Global Journal of Engineering and Technology Advances

eISSN: 2582-5003

Cross Ref DOI: 10.30574/gjeta

Journal homepage: https://gjeta.com/

Global Journal of Engineering and Technology Advances
eISSN:2582-5003
Global Scholar Press, INDIA

(Review Article)

Subject review: Human activity recognition based on depth images

Ahmed Kawther Hussein *

Department of Computer Science, College of Education, Mustansiriyah University, Baghdad, Iraq.

Global Journal of Engineering and Technology Advances, 2024, 19(02), 029–038

Publication history: Received on 20 March 2024; revised on 29 April 2024; accepted on 02 May 2024

Article DOI: https://doi.org/10.30574/gjeta.2024.19.2.0072

Abstract

Human activity recognition (HAR) has becomes a well-known area of study for researchers due to its multiple possible applications. HAR systems are applied in many areas including supervisory nursing care, medical supervision, surveillance, human- machine interaction, gaming and entertainment, etc. However, developing a robust HAR system that incorporates a comprehensive and scalable set of feature spaces is a challenging task. Much research has been done in this area to develop efficient HAR systems.

In this article, a review of the developments in Human Activity Recognition (HAR) will be presented. Human activity recognition works on two basic principles; feature extraction from sensor data and classification of features into action classes. The review presented in this article is hence divided into two parts. The first part deals with feature representation in HAR systems such as Space-time Features, Frequency Features, Local Descriptors, Optical Flow Based, and Skeleton Joints. The second part deals with the classification approaches used in the literature such as Bayesian, and HMM.

Keywords: HAR; local descriptors; HMM; Optical flow based; DBN; 3D spatio-temporal volume

1. Introduction

In recent years, many researchers have done intensive work in the area of human activity recognition. A hierarchical dynamic Bayesian network was proposed to cooperatively recognize the environment and the activity of a human [1]. The hierarchical nature of the model was used to learn data-driven decompositions of complicated activities into sub-activities. The research proved that the hierarchical nature of the model is capable of describing the observed data well and leads to better performance.

The trajectory-based hierarchical model was presented for spatio-temporal context representation [2]. Three levels of context were presented in the model; point-level context, intra-trajectory context, and inter-trajectory context. The spatio-temporal context information for the last two levels was encoded into the transition matrix of a Markov process and then the stationary distribution as the final context descriptor was extracted.

The pose descriptor known as Histogram-of-Oriented-Rectangles (HOR) is introduced for the representation and recognition of human actions in videos [3]. In the work presented, human actions were represented by poses without dealing with complex representations of dynamics. Each pose in an action sequence was represented by oriented rectangular patches and spatial oriented histograms were then formed to represent the distribution of the patches. Next, matching techniques like the nearest neighbour classification were used to carry the information from the spatial domain to the temporal domain. [4] presented a hierarchy-based discriminative space-time neighbourhood features for human action recognition. The proposed method aims at learning the shapes of space-time feature neighborhoods that are most discriminative for a given action category. Local motion and appearance features are first extracted from a set

[*] Corresponding author: Ahmed Kawther Hussein

of training videos, quantized into a visual vocabulary and then grouped into candidate neighborhoods consisting of the words associated with nearby points. Furthermore, a hierarchy of space-time configurations is formed using the descriptors for the variable-sized neighborhoods.

The idea of decomposing an action into multiple finer-grained elements in space and time is further explored and hierarchy-based mid-level action elements (MAEs) were introduced [5]. Each MAE was represented as an action-based spatiotemporal segment in the video. Furthermore, action-related segments were distinguished from background segments. A discriminating algorithm was then introduced to automatically cluster MAEs at multiple levels of granularity.

2. Feature Types for Human Activity Recognition

Feature extraction is required for the reduction of data dimensionality to simplify and speed up computation. Five main types of features can be extracted: space-time features, frequency features, local descriptors in terms of features, skeleton joints, and, optical flow-based features.

2.1. Space-time Feature

Many algorithms have been proposed to recognize human actions using representations built from 3D silhouettes [6]. Researchers have attempted to extract silhouette features which are used to represent spatial and temporal domains. These features are extracted by building a vector of silhouette features sequenced in time. This vector of features can capture human motion and action. One of the common spatial-time features that is extracted from silhouette and has been used for video-based human action recognition is a Space-Time Volume feature [7]. This vector of features can capture human motion and action by building a vector of silhouette features sequenced in time as shown in Figure 1. This vector of features can capture human motion and activity. [8] proposed a combination of Symbolic Aggregate approximation (SAX) and time-series representation for the silhouette.

