning discriminative representations from RGB-D video data," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[10] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558. https://doi.org/10.1109/ICCV.2013.441

[11] L. Shao, L. Liu, and M. Yu, "Kernelized multiview projection for robust action recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 115–129, 2016. https://doi.org/10.1007/s11263-015-0861-6

[12] E. Hato and M. E. Abdulmunem, "Fast algorithm for video shot boundary detection using SURF features," *SCCS 2019–2019 2nd Sci. Conf. Comput. Sci.*, 2019, pp. 81–86. https://doi.org/10.1109/SCCS.2019.8852603

[13] X. Zhen and L. Shao, "Action recognition via spatio-temporal local features: A comprehensive study," *Image Vis. Comput.*, vol. 50, pp. 1–13, 2016. https://doi.org/10.1016/j.imavis.2016.02.006

[14] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488. https://doi.org/10.1109/CVPR.2018.00678

[15] K. Meena, A. Viji, J. J. Athanesious, and V. Vaidehi, "Detecting abnormal event in traffic scenes using unsupervised deep learning approach," in *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2019, pp. 355–362. https://doi.org/10.1109/WiSPNET45539.2019.9032774

[16] A. Mehmood, "LightAnomalyNet: A lightweight framework for efficient abnormal behavior detection," *Sensors*, vol. 21, no. 24, p. 8501, 2021. https://doi.org/10.3390/s21248501

[17] A. Bamaqa, M. Sedky, T. Bosakowski, B. B. Bastaki, and N. O. Alshammari, "SIMCD: SIMulated crowd data for anomaly detection and prediction," *Expert Syst. Appl.*, vol. 203, p. 117475, 2022. https://doi.org/10.1016/j.eswa.2022.117475

[18] A. Taha, H. H. Zayed, M. E. Khalifa, and E.-S. M. El-Horbaty, "Human activity recognition for surveillance applications," in *Proceedings of the 7th International Conference on Information Technology*, 2015, pp. 577–586. https://doi.org/10.15849/icit.2015.0103

[19] G. Sebestyen, I. Stoica, and A. Hangan, "Human activity recognition and monitoring for elderly people," in *2016 IEEE 12th international conference on intelligent computer communication and processing (ICCP)*, 2016, pp. 341–347. https://doi.org/10.1109/ICCP.2016.7737171

[20] T. Subetha and S. Chitrakala, "A survey on human activity recognition from videos," in *2016 international conference on information communication and embedded systems (ICICES)*, 2016, pp. 1–7. https://doi.org/10.1109/ICICES.2016.7518920

[21] H.-K. Lee and J.-H. Kim, "An HMM-based threshold model approach for gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 961–973, 1999. https://doi.org/10.1109/34.799904

[22] L. Xia, I. Gori, J. K. Aggarwal, and M. S. Ryoo, "Robot-centric activity recognition from first-person rgb-d videos," in *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 357–364. https://doi.org/10.1109/WACV.2015.54

[23] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Computer Vision—ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1–5, 2014, Revised Selected Papers, Part V 12*, 2015, pp. 50–65. https://doi.org/10.1007/978-3-319-16814-2_4

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

[25] V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 7–12. https://doi.org/10.1109/CVPRW.2012.6239175

[26] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *workshops at the twenty-fifth AAAI conference on artificial intelligence*, 2011.

[27] R. A. Lateef and A. R. Abbas, "Human activity recognition using smartwatch and smartphone: A review on methods, applications, and challenges," *Iraqi J. Sci.*, vol. 63, no. 1, pp. 363–379, 2022. https://doi.org/10.24996/ijs.2022.63.1.34

[28] R. Lateef and A. Abbas, "Tuning the hyperparameters of the 1D CNN model to improve the performance of human activity recognition," *Eng. Technol. J.*, vol. 40, no. 4, pp. 547–554, 2022. https://doi.org/10.30684/etj.v40i4.2054

[29] M. N. Abdullah and Y. H. Ali, "Vehicles detection system at different weather conditions," *Iraqi J. Sci.*, vol. 62, no. 6, pp. 2040–2052, 2021. https://doi.org/10.24996/ijs.2021.62.6.30

[30] H. Xu, L. Li, M. Fang, and F. Zhang, "Movement human actions recognition based on machine learning," *Int. J. Online Eng.*, vol. 14, no. 4, pp. 193–210, 2018, https://doi.org/10.3991/ijoe.v14i04.8513

[31] W. H. Ali, "A new method using naive bayes and RGBD facial identification based on extracted features from image pixels," *Eng. Technol. J.*, vol. 39, no. 4A, pp. 632–641, 2021. https://doi.org/10.30684/etj.v39i4A.1936

[32] S. Pinitkan and N. Wisitpongphan, "Abnormal activity detection and notification platform for real-time Ad Hoc network," *Int. J. online Biomed. Eng.*, vol. 16, no. 15, pp. 45–63, 2020. https://doi.org/10.3991/ijoe.v16i15.16065

[33] I. W. Ghindawi, "A proposed registration method using tracking interest features for augmented reality," vol. 34, 2018.

