DOCUMENT

SCORE

93 of 100

ISSUES FOUND IN THIS TEXT

54

PLAGIARISM

Checking disabled

Contextual Spelling 5

Confused Words	3
Mixed Dialects of English	1
Misspelled Words	1

Grammar 25

Determiner Use (a/an/the/this, etc.)	11
Faulty Subject-Verb Agreement	5
Incorrect Verb Forms	4
Incorrect Noun Number	2
Wrong or Missing Prepositions	1
Pronoun Use	1
Incorrect Phrasing	1

Punctuation 11

Comma Misuse within Clauses	6
Punctuation in Compound/Complex Sentences	5

Sentence Structure 1

Incomplete Sentences	1

Style 12

Passive Voice Misuse	4

Wordy Sentences	4
Improper Formatting	3
Politically Incorrect or Offensive Language	1

Vocabulary enhancement

No errors

Abstract

More and more datasets made for Human Activity Recognition (HAR) have been made available for publics in recent years. And Human Activity Recognition has gain attention due to its wide range of application from surveillance, medical personal [1] assisted tool, robotic to the interaction between human and machine. And with deep learning technics applied recently especially for image classification researchers have switch [2] and focus more and more from [3] traditional processing to deep learning technics. Although, [4] extracting the correct [5] features for further processing still a challenge, traditional technics [6] still been used for in HAR to avoid computational complexity that come with deep learning methodologies. Understanding human behaviors is a challenging problem in computer vision, we have witnesses [7] recently significant advances with proposed novel methodologies [8] for tracking, pose estimation, [9] and movement recognition. This survey is a succinct description of different existent technics and methods apply in HAR, following previous survey [10] and papers.

Keywords: Human action recognition, Activity recognition, feature extraction

1. Introduction.

Since the early fourteen hundred with Da Vinci work and studies which was [11] interested in Human Appearances to help his student drawing perfectly [12] Human action such as people climbing, going upstairs or going downstairs[https://www.slideshare.net/zukun/cvml2011-human-action-recognition-ivan-laptev-9017571]. With his work, one of well documented [13] research in early [14] Human Action Recognition Da Vinci insist that a painter

1 Possibly confused word: *personal*

2 [~~switch~~ → switch**ed**]

3 [more **~~from~~** → more **on**]

4 [Although,/]

5 Unusual word pair

6 [technics **has**]

7 Possibly confused word: *witnesses*

8 Repetitive word: *methodologies*

9 [estimation,/]

10 Repetitive word: *survey*

11 [~~was~~ → were]

12 Overused word: *perfectly*

should be fully [15] aware of the body structure (nerves system, muscles and bones structures, etc.) to understand various motions.

Intelligent environment (intelligent [16] home, intelligent electronic devices) exploit data collected from users and anticipate the probability of the end result [17] whether bad or worst case scenario. The system is able to [18] get the information, interpreted it and then take an action [19] or suggest an action. As we are in the era of intelligent automate system [21]. [20] And common [22] tasks: walking, standing, running, sleeping, etc. are being study [23] and interpreted by computer [24] system.

Identify humans from video sources has attracted increasing attention in several application domains, such as for content-based video annotation and retrieval, video surveillance, and other applications[1]–[3], but giving semantic meaning to human action or behavior is so challenging, in fact it not necessarily easy to understand what an action [26] really [27] mean. [25] This complexity is source [28] of challenges from an academic point of view. In fact, there is no better way to categorized research due to its complexity, but mainly following [4] we can categorize [29] in three type : [30] Surveillance, Control [31] and Analysis.

People counting or crowd flux, flow, and congestion analysis in public area [33] such as train, bus station or mall[5] can be grouped [34] in Surveillance applications [35], Human Computer Interfaces[6] or virtual reality can be grouped [36] in Control applications and Diagnosis of patient can be grouped [38] [37] as such in Analysis applications [39] of Human Action Recognition or Computer vision field. [32]

The potential [40] amount of applications [41], the speed [42] and price of current hardware especially in poor countries [43] and the focus on security issues have intensified the work within the computer vision community towards retrieving, collecting and analyzing human behavior. Furthermore, the

13 [~~well-documented~~ → well-documented]
14 Unusual word pair
15 Overused word: *fully*
16 Unusual word pair
17 [~~end result~~ → result]
18 [~~is able to~~ → **can**]
19 [~~take an action~~ → take action]
20 Sentence fragment
21 [the system]
22 Overused word: *common*
23 [~~are being study~~ → are being stud**ied**]
24 [a computer or the computer]
25 Wordiness
26 Repetitive word: *action*
27 Overused word: *really*
28 [a source or the source]
29 Repetitive word: *categorize*
30 [~~type :~~ → type:]
31 [Control,]
32 Wordiness
33 [~~area~~ → areas]
34 Passive voice
35

rise of terrorism and securities issues has tremendously increase [44] the research field especially in security[7], means in surveillance.