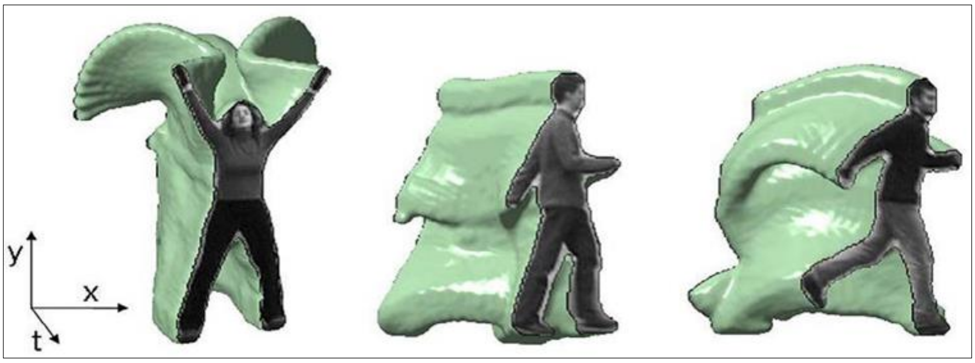

Figure 1 Space-time shapes of ―jumping-jack‖, ―walking‖ and ―running‖ actions [6]

[6] applied the properties of the solution to the Poisson equation for extracting space-time features such as action dynamics, local space-time saliency, orientation, and shape structure. They indicated that these features are valuable for action clustering, detection, and recognition., [9] used sampling bag of 3D points to get a set of postures that correspond to the nodes of the action graph as shown in Figure 2. The application of the 3D points provides the possibility to extend the proposed approach into a view-invariant one. [10] proposed a Three- Three-dimensional motion History Image (3D-MHI) which has the capability to encode the dynamics of a sequence of moving human silhouettes. The original Motion History Image (MHI) is combined with two additional channels, i.e. two depth change-induced motion history images (DMHIs): forward-DMHI and backward-DHMI which encode forward and backward motion history respectively. These combined feature representations are known as 3D-MHIs.

The depth image can be divided into sectors and the average distance from the hand silhouettes in each sector to the center of the normalized hand contour can be calculated. This created a feature vector to recognize hand gestures using the small dataset of twelve dynamic American Sign Language (ASL) gestures [11]. In general, the existing algorithms of 3D silhouette are suitable for single-person action recognition due to information loss. Occlusion and noise can harm the silhouettes and the interaction between the person and background object makes it difficult to extract accurate silhouettes (e.g. pick up the bag). the 3D silhouettes-based algorithm is usually view-dependent, even though they are not limited to only modelling parallel motions as in intensity images.

2.2. Frequency Feature

Most frequency-based features of the human activity recognition approach rely on wearable sensors such as inertial measurement units included in smartphones. This is because such sensors provide one-dimensional signals or time series which makes it convenient to use frequency features for data reduction and information extraction [12]. One well-known feature extraction method is Discrete Fourier Transform (DFT).

The DFT represents the spatial and time information of the input images in frequency domain. The Fourier transform preserves the information in the original signal and ensures that important features are not lost in the FFT [13] using discrete Fourier transforms (DFTs) of small image blocks as the features for human activity recognition. [14] present scheme combines both Key Fourier Descriptors (KFDs) and principal component analysis (PCA) techniques to reduce the size of features and increase the accuracy of classification. The drawback of the DFT is its limitation in terms of angle of view, occlusion, and sensitivity of noise and the non-periodic nature of several human activities.

2.3. Local Descriptor Feature

The local spatiotemporal feature (STF) is widely used in action recognition from video sequences. They provide compact representation of a sequence of time-stamped image to reflect or represent events in an abstract way [15]. The main advantages of the Local spatio-temporal features are invariant to spatio-temporal scales and more robust to noise, and occlusions and deal with multiple motions, the interaction between person and person or interaction between person and object. The limitation in local spatio-temporal is its view-dependency. Besides, the current local spatial methods require the whole video as input. However, the feature extraction algorithm is not fast, which causes a limitation in real-time applications. [16] obtained semi-local features called random occupancy pattern (ROP) features, which employ a novel sampling system that efficiently studies an extremely large sampling space.

