ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Special Issue (April)

Online - 2455-3891
Print - 0974-2441

Full Proceeding Paper

A REVIEW ON MACHINE LEARNING ALGORITHMS ON HUMAN ACTION RECOGNITION

ANKUSH RAI, JAGADEESH KANNAN R

School of Computing Science & Engineering, VIT University, Chennai, Tamil Nadu, India. Email: ankushressci@gmail.com

Received: 28 December 2016 , Revised and Accepted: 10 May 2017

ABSTRACT

Human action recognition is a vital field of computer vision research. Its applications incorporate observation frameworks, patient monitoring frameworks, and an assortment of frameworks that include interactions between persons and electronic gadgets, for example, human–computer interfaces. The vast majority of these applications require an automated recognition of abnormal or anomalistic action states, made out of various straightforward (or nuclear) actions of persons. This study gives an overview of different best in class research papers on human movement recognition. Open datasets intended for the assessment of the recognition procedures are also discussed in this paper too, for comparing results of several methodologies on this datasets. We examine both the approaches produced for basic human actions and those for abnormal action states. These methodologies are taxonomically classified based on looking at the points of interest and constraints of every methodology. Space-time volume approaches and sequential methodologies that represent actions and perceive such action sets straightforwardly from images are discussed. Next, hierarchical recognition approaches for abnormal action states are introduced and looked at. Statistic-based methodologies, syntactic methodologies, and description-based methodologies for hierarchical recognition are examined in the paper.

Keywords: Algorithms, Computer vision, Human activity recognition, Event detection, Activity analysis, Video recognition.

 DOI: http://dx.doi.org/10.22159/ajpcr.2017.v10s1.19977

INTRODUCTION

Human action recognition is a dynamic point in the field of computer vision. This is because of the quickly expanding measure of video records and the huge number of potential applications taking into account programed video examination, for example, visual observation, human-machine interfaces, sports video investigation, and video recovery. Among these applications, a standout among the most fascinating is human action recognition particularly abnormal state behavior recognition. An action is a succession of human body developments and might include a few body parts simultaneously. From the perspective of computer vision, the recognition of action is to coordinate the perception (e.g., video) with beforehand characterized patterns and after that relegate it a label, i.e., action type. Contingent upon multifaceted nature, human activities can be arranged into four levels: Gestures, actions, interactions and group activities [1], and much research takes after a base up development of human movement recognition. Significant segments of such frameworks incorporate feature extraction, action learning, classification, action recognition, and segmentation [2]. A straightforward procedure comprises three stages, in particular discovery of human and/or its body parts, following, and after that recognition utilizing the following results. Case in point, to perceive "shaking hands" activities, two man's arms and hands are initially recognized and followed to produce a spatial-temporal description of their development. This description is contrasted and existing examples in the training data to decide the action sort.

This standard of classifying action recognition methods are intensely depends on the exactness of tracking, which is not solid in cluttered scenes. Numerous different systems were proposed and can be ordered by distinctive criteria as in existing survey papers. Poppe [2] examined human action recognition from picture representation and action classification independently. Weinland *et al*. [3] surveyed systems for action representation, segmentation and recognition. Turaga *et al*. [4] isolated the recognition issue energetically and action as indicated by its unpredictability, and arranged methodologies as indicated by their capacity to handle fluctuating degrees of many-sided quality. There exist numerous other classification criteria [1,5,6]. Among them, Aggarwal and Ryoo [1] are one of the most recent thorough outline and examination of the most noteworthy advancement here. In light of whether the action is perceived from information pictures specifically, Aggarwal and Ryoo [1] isolate the recognition procedures into two noteworthy classes: Single-layered methodologies and hierarchical methodologies. Both are further subarranged relying on the feature representation and learning systems, as the progress is summed and represented in Fig. 1 [1].

Fig. 1 delineates an outline of the tree-organized scientific classification that our audit takes after. We have picked a methodology based scientific classification. All action recognition techniques are initially characterized into two classifications: Single-layered methodologies and hierarchical methodologies. Single-layered methodologies are methodologies that represent and perceive human activities specifically in view of groupings of pictures. Because of their temperament, single-layered methodologies are suitable for the recognition of gestures and actions with sequential qualities. Then again, hierarchical methodologies represent abnormal state human activities by portraying them as far as other more straightforward activities, which they for the most part call sub-occasions. Recognition frameworks made out of numerous layers are developed, making them suitable for the investigation of complex actions.

Single-layered methodologies are again characterized into two sorts relying on how they display human activities: Space-time approaches and sequential methodologies. Space-time approaches view a data video as a three-dimensional (3D) (XYT) volume while sequential methodologies translate it as a grouping of perceptions. Space-time methodologies are further isolated into three classes taking into account what features they use from the 3D space-time volumes: Volumes themselves, directions, or nearby intrigue point descriptors. Sequential methodologies are characterized relying on whether they utilize exemplar based recognition techniques or model-based recognition techniques.

Fig. 2 demonstrates a nitty gritty scientific categorization utilized for single-layered methodologies secured in the audit together with various productions comparing to every classification. Hierarchical

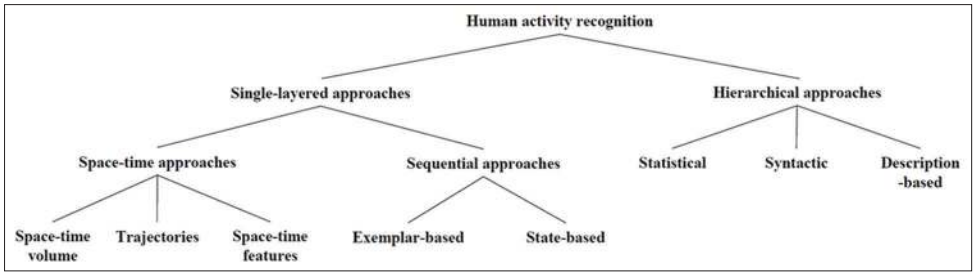

Fig. 1: The hierarchical taxonomy of various approaches for action recognition

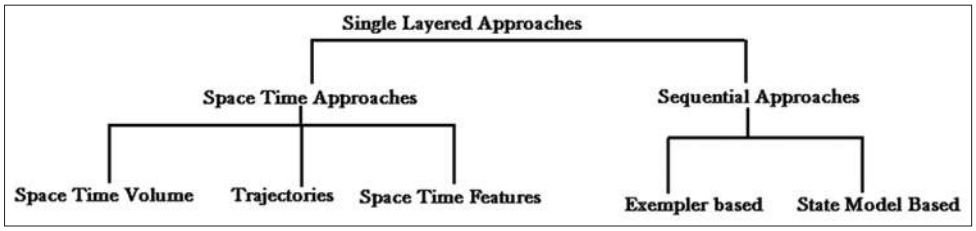

Fig. 2: Detailed taxonomical sub-classification of single-layered approaches

methodologies are grouped in view of the recognition techniques they utilize: Measurable methodologies, syntactic methodologies, and description-based methodologies. Factual methodologies build measurable state-based models linked hierarchically (e.g., layered concealed Markov models) to represent and perceive abnormal state human activities. Thus, syntactic methodologies utilize a linguistic use sentence structure, for example, stochastic context-free grammar (SCFG) to display sequential activities. Basically, they are displaying an abnormal state action as a string of nuclear level activities. Description-based methodologies represent human activities by depicting sub-occasions of the activities and their temporal, spatial, and consistent structures. Fig. 3 presents arrangements of representative distributions comparing to classes.

In this paper, we concentrate on the cutting edge research not talked about in past surveys. Furthermore, all together for an examination with past systems, we utilize a comparative scientific classification as in Aggarwal and Ryoo's survey [1]. For each of the class in Fig. 1, late improvements are given together the correlation in the middle of it and beforehand reported techniques. The rest of this paper is organized as takes after. Freely accessible datasets for human action recognition are audited in Section 2, trailed by two areas that survey recognition approaches. In Section 3, single-layered recognition methodologies are surveyed with distinctive representation and mix routines. Section 4 talks about the advances in hierarchical systems. Section 5 finishes up this survey.