Major applications of HAR are found in security, medical, entertainment, interaction. Thanks to previous studies counter terrorism team can detect and predict from a certain number of patterns and technics a suspicious behavior. In medical, personal devices can help provide live and accurate health status of a patient (in particular old people) as such provide a good direct and quick response from the doctor. In entertainment, the HAR methods applied can help identify and even predict a player next move and in Interaction the application of HAR methods provide good robotics system that come close to the perfection of expressing, understanding and reflecting human behavior. So according to the complexity of the facing situation categories may be determine like: action behavior, gestures behavior and interactions behavior[8] as in Figure 1 bellow.

An action it's a form of expression with is compose of different gestures: running, climbing are examples of common actions and has variable timing. A Gesture it's a non-vocal form of communication where the actor express and exchange information via one part or a combination of some part of the body mostly hands, foot, and head. Often, the gesture does not exist in a long period time. And an Interaction it's an action during which actors (humans or inhuman) exchange information or interact such in hugging, scanning QR code using one device over another device.

Due to challenges and issues surrounding Human Activity Recognition: intra-class variations, viewpoint variations, environmental complexities, occlusions, and more. Current system, still not working with accuracy result. The studies in HAR remote to early decades, researchers are still trying to come close to human nature

36 [~~a~~pplications → Applications]
Passive voice

37
Passive voice

38
Repetitive word: *grouped*

39
[~~a~~pplications → Applications]

40
Unusual word pair

41
Repetitive word: *applications*

42
[speed**,**]

43
Possible politically incorrect language

44
[~~increase~~ → increase**d**]

of getting few item series and categorize it which will be called filter or training set later and from these filter being able to classify any other element that they may be facing. So, in computer vision researcher are trying to match that human particularity. But, we must acknowledge that great significant advances have been made so far even though it still can't match human vision system.

There are methods with manual design features and data driving based approaches which are distinctive by the way classification is applied such as: Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), Scale-Invariant Feature Transform (SIFT), Hessian3D, and Enhanced Speeded-Up Robust Features (ESURF) applied in manual design features and data driving based approaches mostly using deep learning where the feature are detect, interpret and process automatically by the system compare to old approaches where the feature are chosen by the human.

In general, traditional approaches apply bottom-up methodology in 3 steps foreground, feature extraction and finally classification Figure 2. As previously noted, multiple surveys, reviews have been published with different taxonomy and approach to deal with the Human Action Recognition. [8] classify HAR into two categories single layered approach and hierarchical approach where single layered focus on gesture and action or in other word low level human activities in contrast to hierarchical that focus on more complex activities or high level human activities sometimes called sub-events. With subcategories of space time approach and sequential approach for single layered method and statistical, syntactic and description based for hierarchical layered approach.

[9] presented available resources, datasets and libraries and challenges of HAR to deal with problems of background subtraction: change detection and salient motion detection. Other researchers study video base

representation with the particularity of [10] categorizing global and local features extraction where background construction-based methods and foreground extraction-based methods was used in the research. [11] and [12] respectively providing a review covering stages process of HAR from low-level processing stages to high-level feature processing applications with a focus on healthcare and last providing various object segmentation, image processing and activity recognition by briefing on sensor-based vision-based, Hidden Markov Model (HMM) also Principal Component Analysis (PCA).

Occlusion, variation in execution rate, anthropometry, camera motion, and background clutter are some of challenges as mentioned early, faced in HAR as noted in [13]. Mid-Level feature representation by applying sparse classifier for discriminative parts selection was study in [14] similarly [15] study confident based in HAR by proposing a method of making choice between the Dense Trajectories (DT) feature level and the high-level pose features. A literature review on semantic based HAR system using semantic features is presented in [16].

Acquiring data is one of the most require step in computer vision and can be obtain from multiple source. As such, the overall functionality of the system is impacted by the use of appropriate tool. And convincing improvement have been made toward these end[17][18]. Depending on the dimensionality and the depth the data obtain from these devices are classify into 2D and 3D tool. When acquiring data into 2D form, there is a loss of information from one dimension because in reality data are in 3D dimension frequently. Which imply too that system applying 3D approach are more accurate than 2D system. Existent reviews and surveys exist on HAR but due to the popularity that the field is gaining those documents are getting outdated, intrinsically writing a review in a field which improvement are too ubiquitous is challenging. In

this paper we contribute with discussion and comparison of methods applying in HAR the rest of the survey is organized as follows following the introduction: Section 2 discuss manual design features approach, Section 3 discuss Data Driving Based approach (deep and non-deep learning), Section 4 some discussion Section 5 introduce some existent dataset ending in Section 6 with the conclusion.

2. Manual design features

Manual design features approach applied in HAR has accomplish impressive result over the years of it application. The approach use feature detector (global or local feature) in case of low-level feature or high-level feature passing middle-level feature to extract important features (portion property of the overall image or sequence of images). Then, it classifies by training classifier like the Support Vector Machine (SVM)[19][20][21][22]; the approach includes space-time based, space time volumes, space time trajectories, space time features, appearance-based, shape based, motion based, hybrid, local binary patterns, and fuzzy logic-based techniques as shown in Figure 2 with accentuation on low-level features, mid-level features, and high-level features[23] spatio-temporal features as inspired in data model of [24][25] and many more which have attained good result for action recognition.