In local STF, the frames are regarded as (x, y) and t is regarded as the temporal axis. Local spatiotemporal interest points (STIPs) are first discovered, and the descriptors are then built around the STIPs. [17] proposed a Comparative Coding Descriptor (CCD) to represent the depth of information for action analysis. Depth information is considered as a spatio-temporal volume of depth values. Cuboids with size of 3 × 3 × 3 can be extracted from the volume. The value on the reference point is compared with the nearby 26 points respectively. The extraction of STIPs is carried out from depth videos by using a filter called depth cuboid similarity feature (DCSF), [18] used MSR actions data set and compared the experiment results of depth cuboid similarity feature (DCSF) with Harris3D [19] and Cuboid descriptor [20].

Local spatio-temporal features capture the shape and motion characteristics in video and provide an independent representation of events. It is invariant to spatio-temporal shifts and scales, and it naturally deals with occlusions, multiple motions, person-to-person interaction, and person–to–object interaction. Because such features are usually directly extracted without the need for motion segmentation and tracking, it makes the algorithm

more robust and has a wider range of applications. The limitation falls in the following aspects. First, the feature is view-dependent, since the cuboid is extracted directly from the x, y, and t volume. Secondly, the current method requires the whole video as input, and the feature computation algorithm is not very fast, thus limiting its real-time application [21] In addition, the current spatio-temporal interest points extraction and feature description techniques regard the depth channel and RGB channels independently. This may not be the theoretically optimal way since the RGB in the 3D world; it might be interesting to develop algorithms with a better fusion of the two types of data.