DATASETS

In this segment, we talk about and portray datasets being used subsequent to 2009. Datasets that have been used sooner than 2009 can be found in Aggarwal and Ryoo's study [1] in more detail. We concentrate on new datasets gathered and we encourage break down and think about them over a few perspectives.

The KTH dataset

The present database covers six actions (strolling, running, running, boxing, hand waving, and hand applauding) performed a few times by 25 subjects in four distinct situations: Outside, outside with scale variety, outside with diverse garments, and inside. It contains a sum of 2391 groupings. All arrangements are brought with a static camera with 25 fps edge rate, down inspected to the spatial determination of 160 × 120 pixels. In the first paper [7], arrangements were isolated into a training set (eight persons), an acceptance set (eight persons), and a test set (nine persons). The dataset does not give silhouettes models and removed outlines.

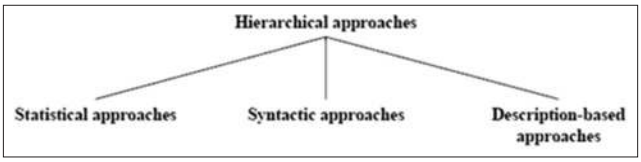

Fig. 3: Detailed taxonomical sub-classification hierarchical approaches

The Weizmann dataset

The database covers 10 normal actions (running, strolling, skipping, bouncing jack, hopping forward-on-two-legs, hopping set up on-two-legs, jogging sideways, waving-two-hands, waving one-hand, and twisting) performed by nine subjects [8]. It contains an aggregate of 93 successions. All arrangements are brought with a static camera with 25 fps edge rate, down examined to the spatial determination of 180 × 144 pixels. The dataset likewise has 10 extra groupings of strolling caught from an alternate perspective shifting somewhere around 0 and 81 in respect to the picture plane. The extricated veils after foundation subtraction and foundation groupings are given.

The INRIA xmas motion acquisition sequences (IXMAS) dataset

IXMAS covers 13 day by day life actions (checking watch, crossing arms, scratching head, taking a seat, getting up, pivoting, strolling, waving, punching, kicking, guiding, picking, overhead tossing and base up tossing) performed 3 times by 11 subjects [9]. It contains an aggregate of 2145 successions. All successions are taped with 5 aligned and synchronized free wire cameras. Dataset gives the extricated silhouettes furthermore recreated visual bodies.

CMU motion of body (MoBo) dataset

The CMU MoBo dataset covers four distinct actions (moderate strolling, quick strolling, slanted strolling, and strolling with a ball) performed by 25 subjects strolling on a treadmill in the CMU 3D room [10]. More than 8000 pictures are caught per subject. All arrangements are taken utilizing six high determination shading cameras. The groupings are 11 seconds long at 30 fps outline rate with determination of 640 × 480 pixels. The extracted silhouettes are given.

Hollywood human actions I (HOHA-I) dataset

The database contains video tests covering eight actions (noting telephone, getting out an auto, hand shaking, embracing, kissing, taking a seat, sitting up, and standing up) from 32 motion pictures [11]. The two training sets are begun from 12 motion pictures with

Table 1: Human action dataset

Dataset	Challenges	Year	Accuracy achieved (%)	Class
KTH	Homogeneous backgrounds with a static camera	2004	97.6 (Ziaeefard *et al.*'10)	General purpose action recognition
Weizmann	Partial occlusions, non-rigid deformations, significant changes in scale and viewpoint, high irregularities in the performance of an action and low quality video	2005	100 (Zhu *et al.*'09; Lin *et al.*'09; Zeng and Ji,'10)	General purpose action recognition
IXMAS	Multi view dataset for view invariant human actions	2006	89.4 (Wu *et al.*'11)	Motion acquisition
CMU MoBo	Human gait	2001	78.07 (Shi *et al.*'11)	Motion capture
HOHA	Unconstrained videos	2008	56.8 (Gilbert *et al.*'11)	Movie
HOHA-2	Comprehensive benchmark for human action recognition	2009	58.3 (Wang *et al.*'11)	Movie
Human Eva	Synchronized video and ground-truth 3D motion	2009	84.3 (Yoon *et al.*'10)	Pose estimation and motion tracking
CMU MoCap	3D marker positions and skeleton movement	2006	100 (Hu *et al.*'9)	Motion capture
UCF sports	Wide range of scenes and viewpoints	2008	93.5 (Jones *et al.*'11)	Sports action
UCF YouTube	Unconstrained videos	2008	84.2 (Wang *et al.*'11)	Sports action
i3DPost multi-view	Synchronized/uncompressed HD 8 view image sequences	2009	80 (Holte *et al.*'11)	Motion acquisition

HOHA: Hollywood human actions, IXMAS: The INRIA xmas motion acquisition sequences, 3D: Three-dimensional, MoBo: Motion of body

219 examples, and test set is started from 20 motion pictures other than utilized as a part of training with 211 specimens with names checked physically.

HOHA-II dataset

This dataset is an expansion of the HOHA dataset. The database contains video tests covering 12 actions (noting telephone, getting out an auto, hand shaking, embracing, kissing, taking a seat, sitting up, standing up, driving auto, eating, battling, and running) and 10 classes of scenes from 69 motion pictures [12]. The classes of scenes are going out, street and entering room, auto, lodging, kitchen, lounge room, office, eatery, and shop. It contains a sum of 3669 examples. The training set starts from 33 films with 823 examples. The test set begins from 36 motion pictures other than those utilized as a part of training with 884 examples having names confirmed physically.

Human EVA dataset

The human Eva-I dataset covers four dim scale video groupings and three shading video arrangements from a movement catch framework which are adjusted and synchronized with 3D body postures. The database contains 4 subjects covering 6 actions (strolling, running, signaling, finding, boxing and mix of strolling and running) [13]. The groupings are with determination of 640 × 480 pixels caught at 60 Hz. The Human Eva II dataset covers developed arrangement of mix of strolling and running actions with two subjects.

CMU MoCap dataset

The CMU Mocap dataset has six classifications (human interaction, interaction with environment lokomotion, physical activities and sports, situations, scenarios, and test motions) performed by 144 subjects. These six classifications are subdivided into 23 subcategories. The actions are caught by 12 Vicon infrared MX-40 cameras with a determination of 120 megapixel [14]. Above datasets and different datasets (UCF Sports action, UCF YouTube action, and i3DPost Multi-View) are outlined in Table 1. Also the performance of several space –time approaches are shown in Table 2

Table 2: Performance comparison of space-time approaches

Approach	Category	KTH (%)	WZMN (%)	Other (%)
Hu'09	Volume			CMU: 100
Ikizler'09	Volume	90	100	
Wang'09	Volume	91.2	100	
Guo'09	Volume	95.33		
Kim'09	Volume	95.33		Gesture: 82
Cao'09	Volume			CMU: 88.1
Liu'10	Volume	81.5	98.3	
Ziaeefard'10	Volume	97.6		
Fang'10	Volume		90.21	
Qian'10	Volume	88.69		
Kim'10	Volume	96.4		
Messing'09	Trajectory	89		Daily action: 67
Wang'11	Trajectory	94.2		HOHA2: 58.3 UCF: 88.2
Bregonzio'09	Local	93.17	96.66	
Rapantzikos'09	Local	88.3		
Minhas'10	Local	94.83	99.44	
Thi'10	Local	93.83	98.2	HOHA: 26.63 TRECVid: 23.25
Ikizler-Cinbis'10	Local			YouTube: 72.51
Yu'10	Local	95.67		UT-Itrctn: 83.33
Le'11	Local	93.9		UCF: 86.5 HOHA2: 53.3 Youtube: 75.8
Jones'12	Local	93.2		UCF: 93.5 HOHA: 48.4
Sadek'11	Local	93.6	97.8	
Gilbert'09	Local	94.5		HOHA: 31.4 mKTH: 68.8
Oikonomopoulos	Local	81	92	Aerobics: 95
Lui'11	Local	97		UCF: 88

HOHA: Hollywood human actions

SINGLE-LAYERED APPROACHES

This segment surveys the single-layered methodologies as shown in Fig. 4. The strategies are described by the activities to be perceived specifically from the crude video data rather than primitive sub-actions or sub-activities. Subsequently, most single layered methodologies manage basic video or datasets, for example, KTH to perceive the actions contained. The picture arrangements from recordings are viewed as being produced from a particular class of actions, and consequently, such methodologies essentially include how to represent the recordings (i.e., extricating features) and coordinate them. All things considered, single-layered methodologies essentially perceive common actions and these perceived straightforward primitive actions can be utilized to identify more intricate action recognition utilizing hierarchical blends, as examined in Section 4.