[34] S. H. Shaker and F. Q. Al-Khalidi, "Human gender and age detection based on attributes of face," *Int. J. Interact. Mob. Technol.*, vol. 16, no. 10, pp. 176–190, 2022. https://doi.org/10.3991/ijim.v16i10.30051

[35] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *J. Healthc. Eng.*, vol. 2017, 2017. https://doi.org/10.1155/2017/3090343

[36] Y. Dedeoğlu, B. U. Töreyin, U. Güdükbay, and A. E. Çetin, "Silhouette-based method for object classification and human action recognition in video," in *European Conference on Computer Vision*, 2006, pp. 64–77. https://doi.org/10.1007/11754336_7

[37] S. M. Hamandi, A. M. S. Rahma, and R. F. Hassan, "A new hybrid shape moment invariant techniques for face identification in thermal and visible visions," *Proc.—2020 21st Int. Arab Conf. Inf. Technol. ACIT 2020*, 2020. https://doi.org/10.1109/ACIT50332.2020.9300069

[38] J. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe, "Video classification with densely extracted hog/hof/mbh features: An evaluation of the accuracy/computational efficiency trade-off," *Int. J. Multimed. Inf. Retr.*, vol. 4, no. 1, pp. 33–44, 2015. https://doi.org/10.1007/s13735-014-0069-5

[39] A. A. B. Badr and A. K. Abdul-Hassan, "Gender detection in children's speech utterances for human-robot interaction," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 5, pp. 5049–5054, 2022. https://doi.org/10.11591/ijece.v12i5.pp5049-5054

[40] A. Noguchi and K. Yanai, "Extracting spatio-temporal local features considering consecutiveness of motions," in *Computer Vision–ACCV 2009: 9th Asian Conference on Computer Vision, Xi'an, September 23-27, 2009, Revised Selected Papers, Part II 9*, 2010, pp. 458–467. https://doi.org/10.1007/978-3-642-12304-7_43

[41] X. Yang, C. Yi, L. Cao, and Y. Tian, "MediaCCNY at TRECVID 2012: Surveillance event detection," in *TRECVID*, 2012.

[42] A. Benoit, A. Caplier, B. Durette, and J. Hérault, "Using human visual system modeling for bio-inspired low level image processing," *Comput. Vis. Image Underst.*, vol. 114, no. 7, pp. 758–773, 2010. https://doi.org/10.1016/j.cviu.2010.01.011

[43] S. T. Strat, A. Benoit, and P. Lambert, "Retina enhanced bag of words descriptors for video classification," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1307–1311.

[44] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *european conference on computer vision*, 2006, pp. 428–441. https://doi.org/10.1007/11744047_33

[45] V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Human activity recognition using a dynamic texture based method," in *BMVC*, vol. 1, p. 2, 2008.

[46] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2834–2841. https://doi.org/10.1109/CVPR.2013.365

[47] H. Wang, A. Klaser, C. Schmid, and L. Cheng-Lin, "Action recognition by dense trajectories. Computer vision and pattern recognition (CVPR)," in *2011 IEEE Conference on*, 2011, pp. 3169–3176. https://doi.org/10.1109/CVPR.2011.5995407

[48] M. Koperski, P. Bilinski, and F. Bremond, "3D trajectories for action recognition," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4176–4180. https://doi.org/10.1109/ICIP.2014.7025848

[49] Y. Xiao, G. Zhao, J. Yuan, and D. Thalmann, "Activity recognition in unconstrained rgb-d video using 3d trajectories," in *SIGGRAPH Asia 2014 Autonomous Virtual Humans and Social Robot for Telepresence*, pp. 1–4, 2014. https://doi.org/10.1145/2668956.2668961

[50] V. Carletti, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "Recognition of human actions from rgb-d videos using a reject option," in *International Conference on Image Analysis and Processing*, 2013, pp. 436–445. https://doi.org/10.1007/978-3-642-41190-8_47