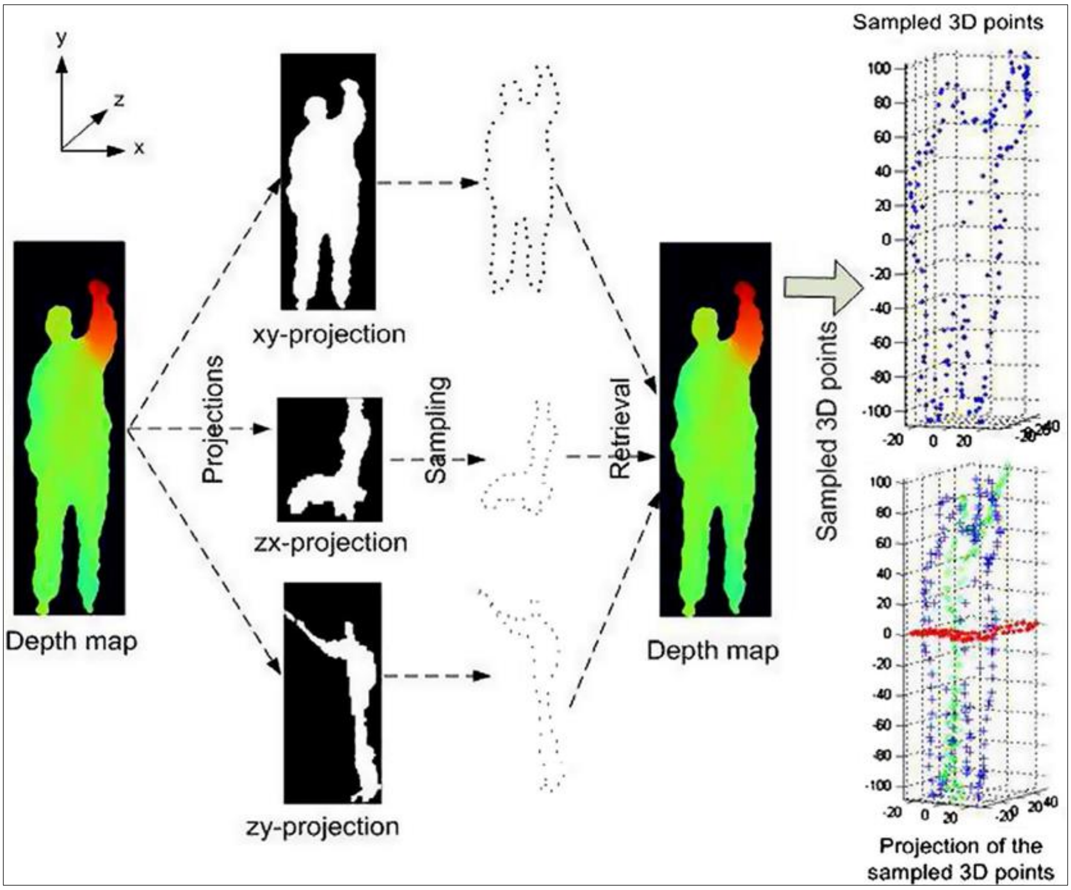

Figure 2 A depth map shows the process of sampling 3D representative points [29]

2.4. Skeleton Joints Feature

Many researchers have worked on extracting joints features from skeletons. They focused on automatic skeleton feature extraction and representing the human body using several human body parts. In Ben-Arie, [22], an efficient automatic skeleton feature extraction was employed to extract robust features from a target object. The skeleton joint features are invariant to the camera position and subject. [23] extracted the 3D body joint locations from depth images using an object recognition scheme. The human body is labelled as body parts based on the per-pixel classification results. The 3D joints Cartesian can be mapped to a spherical Coordinate System (r, θ, ϕ), where θ is the inclination from the z-axis, ϕ is the azimuth from the x-axis in the xy-plane, and r is the radius which is used for action recognition [24]. Kinect device provides the easiest access to skeleton joints locations (see Figure.3). The most discriminative feature is the pair-wise joint location difference feature. By computing the difference between the joint positions from the existing (or current) frame and former frame, one can get the joint motion between the two frames. Other researchers supposed Eigen Joints based action recognition which merges action information including static posture, motion and offset for each frame [25].

The location of certain joints can be related to certain types of activities. Joint angles between every pair of joints are used to detect abnormal actions (falling) of elderly people living in the community [26]. The joint orientation feature is stable to human body volume. [27] built a feature vector using the joint orientation of each body part. The 3D skeletal joint locations are obtained from Kinect depth maps using the method in [28] They presented a novel approach for human action recognition with histograms of 3D joint locations (HOJ3D).

Compared to the features from 3D silhouettes, the skeletal joint features are invariant to the camera location and subject appearance. In a good setting, the skeletal tracking algorithms can accurately extract the joint positions from either the front view, side view, or back view making the skeletal joint feature-based algorithm view-invariant. Furthermore, the joint orientation feature is invariant to human body size. If the skeletal joint locations of multiple persons are known, the features can be extended to model person-to-person interaction. The recognition scheme based on the skeletal joint features is better at modelling finer activities compared to the 3D silhouette-based algorithms. The limitation of the skeletal feature is that it does not give information about the surrounding objects. When modelling person-to-object interaction, object detection and tracking have to be combined. In addition, the current algorithm to estimate skeleton joint position from the depth image is not perfect. The skeleton tracking from the Kinect works well when the human subject is in an upright position facing the camera and there are no occlusions. However, the result is not very reliable when the human body is partly in view, the person touches the background, or the person is not in an upright position (e.g. patient lying on a bed). In surveillance or senior home monitoring scenarios, the camera is mounted in a higher location and the subjects are not facing the camera, this may also create difficulty for the skeletal estimation algorithm.