As appeared in a past survey [1], different methodologies have been proposed for representation and coordinating in single-layered frameworks. They can be extensively arranged into two classes: Space-time approaches and sequential methodologies. The center contrast between space-time and sequential methodologies is the manner by which the temporal measurement (i.e., the third-measurement in a 3D XYT space) is dealt with. Space-time approaches regard time as a customary measurement as spatial measurements and separate features from the 3D volumetric recordings, while sequential methodologies

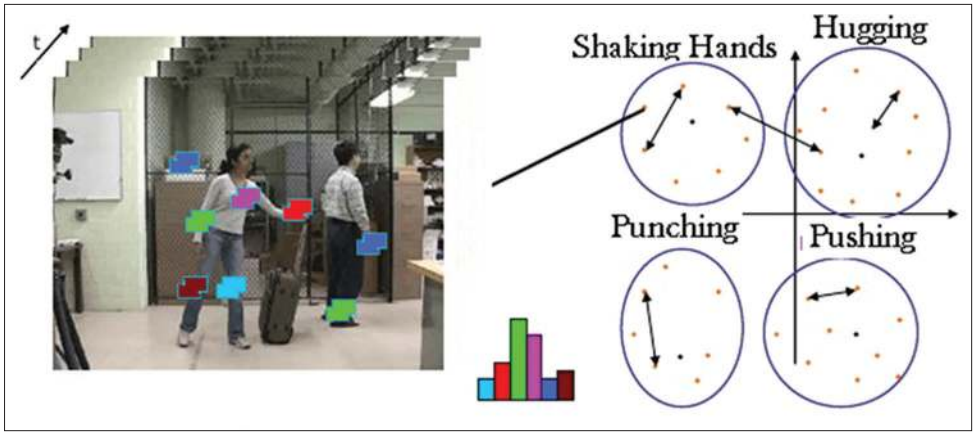

Fig. 4: Sample illustration of a basic space time approach for action recognition

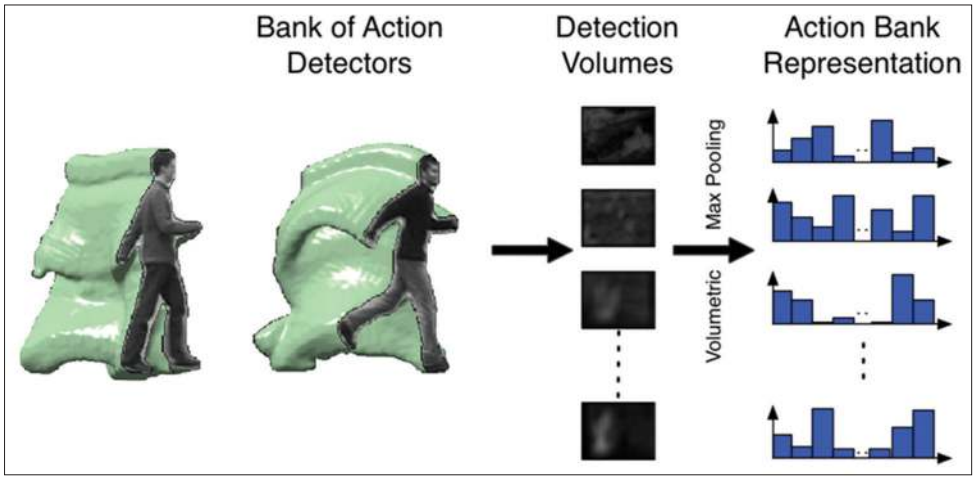

Fig. 5: Sample Illustration of a basic space-time volume approach for action recognition

consider a human action as requested perceptions along the timeline. Since they think about sequential connections, sequential methodologies by and large accomplish preferred results over its space-time partner.

In the following sub-section, we introduce an audit to the latest advancement in this branch of action recognition and made correlation among them and past surveyed strategies. Space-time methodologies are examined in Section 3.1 and sequential methodologies in Section 3.2.

Advances in space-time approaches

For most action recognition frameworks (additionally the extent of this survey), the data are from recordings. All recordings examined here comprise a temporal (T) arrangement of two-dimensional (2D) spatial (XY) pictures or proportionally an arrangement of pixels in 3D XYT space. In this manner, a video can be represented as a spatial-temporal volume, and this volume contains important data for human creatures and machines to perceive the actions and activities in the volume. In view of this suspicion, different representation and correspondence coordinating calculations have been advanced to minimalistically describe the fundamental movement designs. As appeared in Fig. 1, we talk about the advancement of space-time approaches utilizing the same representation-based scientific classification. Aside from systems utilizing the crude volume as a feature, every one of the three representations use movement related data to portray the actions or activities as shown in Fig. 5.

Action recognition with space-time volumes

The most instinctive space-time volume methodology would utilize the whole 3D volume as feature or layout, and match obscure action

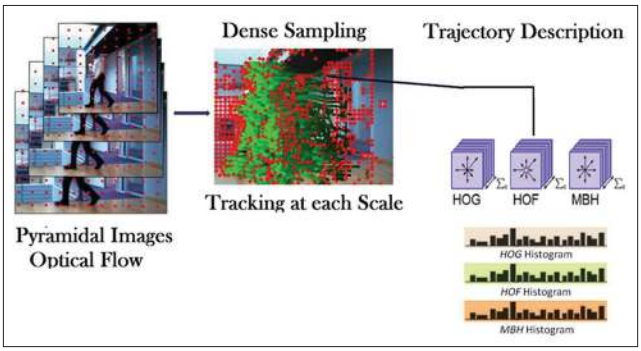

Fig. 6: Sample illustration of a basic space time trajectory approach for action recognition

recordings to existing ones to acquire the classification, as shown in Figs. 6 and 7. Nonetheless, the system experiences the clamor and good for nothing foundation data, and in this way, some exertion has been made to show the closer view development.

In view of Bobick and Davis' [15] take a shot at development, different methodologies have been investigated to expand it for action recognition. Hu *et al.* [16] proposed to consolidate both motion history image (MHI) and appearance data for better portrayal of human actions. Two sorts of appearance-based features were proposed. The main appearance-based feature is the forefront image, acquired by foundation subtraction. The second is the histogram of oriented gradients feature

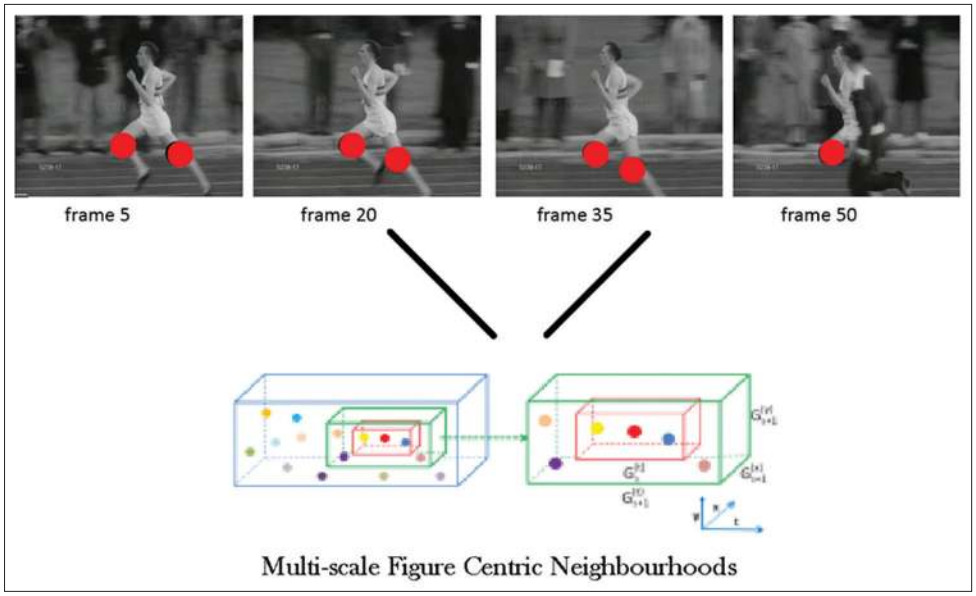

Fig. 7: Sample illustration of a basic space time local features approach for action recognition

(HOG), which portrays the headings and extents of edges and corners. Grin support vector machine (SVM) (simulated annealing multiple instance learning [MIL] SVMs) was proposed for classification. It plans to acquire a global ideal through simulated annealing system without depending on model introduction to maintain a strategic distance from neighborhood minima. Qian *et al*. [17] joined global features and nearby features to order and perceive human activities. The global feature depended on paired motion energy image (MEI) and its form coding of the MEI was utilized rather than MEI as a superior global feature in light of the fact that it defeats the impediment of MEI where hollows exist for parts of human blob are undetected. For nearby features, an item's jumping box was utilized. The feature focuses were grouped utilizing multi-class SVMs. Roh *et al*. [18] additionally extended Bobick and Davis' [15] MHI from 2D to 3D space and proposed volume motion format for perspective autonomous human action recognition utilizing stereo recordings.