The reputation of human action recognition or human behavior recognition has led to numerous published articles and papers[6], [26]–[31]. These articles focus on different features and classifiers used in human behavior recognition. In practice considerable hardware resources and vision algorithms are required to compute the data (acquiring, saving, processing 2D, 3D fix and moving data inputs). And 3D data can be obtained through mostly tree components categories: marker-based motion capture systems MoCap[http://mocap.cs.cmu.edu/] it’s the

perfect illustration, then we have stereo cameras and finally range or depth sensors such as Microsoft Kinect. Despite the fact that vision-based action recognition continues to growth, various challenges still not resolve completely: various actions, mood of the actor, occlusion, camera position, background etc. Wher eas some researchers have utilized wearable inertial sensors including accelerometers and gyroscopes (mostly smartphone)[32]–[37] to solve these issues. Even if, there are many papers related to Human Behavior Recognition using whether depth sensor or inertial sensors, the purpose of this survey it's to inform on the current state of application in computer vision field.

Acquiring 3D data require tools, the basic on which is almost affordable to all is the Kinect (Microsoft or xBox)but the cheap and easy tool is the smartphone with the latest s tatistic reporting 2.32 billion user's Figure 3 worldwide[https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/ (access April 2017)]. The Kinect sensor include: a camera, an Infrared depth sensor, a microphone and an LED light as shown in Figure 4 and Figure 5. It can capture 8 and 16-bits with a resolution of 320×240 and 640x480 pixels properties resolution per channel. Heterogeneous method has been applied to compute the obtained data from these tools [2], [3], [38]–[41].

And for wearable inertial sensors which is often directly connected or placed on human (smartphone and other sensors equipment) and in other case (rarely) indirectly connected or placed on the human; they generate accelerometer and rotation signals corresponding to an action performed by the actor (human mostly), Figure 4 shows a capture image of a 3D skeleton source of data.

And acquiring 2D information, require an easy an accessible tool to all such as mobile phone incorporating a camera. This show how accessing to 2D data is more

simple compare to accessing 3D data.

2.1. Appearance based approach

Shape, motion and hybrid based approach are discussing in this part, where methodologies and technics are applied on 2D as well as on 3D data. Shape based are pursuit objective in [42] with authors proposing the use of bag of words (BOW) framework to classify each frames of a video and in [43], tensor shape descriptor and tensor dynamic time warping was use by [44]. More articles also applied appearance based approach in their founding: gesture recognition [45], blob analysis [46].

45 [the mood or a mood]

2.1.1. Shape based

In this approach features are obtained from shape feature silhouette. [47] obtained 3D data which is convert to 2D data using spatial distribution of gradients the data is then compute with R-transform the technic is applied on Weizmann, KTH, and Ballet dataset. In [17] the authors analyzed maps feature to separate silhouette from noisy background later the framework perform a tracking to check the silhouette movement in the scene. The method creates sequence of scene from the human silhouette maps representation and used a hybrid classifier. In practice HAR method should be computationally lean. Similarly method was proposed using K-neighbor in [48].

46 [),]

In [49], proposed a pose-based view invariant HAR method based on the contour points with sequence of the multi-view key poses. In [50]. the authors employ the contour points of the human silhouette and radial scheme with the SVM as classifier. [51], [52] build a region-based descriptor from extracting features from surrounding regions of the silhouette in the image. [52] used pose information by firstly, extracting the scale invariant features, and then clustered it to build the key poses, finishing by classifying using a weighted voting scheme.

2.1.2. Motion based

For the approach features are obtained from motion

features applied with generic classifier. A motion descriptor was proposed in [53] for unconstrained videos representation. The motion descriptor is based on motion explicit motion modeling operating on codewords generated by dense local patch trajectories, and, so doesn't need foreground-background separation. Another motion-based method was introduce by [54] using histogram of oriented gradients. In [55], action recognition method was proposed based on Human Object Interaction descriptor and pose estimation. Other authors applied kinematic spline curves [56], multiple key motion history images [57], motion trajectories [58] and joint motion similarity [59].

2.1.3. Hybrid

Approaches combining shape-based approach and motion-based approach features. An map level and silhouette-based shape features were used for separating the noise from the actual silhouette in [54] followed by an histograms of oriented gradients to better classify. Other methods based on hybrid approach were proposed in [60] [61]. The BOW and a block-wise weighted kernel function matrix were used for multi-view in [62]. While, [63] applied shape-motion prototype trees. Representing action as a sequence of prototypes and distance measure was used for sequence matching. Method tested on 5 datasets. [64] proposed key poses method as variant of motion energy and motion history Images with simple nearest-neighbor classifier.

2.2. Space time based approach

Approaches that focus on recognizing activities based on space-time features or also on trajectory matching. And an activity is represented by a set of space-time features. It has four major components: the space time interest point with two sub-categories dense detectors and sparse detectors; feature descriptor with local and global features as type; vocabulary comprise of BOW and model based

and finally the classifier with supervised and unsupervised categories. Figure 6 show an example of a human actions with dense trajectories applied in [65]

And Figure 7 show the different major component available and applied in space time approach. Moreover, [66] employed motion features as input to hidden conditional random fields (HCRF) to tackle much broader range of complex hidden structures whereas [67] proposed a Real time classification and prediction of actions.

An action descriptor of HIP, relying on the work of [68]was propose by [69] and [70] proposed to incorporate information from human – objects interactions applied over several datasets.