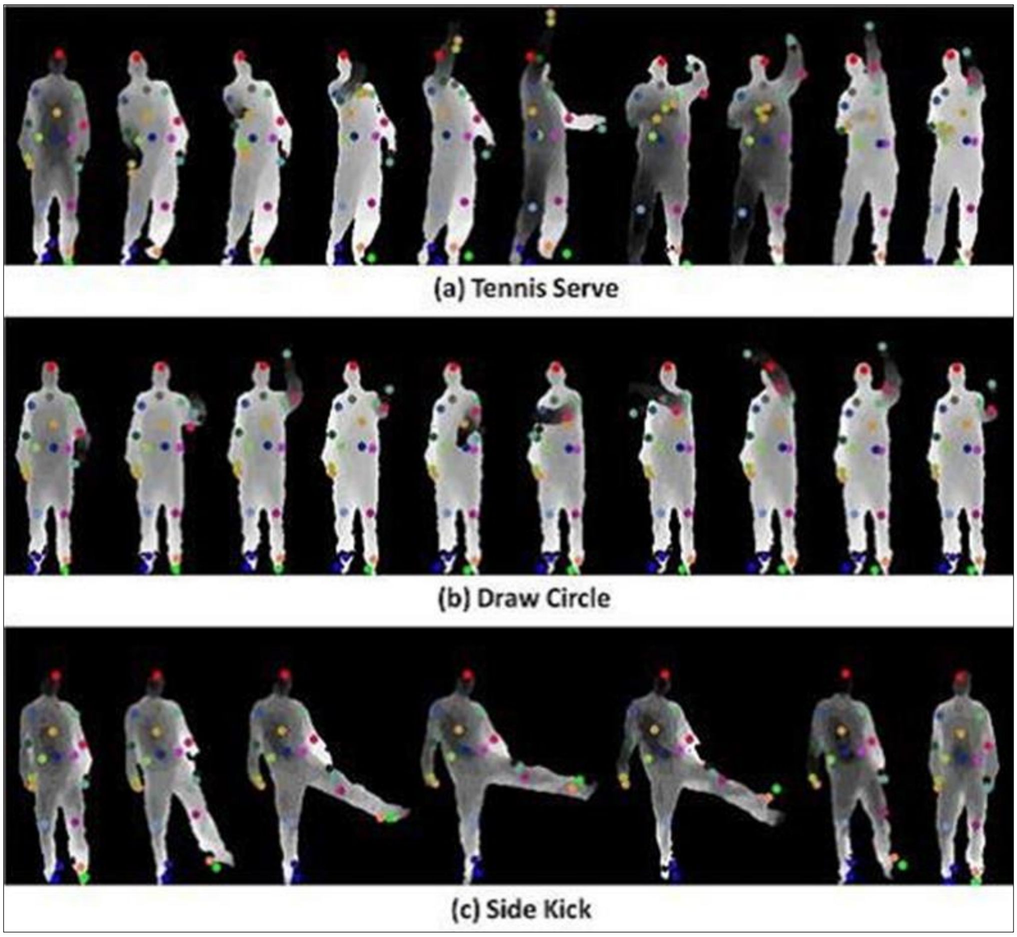

Figure 3 Examples of the sequences of depth maps for actions of (a) Tennis Serve, (b) Draw Circle, and (c) Side Kick [48]

2.5. Optical Flow-Based Features

Some researchers have used Gaussian Mixture Model (GMM) to identify the probability density function of the silhouette features for the actions [29]. Meanwhile, others have used motion flow patterns to represent human actions as in [30] and [31] Optical flow is the distribution of apparent velocities of movement of brightness patterns in an image, which arises both from the relative objects‘ and the viewer‘s motion. It is widely used in intensity images for motion detection, object segmentation and stereo disparity measurement. Also, it is a popular feature in activity recognition from videos [32] [21]. When multiple cameras are available, the integration over different viewpoints allows a 3D motion field, the scene flow [21] However, intensity variations alone are not sufficient to estimate motion and additional constraints such as smoothness must be introduced in most scenarios. Some of the promising works on estimating 3D scene flow from stereoscopic. Motion patterns are calculated based on the estimation of the motion parameters. Motion is calculated based on the optical flow of the consecutive video frames. Another type of feature is a local descriptor feature such as scale-invariant-feature-transform (SIFT) and Histogram of Oriented Gradient (HOG) as mentioned in [33]. Histogram Oriented Gradient (HOG) is extracted from a sequence of consecutive video frames and is used to create a Histogram of Oriented Gradient Pattern History (HOGPH).

Volume Motion Template (VMT) has been proposed by [34]. This method solves the problem of 2D-based human action recognition. VMT is an extension of Motion History Image (MHI) using stereo-depth images. Each human action is modelled with a Projected Motion Template (PMT) by projecting a Volume Motion Template (VMT) onto a 2D plane orthogonal to an optimal virtual viewpoint. The main reason for this is to exclude any camera viewpoint problem. It is only suitable for very simple gestures or actions without a lot of occlusion. Hence, this method needs to deal with complex actions in further research. Due to its robustness anti variation in speed or style in running action, an exemplar-based sequential single-layered approach is applied by using Dynamic Time Warping [35]. They performed body part tracking using a depth camera to improve human joints' body part information in a 3D real-world coordinate system and also to improve the recognition rate. Overall, the exploration of 3D optical flow or scene flow using RGB and depth imagery has been quite limited. Compared to the success of traditional 2D optical flow, the research on scene flow is still in its preliminary stage [21]. Currently, 3D scene flow is often computed for all the 3D points for the subject or scene, resulting in a large computational cost. Computing the 3D scene flow with real-time performance is a challenging task. We can imagine after the emergence of more effective ways to compute 3D scene flow, it has the potential to be a more popular type of feature for human action recognition and benefit more applications.