Correspondingly, roused by a stride energy image [19], Kim *et al*. [20] proposed a collected motion image (AMI) to represent spatiotemporal features of happening actions. The AMI was the normal of image contrasts. A rank lattice was acquired utilizing ordinal estimation of AMI pixels. The separation between rank lattices of question video and hopeful video was registered utilizing L1-standards, and the best match, spatially and temporally, was the competitor with the base separation.

Different researchers attempted to fuse individual models, for example, outlines or skeletons for action recognition. Ikizler and Duygulu [21] proposed another posture descriptor called histogram of oriented rectangles (HOR) for action recognition. They represented every human posture in an action succession with oriented rectangular patches separated over the human outline, which then framed spatial oriented histograms to represent the circulation of these rectangular patches. The nearby progress was caught with the summation of the HOR inside of a sliding window. Four coordinating routines were performed for classification, to be specific closest neighbor, global histogramming, SVM, and element time twisting.

Fang *et al*. [22] initially mapped the high dimensional outlines to low dimensional focuses as spatial motion description utilizing territory saving projection. This low-dimensional motion vector was accepted to depict the natural motion structure. At that point, three distinctive temporal data - i.e., temporal neighbor, motion distinction, and motion direction - were connected to the spatial descriptors to acquire the feature vectors, which were sustained with k-closest neighborhood classifier.

Ziaeefard and Ebrahimnezhad [23] proposed the cumulative skeletonized image (CSI) crosswise over time as features, and built 2D rakish/separation histograms in light of it. A hierarchical SVM was utilized for the coordinating procedure. Initial a coarse classification of CSI histograms utilizing a SVM classifier was gotten with unique actions, and afterward the second SVM was connected to befuddled actions utilizing remarkable features among comparative actions. Wang and Mori [24] proposed semi latent topic models (STM) taking after the sack of-words structure, where a "word" relates to an edge and an "archive" compares to a "video grouping." Subsequent to acquiring settled persons in a video arrangement, optical stream was figured, and half-wave amended into four channels took after by sifting to frame the motion descriptor, in view of which codebook was built. Taking into account latent topics models, for example, latent Dirichlet allocation (LDA) [25] and correlated topic model [26], segmented topic model does not require a decision for the quantity of latent topics, yet gave better training productivity and recognition exactness.

Guo *et al*. [27] saw an action as a temporal succession of nearby shape-distortions of centroid-focused item outlines. Every action was represented by the exact covariance network of an arrangement of 13-dimensional standardized geometric feature vectors that caught the state of the outline burrow. The similitude of two actions was measured as far as a Riemannian metric between their covariance frameworks. The outline passage of a test video is broken into short covering portions, and every section was arranged utilizing a word reference of marked action covariance networks and the closest neighbor principle.

Efforts in other directions have also occurred. Kim and Cipolla [28] extended canonical correlation analysis (CCA) to measure video-to-video similarity. The method acted on video volumes avoiding the difficult problems of explicit motion estimation and provided a way of spatiotemporal matching that is robust to intraclass variations of action due to CCA. Liu and Yuen [29] applied principal component analysis (PCA) to a salient action unit (i.e., one cycle of repetitive action in a video), and AdaBoost classifier was used to classify the action in a query video. Cao *et al*. [30] provided a new way to combine different features using a heterogeneous feature machine.

Action recognition with space-time trajectories

Trajectory construct methodologies are situated in light of the perception that the tracking of joint positions is adequate for humans to perceive actions [31]. Directions are typically built by tracking joint focuses or other interest focuses on human body. Different representations and relating calculations coordinate the directions for action recognition.

Messing and Kautz [32] removed feature directions by tracking Harris3D interest focuses utilizing a KLT tracker [33], and the directions were represented as groupings of log-polar quantized speeds. It utilized a generative blend model to take in a speed history dialect and grouped video arrangements. A weighted blend of sacks of enlarged direction groupings was demonstrated for action classes. These blend components can be considered as speed history words, with every speed history feature being created by one blend component, and every action class has a conveyance over these blend components. Further, they demonstrated how the speed history feature can be developed, both with a more refined latent speed model, and by joining the speed history feature with other valuable data, similar to appearance, position, and abnormal state semantic data.

Wang *et al*. [34] proposed a way to deal with portray recordings by thick directions. They inspected thick focuses from every edge and followed them in light of relocation data from a thick optical stream field. Neighborhood descriptors of HOG, histograms of optical flow, and motion boundary histogram (motion limit histogram) around interest focuses were processed.

Action recognition with space-time local features

The use of neighborhood features in real life recognition was stretched out from article recognition in images. The nearby features allude to the description of focuses and their surroundings in the 3D volumetric data with one of the kind discriminative attributes. These focuses and comparing neighborhood feature descriptors are most enlightening and more powerful. As far as the thickness of extricated feature focuses, the representation of nearby feature methodologies can be partitioned into two general classifications: Inadequate and thick. The Harris 3D identifier [35], and the Dollar *et al*. indicator [36] are representative of the previous, and optical stream based routines the recent. Most calculations are gotten from them. Other novel systems have additionally been connected for finding interest focuses to perceive actions.

Bregonzio *et al*. [37] proposed billows of space-time premium focuses to conquer the impediments of the Dollar *et al*. finder [36]. Utilizing the recognized interest focuses from Dollar *et al*.'s study [36], this was accomplished through separating all encompassing features from billows of interest focuses gathered over multiple temporal scales took after via programed feature determination. SVMs and Nearest Neighbor Classifiers were utilized for classification. One sample of billows of interest focuses. Jones *et al*. [38] additionally construct their research with respect to the Dollar *et al*. indicator [36] to identify and portray premium focuses which were then grouped utilizing k-implies. The advancement is that it consolidated importance input component by utilizing asymmetric bagging and random subspace SVM.

In Thi *et al*.'s study [39], space-time interest focuses are recognized with the Harris3D identifier [35], and appointed names; demonstrating on the off chance that it fits in with the class of interest action by utilizing a Bayesian classifier. The feature vectors of interest point descriptors and names are then given to a PCA-SVM classifier to perceive the action sort. In this work, the action is likewise confined taking into account condition random fields (CRFs) weighting results.

While 3D Harris corners [35] are generally utilized, they endure the issue of sparity. Gilbert *et al.* [40] utilized thick straightforward 2D Harris corners [41] in multiple scales to build features. A two-stage hierarchical grouping procedure was utilized to order features and the actions. Sadek *et al*. [42] additionally utilized a Harris corner locator as a part of every casing and depicted the neighborhood feature focuses with temporal self-likenesses characterized on the fluffy log-polar histograms. Together with global features (i.e., change of gravity focuses), the feature vectors were grouped with SVM. Optical stream is additionally commonly utilized for feature point identification and description [43-45]. Ikizler-Cinbis and Sclaroff [44] utilized optical stream and frontal area stream to concentrate motion features for persons, articles and scenes, taking into account which the shape feature for each was additionally removed. These feature channels were inputs to a MIL system to discover the area of enthusiasm for a given video.