2.2.1. Space time volumes

In [71], an HAR system was proposed using temporal-spatial semantic, instead of using STV the authors used templates composed of 2D observations. The approach was then extended by [72] where motion history image, foreground image approach and HOG were combined, to finally used SMILE-SVM for classification. Applying space time based approach on different datasets have shown outstanding accuracy result output such as in [73] with an accuracy performance of 98.2% applied over the KTH Dataset. And [74] with a performance of 89.4% over the UCF (University of Central Florida) dataset using discriminative clustering, tree mining, tree clustering and ranking to select discriminative tree patterns.

2.2.2. Space time trajectories

Human action can be seen as set of spatio-temporal trajectories, trajectories in Space time trajectories have different levels of abstraction from low-level trajectories to high-level trajectories like handwritten characters.

However, all space time trajectories approach has a common property: time-structured patterns. Space time trajectories is applied on joint position (body joint) to differentiate actions. From these notion many papers have

been published and approaches have been proposed [75], [76].

Inspired [47] by image classification dense sampling method [65] introduced the concept of dense trajectories apply on video action recognition. After sampling and tracked using displacement information, dense points from image frame of dense optical flow field. The approach shows [48] robustness of the proposal to irregular motion changes. [77] Improve [65] work by using SURF descriptor and dense optical flow to optimize the estimation. However, when apply the approach with high density trajectories features in the video the computational cost increase. In fact, there are have been attempts to reduce the cost, to tackle the challenge saliency map method was used to capture salient region within a frame as in [78], [79],[80]. As such, applying the saliency map allow to drop some dense trajectories feature during the process without compromising the frame input.

In 2016 two major publications was made available [81], [82] representing skeleton shapes as trajectories on Kendall's shape manifold. The method uses transported-square root vector fields (TSRVFs) of trajectories and standard Euclidean norm to reduce the computational cost and increase a computational efficiency. And [83] used HOG, HOF, and MBH method for trajectories, recording an highest accuracy. [53] propose the use of explicit motion modelling method to resolve the challenge of HAR in unconstrained videos input data.

2.2.3. Space times features

In general, space time feature are local properties that contain discriminative action characteristics. And can be divided into 2 separate categories: sparse property and dense property. Features detectors based on interest point detectors such as BOW [84], and 3D HOF [85] are grouped in sparse category, while those based on optical flow are grouped into dense category. It (interest point

detectors) provide found 49ation for most recent methods (algorithms) proposed.

[86] build a feature descriptor framework and apply PCA – SVM for classification and [87] used a comparison of Harris3D and Multimodal Decomposable Models for classification. BOW still the most popular method for representation with all the different variations such as BOVW following feature extraction step, codebook generation step, encoding step and pooling step [88],[89], [90], [91]. The performance of BOVW variant of BOW approach is due to effective dense trajectory low level feature. To further improve space time feature method and provide better performance some researchers applied Fisher vector, space time occurrence.

Space time approach with feature detector with a particularity of global feature has a disadvantage of being sensitive to noise and to occlusions. So, detecting the presence of multiple person in a scene make space-time approaches can hard to recognize actions. But, space-time features focus mainly on spatiotemporal information. Other limitations are STVs approaches lack the capacity of recognizing multiple entity (person) in a multiple person image frame. Trajectory-based approaches lack the precision in localize joint position. Space time approach, even though suitable for simple dataset require multiple feature combination to handle complex dataset which also increasing the computational complexity. However, to overcome the limitations we may apply the background subtraction technic, sliding window and more methods.

2.3. Other approaches

Paradoxical to previous paragraph, there are other methods, technics, approach which can be grouped and categorized as traditional approach, but can't fit in formerly appearance or space time approach. For that, we have grouped it in others such as Local Binary Pattern and fuzzy logic-based approach.

47 [published,]

2.3.1. Local binary pattern

This is a type of visual descriptor or Texture Spectrum model used for classification in computer vision, introduce in the field in 1990 by [92][https://en.wikipedia.org/wiki/Local_binary_patterns]. Since its introduction, LBP combined with HOG has shown considerable improvement in detection performance and a full LBP survey of the different versions [50] was proposed by [93] in 2016.

[48] [an image or the image]

Several versions such as have been proposed for different classification [94],[95]. A HAR face recognition was propose by [96] based on Nearest Neighbor Interpolation classifier. This method was applied on the Olivetti Research Laboratory dataset resulting in an accuracy of 97.5% recognition rate performance. Another human action recognition approach using LBP with Gaussian mixture was used in [97], the authors method on top of intensity difference property of LBP introduce the extraction of multiple feature with error correcting output code apply over the simple vector machine classifier.

The linear base pattern approach was also been applied for multi-view HAR, like in [98], where a multi-view based on contour-based pose features and uniform rotation-invariant with simple vector machine classifier. Motion Binary Pattern was introduced for multi-view HAR by [99] in combination of Volume Local Binary Pattern and optical flow. And was tested over the INRIA Xmas Motion Acquisition Sequences dataset with a record performance accuracy of 80.5%.