2.6. Recognition using local 3D Occupancy Features

Instead of representing the depth video as 3D spatio-temporal volume, the points may be projected to the 4D (x, y, z, and t) space. In 4D space, some locations will be occupied by the data points from the video, i.e. the points that the sensor captured from the real world, those locations will have a value of 1, others 0 [21]. In general, the local occupancy pattern is very sparse, that is, the majority of its elements are zero. The local occupancy pattern has been proposed individually by several researchers for activity recognition. In fact, the local occupancy feature can be defined in the (x, y, z) space or (x, y, z, and t), the former one describes the local depth appearance at a certain time instant while the latter describes the local ato mic events within a certain time range [21]. [16] Design 3D Local Occupancy Patterns (LOP). This new feature is also able to capture the relations and interaction between the human body parts and the environmental objects. For instance, the person is drinking a cup of water. When the person fetches the cup, the space around his/her hand is taken by the cup. Later on, when the person lifts the cup to his/her mouth, the space around both the hand and the head is taken. This information can be useful to distinguish this interaction and to recognize the drinking action from other actions. In genral, the number of possible simple features is so large that we are not able to enumerate all of them. The local occupancy feature defined in the (x, y, z, and t) space is same as local spatio-temporal features in that they both describe local “appearance“ in the space–time domains. Local spatio-temporal features treat the z dimension as “pixel values“ in the (x, y, t) volume while local occupancy patterns project the data onto a (x, y, z, and t) 4D space.

3. Classification Approaches for Human Activity Recognition

The classification adoption stage is performed after the data dimensionality reduction in the feature extraction stage. Numerous classification and detection algorithms have been validated for the purpose of human action recognition. Some of them are based on the dynamics time wrapping (DTW) [35]. This technique can measure similarities between two patterns regardless of their dynamic attributes changes such as velocity or acceleration. However, the efficiency of these methods degrades when the number of classes increases. A Hough-transform based voting framework is developed by [36]. The high dimensional space was split into two lower dimensional spaces and low-level features were extracted. However, the work needs to be tested on more sophisticated features.

Other machine learning and artificial neural network based models for this application are support vector machine, decision tree, and Naïve Bayes. RGB-D Kinect based human action recognition has been adapted and a comparative analysis of four methods is provided: Back-propagation Neural Network (BPNN), Naïve Bayes, decision tree, and support vector machine. These classifiers were developed based on body features (joints of human skeleton).

3.1. Dynamic Bayesian Network (DBN)

A Bayesian network is a probabilistic graphical model which is a type of statistical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). Dynamic Bayesian Network, on the other hand, is a Bayesian network with the same structure unrolled in the time axis [37] as shown in Figure 4. To cooperatively recognize the activity and environment of a person, a hierarchical dynamic Bayesian network is proposed [38]. The proposed model produced about 10% absolute enhancement in accuracy over the existing systems. In [39] a Bayesian approach is proposed to simultaneously estimate the object type, the performed action and the type of interaction between them. This approach used to improve the recognition of objects and actions. Another Bayesian approach for recognizing human activities is proposed in In [40], it relies only on objects. In [41] a Bayesian Network models the interaction motions between paired objects in a human object way and uses the motion models to improve the object recognition reliability. In this approach actions are not taken into account. Meanwhile, nursing activities are recognized from nurses interactions with tools and materials using a Dynamic Bayesian Network (DBN) [42].

There are two drawback in dynamic Bayesian networks. The limitation is the computational difficulty of exploring previous unknown network.