Holte *et al*. [43] developed 3D optical stream from eight weighted 2D stream fields to accomplish view-invariant action recognition. 3D motion context (3D-MC) and harmonic MC (HMC) were utilized to represent the removed 3D motion vector fields effectively and in a perspective invariant way. The subsequent 3D-MC and HMC descriptors were grouped into an arrangement of human actions utilizing standardized relationship, considering the performing speed varieties of diverse on-screen characters. Another optical stream based work was Oikonomopoulos' B-spline polynomial descriptor [45]. It was removed as spatiotemporal salient focuses recognized on the evaluated optical stream field for a given image arrangement and depended on geometrical properties of 3D piecewise polynomials, to be specific B-splines. The last was fitted on the spatio-temporal areas of salient focuses that fell inside of a given spatiotemporal neighborhood. The descriptor is invariant in interpretation and scaling in space-time.

Numerous endeavors have been made to discover interest focuses with different standards [46-52]. For instance, Rapantzikos *et al*. [49] proposed a saliency-based interest focuses identifier which consolidates power, shading, and motion. It utilized a multi-scale volumetric representation of the video and included spatiotemporal operations at the voxel level. Interest focuses were chosen as the extreme of the saliency response. Distinctive recognition calculations were utilized, for example, pack of-words with closest neighbor for the KTH dataset and 2 SVM part for HOHA dataset.

Minhas *et al*. [48] proposed new strategies to process the spatiotemporal features utilizing 3D dual-tree discrete wavelet transform. 3D DTDWT was utilized to get the spatiotemporal data (subband vector of wavelet coefficients) productively, and a relative SIFT was utilized for nearby static features. By utilizing mixture spatiotemporal and neighborhood static features, the extreme learning machine classifier came to high exactness for open datasets.

Yu *et al*. [51] presented a structure in view of semantic texton forests (STFs) to accomplish continuous action recognition. The FAST indicator [53] was reached out to V-FAST for video interest point identification. STFs are connected to group neighborhood space-time volumes around interest focuses to create the discriminative codebook. Pyramidal spatiotemporal relationship match (PSRM) was utilized for neighborhood appearance and auxiliary data. An arrangement of 3D relationship histograms were built by investigating each pair of feature focuses utilizing PSRM.

Zhu *et al*. [52] proposed another temporally integrated spatial response (TISR) descriptor, which caught the qualities of individual actions by removing thick spatiotemporal descriptors and representing actions by sack of-words features. With a visual vocabulary of the TISR descriptors, the sack of-words histogram features could endure spatial and temporal varieties.

Le *et al*. [46] introduced an augmentation of the independent subspace analysis (ISA) calculation to take in invariant spatiotemporal features from unlabeled video data hierarchically. All the more particularly, features were first learnt with little information patches swelled into a vector, convolved with a bigger area of the info data, and then utilized as information to the layer above. The features from both layers were

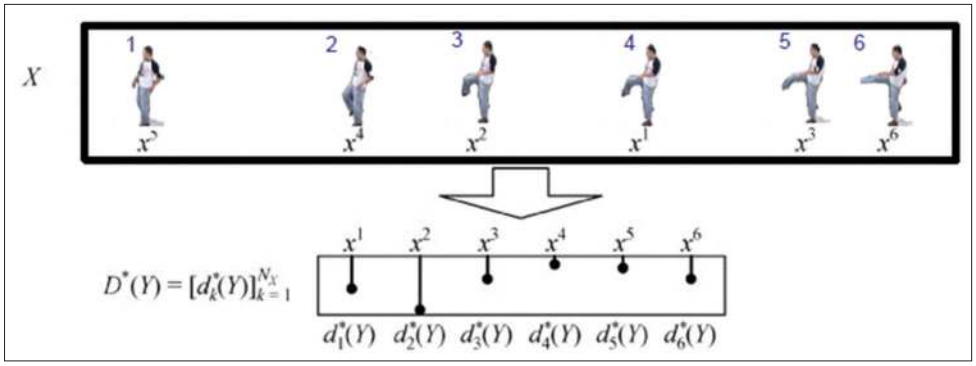

Fig. 8: Sample illustration of a exemplar based approach for action recognition

consolidated as nearby features for classification. This two-layered stacked convolutional ISA model beats the impediment of ISA for vast inputs and performed well on testing datasets.

Sequential approaches

Single-layered sequential methodologies vary with space-time approaches in that they are intended to catch temporal relationships of perceptions. In this way, human actions are integrated as an arrangement of perceptions. For the most part, a perception is connected with neighborhood or global features separated from an edge or an arrangement of casings. As in Aggarwal and Ryoo's study [1], exemplar-based recognition and state model-based analysis are two sub-classes of sequential methodologies as shown in Figs 8 and 9.

Exemplar-based approaches

As we specified before, sequential methodologies characterize actions to be a succession of perceptions and how perceptions are extricated is not restricted. Exemplar-based methodologies represent human actions with a format arrangement of perception or an arrangement of test grouping of action perceptions. Subsequently, the center of exemplar-based methodologies is characterizing how another data video can be contrasted and the format or test succession of action perceptions. In past work, dynamic time warping (DTW) has been generally received for exemplar-based human action recognition in Darrell and Pentland; Gavrila and Davis; Veeraraghavan and Roy-Chowdhury's study [54-56].

The likeness in the middle of information and action layout is measured by looking at coefficients of the action premise after PCA in Yacoob and Black's study [57]. Dynamic feature changes are likewise used to represent a movement as a linear time invariant framework [58]. As of late Lin *et al*. [59] represented actions in recordings as a succession of models. The model depends on a novel shape-motion feature and the matching so as to group is created with a hierarchical model tree developed utilizing K-implies (K=2) bunching connected iteratively. Given an action video, model arrangement will be produced for it with a model grouping estimation. The model matching was satisfied utilizing FastDTW algorithm to increment computational effectiveness.

State model-based approaches

Rather than representing human action as a succession of perceptions state model-based methodologies take in a state model for every action and every action is represented as far as an arrangement of concealed states. It produces groupings of perception and each succession of perception is connected with an instance of the relating action. Standard concealed Markov models have been broadly utilized for state model-based methodologies in Bobick and Wilson, Starner *et al*., and Yamato *et al*.'s study [60-62]. Gee are additionally stretched out to coupled hidden semi-Markov models to model length of time of human activities [63,64].

As of now, hidden Markov models (HMMs) or expansions are still connected in human action recognition. In Yu and Aggarwal's study [65],

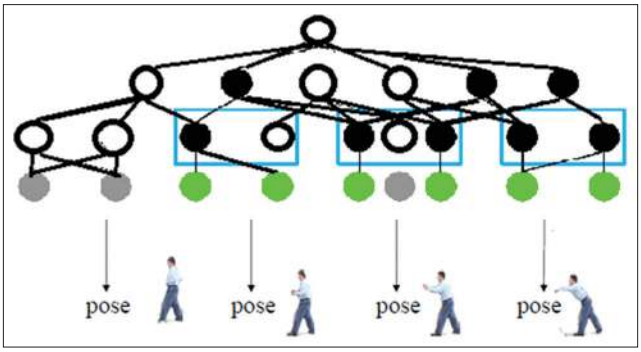

Fig. 9: Sample illustration of state model based approach for action recognition

an adaptable star skeleton is depicted for use in stance representation. The point is to precisely match human limits utilizing shapes and histograms from an image outline. A HMM is used to perceive human actions. In Kellokumpu *et al*.'s study [66], novel composition descriptors are proposed to portray motion and a HMM is utilized to show the temporal improvement of surface motion histograms. In Shi *et al*.'s study [67], a discriminative semi-Markov model methodology is proposed and with a specific end goal to effectively take care of the induction issue of at the same time portioning and perceiving distinctive actions they outlined a Viterbi like dynamic programming algorithm. Examination of sequential methodologies can be found in Table 3.

HIERARCHICAL APPROACHES

As depicted in Aggarwal and Ryoo's study [1], hierarchical methodologies attempt to perceive intriguing occasions (abnormal state activities) in view of more straightforward or low-level sub-activities. As it were an abnormal state action can be deteriorated into a succession of a few sub-activities, for example, "hand shaking" might be integrated as an arrangement of two hands being expanded, converging into one item, and two hands being pulled back. Sub-activities can be further considered as abnormal state activities until deteriorated into nuclear ones.