2.3.2. Fuzzy Logic

Traditional approaches employ spatial or temporal features with generic classifier for representation and classif [51]ication. However, it is challenging to handle uncertainty and complexity involved in real world applications. And, so to resolve this issue the Fuzzy logic approach was introduced, to benefit from it particularity of

considering as truth only integer variables of value in a range of 0 to 1. But the notion and term was firstly introduced in nineteen sixty-five in a fuzzy set theory by Lotfi zadhe[https://en.wikipedia.org/wiki/Lotfi_A._Zadeh].

[a dense or the dense]

To resolve these uncertainty, fuzzy logic based approach has been applied as in [100] based on Interval Type-2 Fuzzy Logic Systems with feature information optimize with Big Bang-Big Crunch algorithm, the experiments were performed on Weizmann human action dataset which outperformed the equivalent Type1 Fuzzy Logic System and non-fuzzy methods regarding recognition accuracy and analysis performance. . In [101] authors utilized silhouette slices features and movement speed features, and employed fuzzy c-means clustering technique to acquire membership function. And in [102] fuzzy logic based classifier method was used to recognize human intention, [103] applied fuzzy view estimation framework to predict squat evolution of scenarios.

Most HAR appro aches depend on the view and recognize an activity through fixed viewpoint. However, in real time world applications the recognition must come from any viewpoint, which introduce the use of multi camera to collect the data, but this solution is difficult in practice because of camera calibration. Following this path [104] propose a method for view invariant using single camera and clustering algorithm, the method was applied over the IXMAS dataset. In addition other approach focus on neuro-fuzzy systems have also been proposed for gesture recognition in particular [105] and other behavior recognition [106] are also very successful in behavior recognition.

3. Data Driving Based Approach

We mention it in previous lines the performance of the HAR depends on the methods and the appropriate chosen feature as well as efficient representation of data.

Dissimilar to traditional approaches where the action is represented by picked (chosen) feature detectors and descriptors; learning-based approach in the other hand have capability to automatically learn the feature from raw data, along this line introducing end-to-end learning concept, meaning conversion from pixel level to action classification level. These approaches are grouped in non-deep learning approach and deep learning approach as shown in Figure 8 bellow.

3.1. Non-Deep Learning-Based

As one of the category dictionary learning approach is a type of representation generally focusing on sparse representation. It has been used in many applications like in image classification or in action recognition [107]. The concept is similar to BOVW methodology because it based on vectors representation. And these vectors also called code words, also called dictionary atoms sometimes. [108], four dataset were subject of the study with the authors applying spatio-temporal motion features. Genetic programming is an evolutionary technique inspired by the process of natural evolution. And may be used to solve problems without having prior knowledge and help maximizing the recognition task performance. Along the way feature descriptor evolved on filling 3D operators such as 3D-Gabor filter and wavelet.

50 [performance,]

[109] propose based on discriminative Bayesian on five dataset to recognize action and face. [110] address the problem of Cross-view action recognition by using transferable dictionary pair. The authors differentiate specific dictionaries where each dictionary equal to one camera view. Moreover, [111] extended [110] work with common dictionary technic which acquire information from different views. A weakly supervised dictionary learning-based approach with trace lasso was proposed in [112]. The approach used dictionary and fully exploiting visual attribute correlations rather than priors label

information. In [111] the authors applied dictionary leaning-based methods for cross-view action recognition. This method used two dictionary learning approaches to learn the sparse representations of videos regardless of the views, by enforcing correspondence videos in a set. It was performed over tree dataset and shows great performance.

3.2. Deep Learning Based

This is part of machine learning algorithms that use cascade nonlinear processing unit layers to extract feature and transform the input into multiple small feature level. And Each layer uses output from previous layer as input. And the algorith 53ms may be supervised for analysis pattern or unsupervised for classification[https://en.wikipedia.org/wiki/Deep_learning#Definitions].

51 [a generic or the generic]

Previous studies applied on different dataset shows that traditional approach does not fulfill totally the process of computer vision and action recognition. As such, HAR system that can offer the possibility of automatically determine feature descriptor, learn and evolve without the intervention of human will be crucial for evolvement of action recognition. This is where deep learning come in handy and it as shown over the past studies how important it is in machine learning with the aimed of learning different multiple levels of representation and abstraction, to make information meaningful and deep learning as also shown it accuracy and performance higher than traditional approach and it is applied in speech, images, videos and text extraction, representation, and classification. As in Figure 8 deep learning can be grouped into two entities: unsupervised approach such as Deep Belief Networks, Deep Boltzmann machines, Restricted Boltzmann Machines, and regularized auto-encoders and supervised approach: Deep Neural Networks, Recurrent Neural Networks, and Convolutional Neural Networks.

But due to the success of models such as the simple

vector machine, non-available data to perform algorithm on for training deep learning approach have received little attention in the beginning of computer vision field and action recognition in particular.

3.2.1. Unsupervised deep learning model

During training process in this model there is no need for class [54] to label, meaning this model is used and apply when facing the unavailability of labelled data. In 2006, [113] work trigger the notion of deep learning by proposing deep belief networks method with the uses of unsupervised algorithm to train DNN a layer at the time. The same year saw [114] following the same path proposing a feature reduction technic for deep learning. Considering the introduction of deep learning approach, there have been an increase concern to apply this approach for divergent application whether it is in image, classification, human action recognition, speech recognition, health care system, intelligent home, object recognition or more.