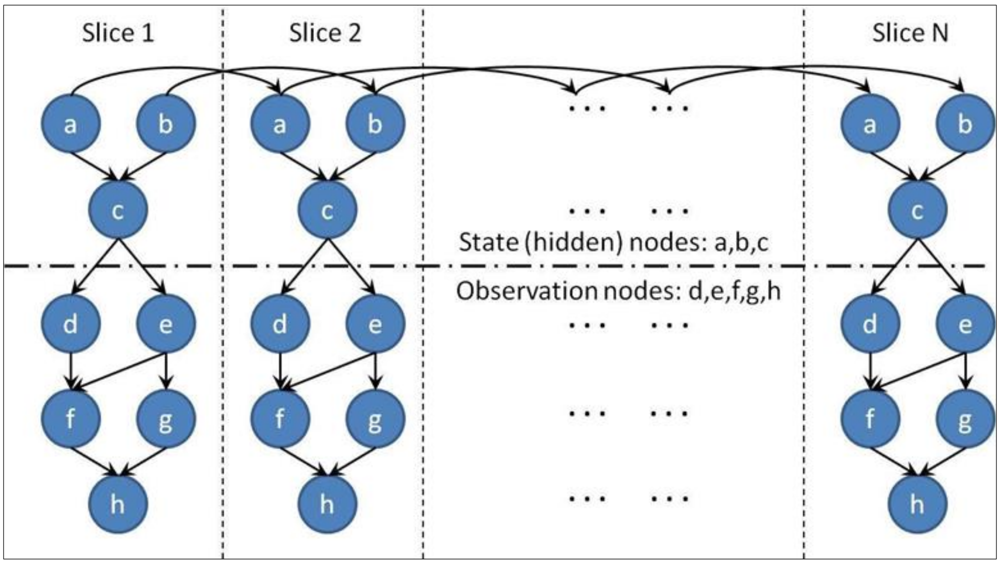

Figure 4 An example of a dynamic Bayesian network (DBN) is unrolled in time axis with state (hidden) notes and observation nodes [43]

3.2. Hidden Markov Model (HMM)

Hidden markov model (HMM) is a statistical Markov model where the modeled system is expected to be a Markov process with undetected states. A HMM can be identified by three terms. The first term P $(x0)$ is the initial probability of hidden states. The second term (A) is the transition matrix, which states a change probability from one hidden state to another hidden state. The third term (B) is the performance matrix, which identifies the probability of the detected symbol given a hidden state [47], see Figure 5 A real-time system is proposed for dynamic hand gesture recognition based on action graph, which shares similar robust properties with standard Hidden Markov Models (HMM) [46]. By accepting states shared among various gestures, less training data is required for this approach.

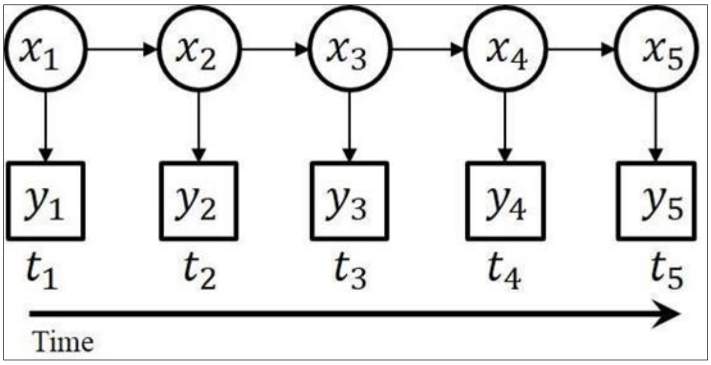

Figure 5 A hidden Markov Model (HMM) inference graph [44]

The depth perception algorithm based on the Kinect depth sensor for performing 3D hand tracking is introduced by [45]. For segmenting the 3D hand gesture from the hand motion trajectory, the authors proposed a novel start/end point detection method. They implemented HMM to categorize and model the hand gesture sequences and the identified gestures are changed into control commands for the collaboration with a robot. The results show high robustness but as result of insufficient performance of outdoor camera sensor, it is recommended to build a stereo vision system. The limitation of a hidden markov model in representation of sequence of images. The representation is not only inefficient but also difficult to interrupt.