The upside of hierarchical methodologies is the ability to display the perplexing structure of human activities and its adaptability for either individual activities, interaction in the middle of humans and/or protests or group activities. In addition, hierarchical models give a natural and helpful interface for incorporating former information and understanding of structure of activities. Hierarchical ways to deal with some degree have a cozy relationship with single layer methodologies. For instance non-hierarchical single layer methodologies can be effortlessly used for low-level or nuclear action recognition, for example, motion location. Some non-hierarchical single layer methodologies can likewise be stretched out to hierarchical models, for example, expanded multi-layered HMMs. Utilizing the scientific classification proposed as a part of Aggarwal and Ryoo's study [1], hierarchical methodologies

Table 3: Comparison of sequential approaches

Approach	Category	KTH (%)	WZMN (%)	Other (%)
Shi'11	State-based	95		CMU: 78 WBD: 94
Yu'09	State-based			Human climbing fences: 97.9 Ballet movie: 93.6
Kellokumpu'09	State-based	93.8	98	
Lin'09	Exemplar	95.77	100	

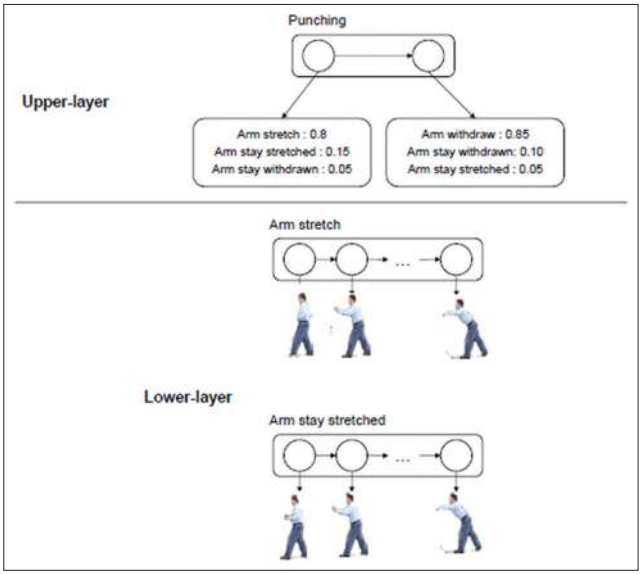

Fig. 10: Sample illustration of statistical approach for action recognition

are ordered into three groups: Measurable methodologies, syntactic methodologies, and description-based methodologies.

Statistical approaches

HMMs can be considered as a straightforward instance of dynamic Bayesian networks (DBNs), a shown in Fig. 10. A HMM represents the condition of the world utilizing a solitary discrete random variable however DBN represents the condition of the world utilizing an arrangement of random variables. Multiple levels of hidden states shape a representation of hierarchical human activities. Past research endeavors on factual methodologies for the most part harp on utilizations of augmented HMMs and DBNs: Two-layered hierarchical HMMs [68-70] and dynamic probabilistic networks otherwise called DBNs [71,72]. Sub-activities can be either simultaneous or sequential. Well based methodologies in the writing handle sequential sub-activities. Along these lines, a hierarchical methodology utilizing an engendering system (P-net) [73] has been proposed to handle both simultaneous and sequential sub-activities. Past HMMs and DBNs another four-layered hierarchical probabilistic latent model are proposed in Yin and Meng's study [74]. To begin with the spatial-temporal features are identified and grouped utilizing hierarchical Bayesian model to frame nuclear actions. At that point, in light of LDA, a hierarchical probabilistic latent model is utilized to recognition the action without the need to determine the quantity of latent states. Neighborhood feature-spatial-temporal features are used rather than global feature, for example, human motion. It is an endeavor to use grouped space-time features as nuclear actions and hierarchical descriptions and representations of complex actions.

Another measurable methodology [75] is to deteriorate the body into a hierarchical structure. A hierarchical complex space is learnt to portray the motion designs. Course CRFs are utilized to anticipate these motion designs. SVMs are utilized to order last human actions in view of the motion designs. Hierarchical representation of human action is proposed as opposed to straightforward non-hierarchical pack of-words representation. In Mauthner *et al.*'s study [76], hierarchical K-implies tree is additionally used to represent the feature signs. The issue of inadequate integrating so as to train data is handled in Zeng and Ji's study [77] with area information. To begin with request rationale based area information is abused for dynamic Bayesian system learning, both the structure and the parameters.

Syntactic approaches

Syntactic methodologies incorporate actions as a series of images. An image in this context is really the nuclear sub-activities said in the past area. Nuclear sub-activities can be perceived utilizing any of the past hierarchical or non-hierarchical systems. However, actions represented as a series of images results in a constraint for simultaneous action recognition. In past work, context-free grammers (CFGs), in view of syntactic methodologies, have been contemplated and connected in human action recognition. A few probabilistic augmentation of CFGs; SCFGs are presented in Ivanov and Bobick, Joo and Chellappa, Minnen *et al*., and Moore and Essa's study [78-81]. For the most part two-layer structures are proposed; the lower layer for the most part capacities to perceive nuclear or low-level actions and the higher layer uses parsing strategies for the abnormal state action recognition.

Another impediment is that client must give an arrangement of creation tenets and so as to overcome such constraints Kitani *et al*. [82] acquainted an algorithm with naturally take in tenets from perceptions. As of late endeavors have been made towards another hierarchical structure. In Kitani *et al*.'s study [83], a four-level order is proposed. Actions are represented by an arrangement of grammar standards sorted into three classes; solid, feeble, and stochastic relations in view of spatio-temporal relations.

Description-based approaches

Description-based methodologies vary from measurable and syntactic methodologies through a capacity to unequivocally express human activities' spatiotemporal structures. In this manner, such routines can perceive both sequential and simultaneous actions rather being constrained to sequential actions. Essentially, description-based methodologies model human activities as an event of implanted sub-activities. Such events must fulfill determined temporal, spatial and legitimate relationships that are signatory of an abnormal state action. Subsequent to the presentation of Allen's temporal interim predicates, they have been embraced for description-based human movement recognition for both sequential and simultaneous relationships. Context free grammars have additionally been used for description-based methodologies. A formal grammar is required for the representation of human activities as in Nevatia *et al*. and Ryoo and Aggarwal's study [84,85].

Transformation from Allen's interim variable based math limitation system to a proprioceptive neuromuscular facilitation-system is proposed in Pinhanez and Bobick's study [86] to portray indistinguishable temporal data. The change accomplishes a structure that is computationally tractable. Bayesian conviction networks and Petri nets are presented, individually, in Intille and Bobick, Ghanem *et al*. [87,88]. Occasion rationale is depicted by Siskind to perceive abnormal state activities in Siskind's study [89]. In request to make up for the disappointments of its low-level components because of

Table 4: Comparison of hierarchical approaches

Approach	Category	KTH (%)	WZMN (%)	Other (%)
Yin'10	Statistical	82		
Zeng'10	Statistical	92.1	100	
Han'10	Statistical			CMU: 98.27
Wang'11	Statistical	92.5		HOHA: 37.6 UCF: 68.3
Ijsselmuiden'10	Description-based			Group activities: 74.4
Morariu'09	Description-based			Basketball: 72

the deterministic attributes of description based methodologies a few probabilistic expansions of the recognition systems are proposed in Aggarwal and Ryoo, Gupta *et al*. [90,91]. Symbolic computerized reasoning procedures Markov logic networks (MLN) was additionally received to induce fascinating activities probabilistically as in Tran and Davis's study [92].

Ijsselmuiden and Stiefelhagen [93] give a brief system to abnormal state human movement recognition. It consolidates diverse info sources and depends on temporal rationale. No probabilistic calculation is utilized in this work. As of late a structure was proposed in Morariu and Davis's study [94], to perceive behavior in balanced b-ball by method for self-assertive directions acquired by tracking the ball, hands, and feet. This system utilizes video analysis and blended probabilistic and legitimate induction to expound occasions. The technique requires semantic descriptions of what by and large happens in different situations. To start with request rationale in light of Allen's interval logic is used to encode spatiotemporal structure learning and MLN is utilized to handle vulnerability low-level perception. Albeit, much exertion has been stretched out as depicted already however common standard dataset has not been used to certain degree so that correlation between description-based methodologies can be communicated regarding practically rather than factually. Correlation between hierarchical methodologies is appeared in Table 4.