[51] F. Al-Azzo, A. M. Taqi, and M. Milanova, "3D Human action recognition using Hu moment invariants and euclidean distance classifier," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, 2017. https://doi.org/10.14569/IJACSA.2017.080403

[52] Y. Wang, Y. Li, and X. Ji, "Human action recognition using compact global descriptors derived from 2DPCA-2DLDA," in *2014 IEEE International Conference on Computer and Information Technology*, 2014, pp. 182–186. https://doi.org/10.1109/CIT.2014.56

[53] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, pp. 1–8. https://doi.org/10.1145/1646396.1646421

[54] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001. https://doi.org/10.1109/34.910878

[55] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005, vol. 2, pp. 1395–1402. https://doi.org/10.1109/ICCV.2005.28

[56] D.-M. Tsai, W.-Y. Chiu, and M.-H. Lee, "Optical flow-motion history image (OF-MHI) for action recognition," *Signal, Image Video Process.*, vol. 9, no. 8, pp. 1897–1906, 2015. https://doi.org/10.1007/s11760-014-0677-9

[57] C. Caetano, J. A. dos Santos, and W. R. Schwartz, "Optical flow co-occurrence matrices: A novel spatiotemporal feature descriptor," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 1947–1952. https://doi.org/10.1109/ICPR.2016.7899921

[58] K. Muchtar, C.-H. Yeh, Z.-Y. Jian, C.-Y. Lin, W.-Y. Lin, and W.-J. Huang, "Moving object detection based on image bit-planes and co-occurrence matrix in video surveillance," in *2017 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2017, pp. 1–2. https://doi.org/10.1109/ICCE-China.2017.7990965

[59] K. Lloyd, D. Marshall, S. C. Moore, and P. L. Rosin, "Detecting violent crowds using temporal analysis of GLCM texture," *arXiv Prepr. arXiv1605.05106*, 2016.

[60] H. Qian, Y. Mao, W. Xiang, and Z. Wang, "Recognition of human activities using SVM multi-class classifier," *Pattern Recognit. Lett.*, vol. 31, no. 2, pp. 100–111, 2010. https://doi.org/10.1016/j.patrec.2009.09.019

[61] B. Solmaz, S. M. Assari, and M. Shah, "Classifying web videos using a global video descriptor," *Mach. Vis. Appl.*, vol. 24, no. 7, pp. 1473–1485, 2013. https://doi.org/10.1007/s00138-012-0449-x

[62] Y. Wang, Y. Li, and X. Ji, "Human action recognition based on global gist feature and local patch coding," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 8, no. 2, pp. 235–246, 2015. https://doi.org/10.14257/ijsip.2015.8.2.23

[63] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, "Combing rgb and depth map features for human activity recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1–4.

[64] S. R. Fanello, I. Gori, G. Metta, and F. Odone, "One-shot learning for real-time action recognition," in *Iberian Conference on Pattern Recognition and Image Analysis*, 2013, pp. 31–40. https://doi.org/10.1007/978-3-642-38628-2_4

[65] S. S. Ghintab and M. Y. Hassan, "CNN-based Visual Localization for Autonomous Vehicles under Different Weather Conditions," *Eng. Technol. J.*, vol. 41, p. 2, 2023. https://doi.org/10.30684/etj.2022.135917.1289

[66] H. Abdullah and H. Abduljaleel, "Deep CNN based skin lesion image denoising and segmentation using active contour method," *Eng. Technol. J.*, vol. 37, no. 11A, pp. 464–469, 2019. https://doi.org/10.30684/etj.37.11A.3

[67] Z. F. Shaaf, M. Mahadi, A. Jamil, and R. Ambar, "A convolutional neural network model to segment myocardial infarction from MRI images," pp. 150–162. https://doi.org/10.3991/ijoe.v19i02.36607

[68] T. H. Obaida, A. S. Jamil, and N. F. Hassan, "Real-time face detection in digital video-based on Viola-Jones supported by convolutional neural networks," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 3, pp. 3083–3091, 2022. https://doi.org/10.11591/ijece.v12i3.pp3083-3091

[69] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE international symposium on circuits and systems*, pp. 253–256, 2010. https://doi.org/10.1109/ISCAS.2010.5537907

[70] M. Z. Alom *et al.*, "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, vol. 8, no. 3, p. 292, 2019. https://doi.org/10.3390/electronics8030292