[52] [the squat]

[115] proposed for video action recognition an unsupervised learning approach, where the authors used a spatial appearance feature and incorporate with CNN technic. The solution proposed was applied on the ImageNet Dataset. [116] proposed DBN with Restricted Boltzmann Machines. Despite the fact that unsupervised approach offer performance higher than traditional approach seen before, there still a challenge faced by researchers, because processing from unlabeled video data still a challenge.

To bring some light to it, [117] used unsupervised approach, whereas data were collected from four different dataset applied with hybrid feature models and active learning. Another study using Deep Belief Networks was proposed by [118] where the authors used skeleton coordinates feature obtain from depth images. Even, though we have seen performance in it application,

unsupervised approach researchers are losing and abandoning the method over the supervised approach, especially with the implication of Convolutional Neural Networks. But [119] study advocate that in the future unsupervised approach will be the most applied approach rather than supervised approach because, as like human recognition and identification of object come by observation and not by the notion of being told, so does future system will be able to recognize unsupervised elements.

3.2.2. Supervised deep learning model

There is a significant increase of studies related to deep learning in recent years whether it applied for classification, modeling texture, regression, information retrieval, robotics, fault diagnosis and many more with deep CNN or RNN. Many reason can be listed for that matter but here we only nominated the access to data, the access to materials and the computational abilities.

Until now, CNN is considered as one of the most effective and powerful solution for action recognition, it has shown great performance in different applications and for different tasks like HAR, image classification or even hand writing recognition [120], [121], [122], [123]. The Convolutional Neural Network consist of dividing the input into multiple layers such: convolutional layers, Rectifier Linear Units, pooling layers and fully connected layer, but in theory only three categories are cited: convolution layers, subsampling layers, and full connection layers as in Figure 9.

[124] elongated [125] work on by applying the technic on video using fixed scene frame as data matrix input, unfortunately the outcome performance was not useful. Later [126] using two-stream convolutional neural network to resolve the issues faced by [124] by combining late fusion and the method produce great result. However, due to computational complexity two stream technic is not

recommended or suitable for real time system application.

In general, deep learning deal with retrieve information express in two dimension, but some application retrieves three-dimension data as such require 3D convolution neural network. [127], [128] works applied 3D CNN, the first performance reaching a high sensitivity of 93.16% with average of 2.74 false positives for detection and recognition of micro bleeds in magnetic resonance images and the second one inspired by VoxNet and 3D ShapeNets applied 3D CNN on the ModelNet dataset to acquire and recognize also classify the data.

There still exist issues such as computational complexity or the amount of require data to create the 100% perfect system for HAR. To follow the path [129] propose a variation of CNN called Factorized Spatio temporal Convolutional network. The approach factorizes standard three-dimension convolutional neural network model as two dimension spatial kernels to reduce the amount of learn parameters and the complexity of the network, and another study was made by [129] still using 3D CNN. And the application of this approach shows that the approach is better for spatio temporal information compare to 2D data and apply with linear classifier exceed the state-of-the-art methods. Other researchers have mixed traditional approach and CNN, arguing that it improves performance [130]. Another variation of convolutional neural network was introduce by adjusting pre-trained convolutional neural network, extracting at frame level feature, applying PCA, SVM in [131].

53 [the previous]

Another method of supervised deep learning is semantic based feature like pose is also use in computer vision to describe an action [132], [133]. Descriptors of this method are based on the motion and appearance information, from joint human body parts. Experimental of the approach were evaluated on Berkeley MHAD dataset, on JHMDB and MPII Cooking datasets. The outcome result shown

better performance that some others technics. [134] utilize contextual information and adapted the region based CNN for classification, and [135] address the task of semantic image segmentation. [136] propose a method to deal with multi view data source by learning from 2D dense trajectories and renders synthetic 3D model and once more its shown how deep learning approach is far better in performance compare to traditional approach. Also, learning based approach use the advantages of learning feature from raw data or unlabeled data.

However, learning based approach have some limitations such as the amount of data needed for training. To solve the problem [137] proposed in 2016 a dataset compose 55 of 200 action or 849 hours of video to help apply learning base approach algorithm. In fact, recent statistic shows increase interest in computer vision, in human action recognition and in convolutional neural network as process, which means we will not be surprise if some researchers found in near future breakthrough algorithms for action recognition.

54 [model,]

4. Discussion

A low level feature it's a portion of an image, that allow to simplify the complexity of an image by getting properties related only to a certain pattern. As such, the input may be an with M value on the X axis, N value on the Y axis and 3 the color property RGB. Which lead us to value of a low level feature entity. Extracting such a valuable information it's one of the first task faced by system in computer vision in particular.

As human, from the day we are born we deal easily with images and the natural instinct always take the lead in categorizing the environment surrounding us[138], so does one may wonder if a computer can also adapt and recognize entity from images. In past years the answer to the question will be no but with recent research discovery it has been made possible for computer to obtain, read and

understand an image as such classify it. To reach the goal, computer does not read the input as an all one single element, that the importance of devise as it has been proven “Divide to better reign”, so the system will reduce and divide the input into multiple smallest entities possible and treat each new as a single element. Commonly there are tree property used during the extraction process: color, shape and texture[25], [27], [139]–[143]. And the performance of the system is a related to a good choice of feature and extraction method.