CONCLUSION

In this study, we give a survey of advances in automated human action recognition. A substantial gathering of techniques are recognized. Among them, 50 particular and powerful recommendations of the most recent 3 years are accounted for. The examination utilizes the same scientific classification as a past survey taking into account whether the action is perceived straightforwardly from the images or low-level sub-actions. Our objective was to cover the best in class improvements in every classification, together with the datasets utilized as a part of approval. The writing surveyed demonstrates that much research has been committed to recognition of human actions specifically from the recordings or images in a solitary layered way. This is particularly valid for the case utilizing space-time volume and neighborhood features. It is regular to amplify 2D image preparing routines, for example, interest point identification, to 3D recordings to concentrate feature descriptors. In the interim, more researchers are starting to investigate routines for abnormal state action recognition. For this situation, most strategies surveyed utilize a hierarchical methodology, taking into account factual, syntactic, or description-based routines to clarify and construe activities from low-level occasions. Especially, it is of enthusiasm to consolidate the formal descriptors and probabilistic thinking to translate human actions, for example, done in Nevatia *et al*.; Ryoo and Aggarwal; Siskind [84,85,89]. While some research has concentrated on complex genuine actions, most prominent test datasets are still basic, obliged, and organized situations. The presentation of more practical datasets, for example, Hollywood films and YouTube recordings are testing. The precision reported is low in the writing surveyed here. In view of the aftereffects of low-level actions, we trust more research will be done in the zone of abnormal state action recognition in datasets and genuine scenes. We know, in any case, that finish audit of all the methodologies is far-off. As a well-known research topic, human action and movement recognition has pulled in much consideration and will stay critical. With more application fields being investigated, on one side, area particular systems will most likely rise. On the other side, a cross-area system would be helpful to the whole community.

REFERENCES

1. Aggarwal J, Ryoo M. Human activity analysis: A survey. ACM Comput Surv 2011;43:1-43.
2. Poppe R. A survey on vision-based human action recognition. Image Vis Comput 2010;28:976-90.
3. Weinland D, Ronfard R, Boyer E. A survey of vision-based methods for action representation, segmentation and recognition. Comput Vis Image Underst 2011;115:224-41.
4. Turaga P, Chellappa R, Subrahmanian VS, Udrea O. Machine recognition of human activities: A survey. IEEE Trans Circuits System Video Technol 2008;18:1473-88.
5. Candamo J, Shreve M, Goldgof DB, Sapper DB, Kasturi R. Understanding transit scenes: A survey on human behavior recognition algorithms. IEEE Trans Intell Transp Syst 2010;11:206-24.
6. Chaudhary A, Raheja JL, Das K, Raheja S. A survey on hand gesture recognition in context of soft computing. In: Meghanathan N, Kaushik BK, Nagamalai D, editors. Advanced Computing. Berlin: Springer; 2011. p. 46-55.
7. Schuldt C, Barbara I. Recognizing Human Actions: A local SVM Approach. IEEE Computer Society; 2004.
8. Blank M, Gorelick L, Shechtman E, Irani M, Basri R. Actions as space-time shapes. In: IEEE International Conference on Computer Vision (ICCV); 2005. p. 1395-402.
9. Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst 2006;104:249-57.
10. Gross R, Shi J. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18. Pittsburgh, PA: Robotics Institute; 2001.
11. Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2008.
12. Marszalek M, Laptev I, Schmid C. Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009.
13. Sigal L, Black MJ. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Int J Comput Vis 2006;87:4.
14. University CM. CMU graphics lab Motion Capture Database; 2006. Available from: http://www.mocap.cs.cmu.edu.Technical Report.
15. Bobick AF, Davis JW. The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 2001;23:257-67.
16. Hu Y, Cao L, Lv F, Yan S, Gong Y, Huang T. Action detection in complex scenes with spatial and temporal ambiguities. In: IEEE International Conference on Computer Vision (ICCV); 2009. p. 128-35.
17. Qian H, Mao Y, Xiang W, Wang Z. Recognition of human activities using SVM multi-class classifier. Pattern Recognit Lett 2010;31:100-11.
18. Roh MC, Shin HK, Lee SW. View-independent human action recognition with volume motion template on single stereo camera. Pattern Recognit Lett 2010;31:639-47.
19. Han J, Bhanu B. Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 2006;28:316-22.
20. Kim W, Lee J, Kim M, Oh D, Kim C. Human action recognition using ordinal measure of accumulated motion. EURASIP J Adv Signal Process 2010;2010:1-12.
21. Ikizler N, Duygulu P. Histogram of oriented rectangles: A new pose descriptor for human action recognition. Image Vis Comput 2009;27:1515-26.