[71] M. A. A. Siddique, J. Ferdouse, M. T. Habib, M. J. Mia, and M. S. Uddin, "Convolutional neural network modeling for eye disease recognition," *Int. J. online Biomed. Eng.*, vol. 18, no. 9, pp. 115–130, 2022. https://doi.org/10.3991/ijoe.v18i09.29847

[72] W. S. Ahmed and A. A. A. Karim, "Motion classification using CNN based on image difference," *CITISIA 2020—IEEE Conf. Innov. Technol. Intell. Syst. Ind. Appl. Proc.*, 2020. https://doi.org/10.1109/CITISIA50690.2020.9371835

[73] H. I. Abdulrazzaq and N. F. Hassan, "Modified siamese convolutional neural network for fusion multimodal biometrics at feature level," *SCCS 2019–2019 2nd Sci. Conf. Comput. Sci.*, pp. 12–17, 2019. https://doi.org/10.1109/SCCS.2019.8852593

[74] M. Kodher, J. H. Saud, and H. S. Hassan, "Wheelchair movement based on convolution neural network," *Eng. Technol. J.*, vol. 39, no. 6, pp. 1019–1030, 2021. https://doi.org/10.30684/etj.v39i6.1615

[75] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813. https://doi.org/10.1109/CVPRW.2014.131

[76] B. Athiwaratkun and K. Kang, "Feature representation in convolutional neural networks," *arXiv Prepr. arXiv1507.02313*, 2015.

[77] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *European conference on computer vision*, 2010, pp. 140–153. https://doi.org/10.1007/978-3-642-15567-3_11

[78] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[79] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1725–1732. https://doi.org/10.1109/CVPR.2014.223

[80] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308. https://doi.org/10.1109/CVPR.2017.502

[81] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *International workshop on human behavior understanding*, 2011, pp. 29–39. https://doi.org/10.1007/978-3-642-25446-84

[82] M. Latah, "Human action recognition using support vector machines and 3D convolutional neural networks," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 1, p. 47, 2017. https://doi.org/10.26555/ijain.v3i1.89

[83] X. Song, L. Herranz, and S. Jiang, "Depth CNNs for RGB-D scene recognition: Learning from scratch better than transferring from RGB-CNNs," in *Thirty-First AAAI conference on artificial intelligence*, 2017. https://doi.org/10.1609/aaai.v31i1.11226

[84] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. Human-Machine Syst.*, vol. 46, no. 4, pp. 498–509, 2015. https://doi.org/10.1109/THMS.2015.2504550

[85] Y. Wang, Y. Zhao, and Y. Chen, "Texture classification using rotation invariant models on integrated local binary pattern and Zernike moments," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp. 1–12, 2014. https://doi.org/10.1186/1687-6180-2014-182

[86] M. Asadi-Aghbolaghi, H. Bertiche, V. Roig, S. Kasaei, and S. Escalera, "Action recognition from RGB-D data: Comparison and fusion of spatio-temporal handcrafted features and deep strategies," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3179–3188. https://doi.org/10.1109/ICCVW.2017.376

[87] J. Liu, N. Akhtar, and A. Mian, "Viewpoint invariant RGB-D human action recognition," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017, pp. 1–8. https://doi.org/10.1109/DICTA.2017.8227505

8 Authors

Ali Abid Hussan Altalbi, received the BSc Computer Sciences in 2006 from the University of Babylon, and MSc degrees in Computer Sciences in 2019 from the University of Technology, Baghdad, Iraq, current Ph.D. student in Department Computer Sciences, University of Technology, Baghdad, Iraq, His research focuses on Computer Vision, Video and Image processing and pattern recognition (CS.20.02@grad.uotechnology.edu.iq).

Dr. Shaimaa Hameed Shaker earned his Ph.D. in Computer science from the Department of Computer Science at the University of Technology Baghdad-Iraq since 2006. Shaimaa earned her bachelor's and master's degree also in Computer Science at the University of Technology(UOT)-Baghdad-Iraq since 1996. Her research interested focus on Image processing and pattern recognition and security visual cryptography systems. (120011@uotechnology.edu.iq).

Dr. Akbas Ezaldeen Ali earned her Ph.D., Msg, and BSc in Computer Science from the Department of Computer Science at the University of Technology, Baghdad, Iraq., Her research interested focus on Video and image Processing, Artificial Intelligent, and Computer Vision (110034@uotechnology.edu.iq).

Article submitted 2023-02-11. Resubmitted 2023-03-12. Final acceptance 2023-03-13. Final version published as submitted by the authors.