In regard to the previous noted properties and based on their importance of extracting feature in computer vision or predictive modeling and probabilistic data mining; There still challenges that need solution to be found for, to completely capture and classify Human Activity or Human Behavior, given the complexity of human action or reaction to the reality, environment, etc. Advance have been made in computer vision by applying different technics and method over multiple properties to overcame the challenge. Some researcher have applied face feature and speech feature [144]–[146], or text feature and speech feature [147], [148], or the combination of multiple feature (posture, speech, face, etc.) [149]–[151]. We have acknowledged the fact that using multiple feature increase the performance and the accuracy but at the same time increase the complexity.

We have to notice that all approaches represent activities (action, behavior) as a frame sequence in time and space locations whether it has been extracted from moving entities or fix images and use different classification models. [70] proposed to transfer from one dataset to another dataset after incorporating information going from human interaction to object interaction. Hidden conditional random fields from motion features input with a combination of large-scale global features and local patch features to distinguish various actions in [66]. [152] and

[153] used Random forests for action representation respectively classify and localize human actions in video using a Hough transform voting framework, and, a vocabulary of local appearance-motion features and fast approximate search in a large number of trees. A real-time algorithm to describe interactions was proposed the early two thousand, with a capacity to detect and track movements, creating a feature vector given as input to Hidden Markov Model for classification that describes the motion[154]. Complex activity recognition with two sequential sub-tasks increasing granularity levels, applying firstly human-to-object interaction techniques, then context-based information to train a conditional random field model was proposed by [155].
Self-organizing maps to learn body posture with fuzzy distances, for time invariant action representation, the algorithm is based on multilayer perceptions was used by [156]. Local occupancy pattern and actionlet ensemble model was proposed by [157] in which the authors first captured the human body parts then captured intraclass variations to allow error handling in depth camera. Interaction between activity and scene to recognize human activities using 3D skeletal representation and geometric representation of the scenes. and, appearance-to-pose mapping for activity problem. Gaussian processes as an online probabilistic feature using sparse representation to reduce complexity in computation was applied by [75] and [158] used sparse representation of skeletal data with dissimilarity space to recognize behavior or activities.

To describe an event, an action with multiple features containing meaningful information can be considered to achieve the goal. As in previous paragraph more and more papers have been published in computer vision field. And for these articles they are mostly based on feature fusion which can whether be early fusion or late fusion. Using one key element is good in the functionality of anything

but using multiple key feature increase dramatically the chance of better performance and great accuracy in the outcome. And, so using multiple feature increases the performance recognition.

[150]proposed a novel method by applying Kernel Canonical Correlation Analysis and Multi-view Hidden Conditional Random Fields for Human Activity Recognition to detect and interpret agreement and disagreement notion from nonverbal audio-visual cues data. However, the proposed methods from previous paper face challenging difficulty when classifying, and, sometimes the audio sample get lost in the procedure. In the other hand [145] applied multiple hierarchical classification models taken from the properties of NN (Neural network) for recognizing audio emotional feature as well as visual emotional feature instead of labels.

[159]used the Hollywood Human Actions dataset and by taking advantages of video sequences to propose a HAR system, the researcher extract firstly visual feature before extracting audio feature and finally apply support vector machines classifiers [160] used audio and visual cues and apply several classifiers to separate the information and categorize whether it an audio or visual content using spatio-temporal features to allow the extend spatio-temporal bag of features with geometry, and, apply kernel-based learning techniques. Similarly, [161] with previously using multiple kernel learning algorithm for better estimation, applied fuzzy techniques and put together support vector machines classifiers output.

But human action or activity are complex and influence by the mood, the emotions, the interactions, etc. This explain the complexity in computer vision field, and choosing the exact parameters or properties or proper features useful for HAR become a key component to advance in recognizing and predict human behavior as in [162]. Some researchers focus on audio data, such [163] where the authors using

the canonical correlation analysis (CCA) propose another way of using and interpret human behavior apply to leap feature and speech synchronization. Whereas [164] use canonical kernel and Space Vector Machine (SVM) in learning and classifying images. Other researchers took advantage of facial expressions and facial action coding system (FACS) [165], to describes all eventuality of behavior with the combination of action units (AU), and audio information to identify their emotion of actors, following the path with a real time 3D system [166]. [167] applied Conditional random network to solve at a certain point the challenging task of recognizing and classifying human behavior by selecting tree main classes friendly, aggressive or neutral employing conditional random field method, the author applied there method over the dataset obtain from speech in the Greek parliament.