22. Fang CH, Chen JC, Tseng CC, Lien JJ. Human Action Recognition Using Spatio-Temporal Classification; 2010. p. 98-109.
23. Ziaeefard M, Ebrahimnezhad H. Hierarchical human action recognition by normalized-polar histogram. In: International Conference on Pattern Recognition (ICPR); 2010. p. 3720-3.
24. Wang Y, Mori G. Human action recognition by Semi latent topic models. IEEE Trans Pattern Anal Mach Intell 2009;31:1762-74.
25. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022.
26. Blei D, Lafferty J. Correlated topic models. Adv Neural Inf Process Syst 2006;18:147.
27. Guo K, Ishwar P, Konrad J. Action recognition in video by covariance matching of silhouette tunnels. In: The 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing; 2009. p. 299-306.
28. Kim TK, Cipolla R. Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE Trans Pattern Anal Mach Intell 2009;31:1415-28.
29. Liu C, Yuen PC. Human action recognition using boosted Eigen actions. Image Vis Comput 2010;28:825-35.
30. Cao L, Luo J, Liang F, Huang TS. Heterogeneous feature machines for visual recognition. In: International Conference on Computer Vision (ICCV); 2009. p. 1095-102.
31. Johansson G. Visual motion perception. Sci Am 1975;232:76-88.
32. Messing R, Kautz H. Activity recognition using the velocity histories of tracked key points. In: IEEE International Conference on Computer Vision (CVPR); 2009. p. 104-11.
33. Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. In: The 7th International Joint Conference on Artificial Intelligence. Vol. 2; 1981. p. 674-9.
34. Wang H, Kläser A, Schmid C, Cheng-Lin L. Action recognition by dense trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). USA: Colorado Springs; 2011. p. 3169-76.
35. Laptev I, Lindeberg T. Space-time interest points. In: IEEE International Conference on Computer Vision (ICCV); 2003. p. 432-9.
36. Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance; 2005.
37. Bregonzio M, Gong S, Xiang T. Recognising function as clouds of space-time interest points. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on; 2009. p. 1948-55.
38. Jones S, Shao L, Zhang J, Liu Y. Relevance feedback for real-world human action retrieval. Pattern Recognit Lett 2012;33:446-52.
39. Thi TH, Zhang J, Cheng L, Wang L, Satoh S. Human action recognition and localization in video using structured learning of local space-time features. In: IEEE International Conference on Advanced Video and Signal Based Surveillance; 2010. p. 204-11.
40. Gilbert A, Illingworth J, Bowden R. Fast realistic multi action recognition using mined dense spatio-temporal features. In: IEEE International Conference on Computer Vision (ICCV); 2009. p. 925-31.
41. Harris C, Stephens M. A combined corner and edge detector. In: Alvey Vision Conference; 1988. p. 189-92.
42. Sadek S, Al-Hamadi A, Michaelis B, Sayed U. An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity. EURASIP J Adv Signal Process 2011.
43. Holte MB, Moeslund TB, Nikolaidis N, Pitas I. 3D human action recognition for multi-view camera systems. In: International Conference on 3D Imaging, Modeling, Processing and Transmission; 2011. p. 342-9.
44. Ikizler-Cinbis N, Sclaroff S. Object, scene and actions: Combining multiple features for human action recognition. In: European Conference on Computer vision (ECCV): Part I; 2010. p. 494-507.
45. Oikonomopoulos A, Pantic M, Patras I. Sparse b-spline polynomial descriptors for human activity recognition. Image Vis Comput 2009;27:1814-25.
46. Le QV, Zou WY, Yeung SY, Ng AY. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2011. p. 3361-8.
47. Lui YM, Beveridge JR. Tangent bundle for human action recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition; 2011. p. 97-102.
48. Minhas R, Baradarani A, Seifzadeh S, Wu QM. Human action recognition using extreme learning machine based on visual vocabularies. Neurocomputing 2010;73:1906-17.
49. Rapantzikos K, Avrithis Y, Kollias S. Dense Saliency-Based Spatiotemporal Feature Points for Action Recognition; 2009.
50. Shao L, Ji L, Liu Y, Zhang J. Human action segmentation and recognition via motion and shape analysis. Pattern Recognit Lett 2012;33:438-45.
51. Yu TH, Kim TK, Cipolla R. Real-time action recognition by spatiotemporal semantic and structural forest. In: Proceedings of the British Machine Vision Conference (BMVC); 2010. p. 52.1-12.
52. Zhu G, Yang M, Yu K, Xu W, Gong Y. Detecting video events based on action recognition in complex scenes using spatiotemporal descriptor. In: 17th ACM International Conference on Multimedia; 2009. p. 165-74.
53. Rosten E, Drummond T. Machine learning for high-speed corner detection. In: European Conference on Computer Vision (ECCV); 2006. p. 430-43.
54. Darrell T, Pentland A. Space-time gestures. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 1993. p. 335-40.
55. Gavrila DM, Davis LS. Towards 3-D model-based tracking and recognition of human movement: A multi-view approach. In: In International Workshop on Automatic Face - And Gesture-Recognition. IEEE Computer Society; 1995. p. 272-7.
56. Veeraraghavan A, Roy-Chowdhury AK. The function space of an activity. In: In Proceedings Computer Vision Pattern Recognition; 2007. p. 959-68.
57. Yacoob Y, Black MJ. Parameterized modeling and recognition of activities. In: IEEE International Conference on Computer Vision (ICCV); 1998. p. 120-7.
58. Lublinerman R, Ozay N, Zarpalas D, Camps O. Activity recognition from silhouettes using linear systems and model invalidation techniques. In: International Conference on Pattern Recognition (ICPR); 2006. p. 347-50
59. Lin Z, Jiang Z, Davis LS. Recognizing actions by shape motion prototype trees. In: IEEE International Conference on Computer Vision; 2009. p. 444-51.
60. Bobick AF, Wilson AD. A state-based approach to the representation and recognition of gesture. IEEE Trans Pattern Anal Mach Intell 1997;19:1325-37.
61. Starner T, Weaver J, Pentland A. Real-time American sign language recognition using desk and wearable computer based video. IEEE Trans Pattern Anal Mach Intell 1998;20:1371-5.
62. Yamato J, Ohya J, Ishii K. Recognizing human action in time-sequential images using hidden Markov model. In: Computer Vision and Pattern Recognition, Proceedings CVPR '92, 1992 IEEE Computer Society Conference on; 1992. p. 379-85.
63. Lv F, Nevatia R. Single view human action recognition using key pose matching and Viterbi path searching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2007. p. 1-8.
64. Natarajan P, Nevatia R. Coupled hidden semi Markov models for activity recognition. In: IEEE Workshop on Motion and Video Computing; 2007. p. 10-7.
65. Yu E, Aggarwal JK. Human Action Recognition with Extremities as Semantic Posture Representation. Vision Research; 2009. p. 1-8.
66. Kellokumpu V, Zhao G, Pietikainen M. Recognition of human actions using texture descriptors. Mach Vis Appl 2009;22:767-80.
67. Shi Q, Cheng L, Wang L, Smola A. Human action segmentation and recognition using discriminative semi-Markov models. Int J Comput Vis 2010;93:22-32.
68. Oliver N, Horvitz E, Garg A. Layered representations for human activity recognition. In: IEEE International Conference on Multimodal Interfaces; 2002. p. 3-8.
69. Yu E, Aggarwal JK. Detection of fence climbing from monocular video. In: The 18th International Conference on Pattern Recognition (ICPR); 2014.
70. Zhang D, Gatica-Perez D, Bengio S, Mccowan I, Lathoud G. Modeling individual and group actions in meetings: A two-layer hmm framework. In: IEEE Workshop on Event Mining in Video (CVPR EVENT); 2004.
71. Dai P, Di H, Dong L, Tao L, Xu G. Group interaction analysis in dynamic context. IEEE Trans Syst Man Cybern B 2008;39:34-42.
72. Gong S, Xiang T. Recognition of group activities using dynamic probabilistic networks. In: IEEE International Conference on Computer Vision (ICCV); 2003. p. 742.
73. Shi Y, Huang Y, Minnen D, Bobick A, Essa I. Propagation networks for recognition of partially ordered sequential action. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2004. p. 862-9.
74. Yin J, Meng Y. Human activity recognition in video using a hierarchical probabilistic latent model. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition - Workshops; 2010. p. 15-20.
75. Han L, Wu X, Liang W, Hou G, Jia Y. Discriminative human action recognition in the learned hierarchical manifold space. Image Vis Comput 2010;28:836-49.

76. Mauthner T, Roth PM, Bischof H. Temporal feature weighting for prototype-based action recognition. In: The 10th Asian Conference on Computer Vision; 2011. p. 566-79.
77. Zeng Z, Ji Q. Knowledge based activity recognition with dynamic Bayesian network. In: The 11th European conference on Computer Vision (ECCV); 2010. p. 532-46.
78. Ivanov Y, Bobick A. Recognition of visual activities and interactions by stochastic parsing. IEEE Trans Pattern Anal Mach Intell 2000;22:852-72.
79. Joo SW, Chellappa R. Attribute grammar-based event recognition and anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop; 2006. p. 107-14.
80. Minnen D, Essa I, Starner T. Expectation grammars: Leveraging high-level expectations for activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2003. p. 626-32.
81. Moore D, Essa I. Recognizing multitasked activities from video using stochastic context-free grammar. In: AAAI National Conference on Artificial Intelligence; 2002. p. 770-6.
82. Kitani K, Sato Y, Sugimoto A. Recovering the basic structure of human activities from a video-based symbol string. In: IEEE Workshop on Motion and Video Computing; 2007. p. 9.
83. Wang L, Wang Y, Gao W. Mining layered grammar rules for action recognition. Int J Comput Vis 2010;93:162-82.
84. Nevatia R, Hobbs J, Bolles B. An ontology for video event representation. In: IEEE Conference Computer Vision and Pattern Recognition Workshop; 2004. p. 119.
85. Ryoo MS, Aggarwal JK. Recognition of composite human activities through context-free grammar based representation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2006. p. 1709-18.
86. Pinhanez C, Bobick A. Human action detection using PNF propagation of temporal constraints. In: In Proceedings of the Conference on Computer Vision and Pattern Recognition; 1997. p. 898-904.
87. Intille SS, Bobick AF. A Framework for Recognizing Multi Agent Action from Visual Evidence. In: Proceedings AAAI-99. AAAI Press; 1999. p. 518-25.
88. Ghanem N, Dementhon D, Doermann D, Davis L. Representation and recognition of events in surveillance video using petri nets. In: Proceedings of Conference on Computer Vision and Pattern Recognition Workshops CVPRW; 2004.
89. Siskind JM. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. J Artif Intell Res 2001;15:31-90.
90. Aggarwal JK, Ryoo MS. Semantic representation and recognition of continued and recursive human activities. Int J Comput Vis 2009;82:1-24.
91. Gupta A, Srinivasan P, Shi J, Davis L. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009. p. 2012-9.
92. Tran SD, Davis LS. Event modeling and recognition using Markov logic networks. In: The 10th European Conference on Computer Vision: Part II; 2008. p. 610-23.
93. Ijsselmuiden J, Stiefelhagen R. Towards high-level human activity recognition through computer vision and temporal logic. In: The 33rd Annual German Conference on Advances in Artificial Intelligence; 2010. p. 426-35.
94. Morariu VI, Davis LS. Multi-Agent Event Recognition in Structured Scenarios. In: CVPR; 2011.