55 [2016,]

Hierarchical Dirichlet Process was applied by [168] which allow the creation of multiple hidden state and used Markov-chain Monte Carlo for sampling the data which gave the opportunity to identify and classify behavior in two type agreement or disagreement from non-verbal features model and cues. [169] paper study the mimicry during human interactions, with a notice on the fact that first and for most these signals where study by psychologist before being used and classify by researcher in computer vision, so, the authors as one of the first in the first in the field to applied computation techniques on such type of feature to capture continuous detection of human behavioral mimicry. And [170] applied psychology [171] notion coupling with computational method to classify human activity by decomposing an activity and use each section of the action as feature or input. And comparing the result to the Hidden Markov Model classifier the author found a significant increasing improvement. Whether it's in a healthcare systems [172], in security [173], or in autonomous prediction [174] computer vision

will keep attracting researcher because, the complexity of human behavior keep putting a big difference between human and machines, Although, there are currently improvement in machine learning while applying technics to understand human behavior it still a challenge to fully understand accurately how human behave or could behave. As such, selecting exact useful and important key element for interpretation of human activity demure an issue. Even though [175], [176] try to characterize or classify human action it still not sufficient, up to date only combination of multiple different features can almost try to describe human behavior. Nonetheless, complex computational classification is the consequent of high level feature, as such there is not enough research applied with these properties. Also, Learning-based approaches have been categorized into dictionary learning and supervised approach, genetic approach as well as unsupervised deep learning approach. However, the categorization boundary may overlap, as such it is not strict boundary limit.

5. Example of some public datasets

Many public domain data set have been made available to all, bellow is a non-exhaustive list of some of the data source.

Common well known public dataset

5.1. Berkeley MHAD dataset[http://tele-immersion.citris-uc.org/berkeley_mhad#about]

Generated as part of the NSF funded project (#0941382), CDI-Type I: Collaborative Research: A Bio-Inspired Approach to Recognition of Human Movements and Movement Styles. The Berkeley Multimodal Human Action Database (MHAD) contains 11 actions performed by 7 males and 5 female subjects in the range 23-30 years of age except for one elderly subject performing a total of 660 actions such as jumping in place, jumping jacks, throwing, waving hands, clapping hands, sit down, stand up [177].

With [178] applying meta-cognitive radial basis function network and its projection based learning algorithm to achieve over 97% recognition accuracy.

5.2. URFD dataset

Created by Michal Kępski from Interdisciplinary Centre for Computational Modelling at the University of Rzeszow in December 2014. The dataset consists of 70 sequence of 30 falls + 40 activities of daily living record with 2 Microsoft Kinect cameras.

[179] and [180] both applied their method on the URFD dataset correspondingly with statistical control chart and neural network for classification and improving HAR system output, and, strategy for fall events detection.

5.3. UTD MHAD dataset

Collected as part of a research on HAR using fusion of depth and inertial sensor data, the dataset was created at the Department of Electrical Engineering, University of Texas at Dallas. Consisting of 300 actions (wave, throw, catch, draw, etc.) perform by six actors (3 males and 3 females) with depth sequence size of 424x512x number of frame.

[181] method applied Spatio-Temporal Interest Point to detect changes. Then, extract appearance and motion features interest points using the HOG and Histogram of Optical Flow (HOF) descriptors. To finally match the SVM by BOW of the space-time interest point descriptor.

[182] encode spatio-temporal information of skeleton sequences with convNets.

5.4. Weizmann Human Action Dataset

Dataset introduced by the Weizmann institute of Science in 2005. This dataset consists of 10 simple actions with static background: walk, run, skip, jack, jump forward or jump, jump in place or pjump, gallop-sideways or side, bend, wave1, and wave2. Consisting of 90 videos of Resolution = 180x144 of Static camera. The dataset has homogeneous outdoor backgrounds. Also provides

irregular versions (with dog, occluded, with bag, etc.) for robustness experiment. Some research has shown an accuracy of hundred percent when applied on this dataset[52].

6. Conclusion.

This survey review 56 different approaches used in Human Action Recognition (HAR) or Human Behavior Recognition along with technics and method applied. Focusing in categorizing traditional representation based and learning base representation. Despite the enormous amount of published papers, methodologies employ or technics applied 57 to collect and process the data, there still challenging problem whether in the interpretation or labeling of action. Human 58 can sometimes make action 60 59 which does not exactly means what it looks like but instead meaning 61 differently according to the mood (e.g. putting both hand 62 behind the neck) or others reasons. As such, there still window for improvement in computer vision field. That being said 63, the accuracy and performance are factors of used features 64 but that also imply that the system become 65 more complex 66 if more features, 67 are extract and more method are applied to it. Next step of this document will be to give more documents 68 and give even more details on the founding 69 so far in computer vision and help new researchers to have a document that 70 reflect 71 everything that need 72 to be known before jumping into the field and have the perfect knowledge foundation. The research will facilitate better judgement 73 in where does the notion of Human Activity recognition come from, what is it 74 current state and final how can future researchers improve and solve different challenges.

56 [~~review~~ → reviews]

57 Repetitive word: *applied*

58 [~~H~~uman → A human]

59 Repetitive word: *action*

60 [an action or the action]

61 Repetitive word: *meaning*

62 [~~hand~~ → hands]

63 Passive voice

64 [features,]

65 [~~become~~ → becomes]

66 Overused word: *complex*

67 [features,]

68 Repetitive word: *documents*

69 Possibly confused word: *founding*

70 Repetitive word: *document*

71 [~~reflect~~ → reflect**s**]

72 [~~need~~ → needs]

73 [judg~~e~~ment → judgment]

74 [~~it~~ → its]