4. Conclusions

The relevant efforts on human activity recognition have been covered in this article, along with a focus on certain feature types: space-time features, frequency features, local descriptors, optical flow-based, and skeleton joints. The paper also examined categorization strategies for identifying human activity. One difficult area of pattern recognition and machine learning is computer vision-based human activity recognition. While many efforts and innovative methods have been made in the field of human activity recognition, there is still much work to be done in developing an algorithm that would perform as error-free as possible in unrestricted environments like abnormal activity recognition, crowd behaviour, and multiple-person interaction. Moreover, in the available literature, most research is focused on common features such as spatial domain and frequency domain features. Such features, however, are lack context and not comprehensive.

Compliance with ethical standards

Acknowledgements

The author would like to thank the Mustansiriyah University (www.uomustansiriyah.edu.iq) in Baghdad - Iraq for its support in the present work.

References

[1] Subramanya, A., Raj, A., Bilmes, J., & Fox, D. (2006). Hierarchical models for activity recognition. In 2006 IEEE 8th Workshop on Multimedia Signal Processing (pp. 233– 237). IEEE

[2] Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., & Li, J. (2009). Hierarchical spatiotemporal context modeling for action recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2009 .CVPR 2009. (pp. 2004–2011). IEEE.

[3] Ikizler, N., & Duygulu, P. (2009). Histogram of oriented rectangles: A new pose descriptor for human action recognition. Image and Vision Computing, 27(10), 1515–1526.

[4] Kovashka, A., & Grauman, K. (Blank). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In 2010 IEEE Conference onComputer Vision and Pattern Recognition (CVPR) (pp. 2046–2053). IEEE.

[5] Lan, T., Zhu, Y., Zamir, A. R., & Savarese, S. (2015). Action Recognition by Hierarchical Mid-level Action Elements. In The IEEE International Conference on Computer Vision (ICCV) (pp. 4552–4560). IEEE.

[6] Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as SpaceTime Shapes. In Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005 (Vol. 29, pp. 1395–1402). IEEE.

[7] Mokhber, A., Achard, C., & Milgram, M. (2008). Recognition of human behavior by space-time silhouette characterization. Pattern Recognition Letters, 29(1), 81-89.

[8] Junejo, I. N., Junejo, K. N., & Aghbari, Z. A. (2014). Silhouette-based human action recognition using SAX-Shapes. Visual Computer, 30(3), 259–269.

[9] Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3D points. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 9–14). IEEE.

[10] Ni, B., Wang, G., & Moulin, P. (2013). GBD-HuDaAct: A color-depth video database for human daily activity recognition. In A. Fossati, J. Gall, H. Grabner, X. Ren, & K. Konolige (Eds.), Consumer Depth Cameras for Computer Vision (Vol. 47, pp. 193– 208). Springer London.

[11] Kurakin, A., Zhang, Z., & Liu, Z. (2012). A real time system for dynamic hand gesture recognition with a depth sensor. In 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO) (pp. 1975–1979). IEEE.

[12] Altun, K., & Barshan, B. (2010, August). Human activity recognition using inertial/magnetic sensor units. In International Workshop on Human Behavior Understanding (pp. 38-51). Springer Berlin Heidelberg.

[13] Kumari, S., & Mitra, S. K. (2011). Human Action Recognition Using DFT. In 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (pp. 239–242). IEEE.

[14] Nassar, H., El-Taweel, G., & Mahmoud, E. (2010). A novel feature extraction scheme for human gait recognition. International Journal of Image and Graphics, 10(04), 575- 587.

[15] Willems, G., Tuytelaars, T., & Van Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. Computer Vision–ECCV 2008, 650-663.

[16] Wang, J., Liu, Z., Chorowski, J., Chen, Z., & Wu, Y. (2012). Robust 3d action recognition with random occupancy patterns. In Computer vision–ECCV 2012 (pp. 872-885). Springer Berlin Heidelberg.

[17] Cheng, Z., Qin, L., Ye, Y., Huang, Q., & Tian, Q. (2012). Human Daily Action Analysis with Multi-view and Color-Depth Data. In A. Fusiello, V. Murino, & R. Cucchiara (Eds.), Computer Vision – ECCV 2012. Workshops and Demonstrations (pp. 52–61). Springer Berlin Heidelberg.

[18] Xia, L., & Aggarwal, J. K. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2834–2841). IEEE.

[19] Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision, 64(2), 107–123.

[20] Dollar, P., Rabaud, V., & Belongie, S. (2005). Behavior Recognition via Sparse SpatioTemporal Features. In 2nd Joint IEEE International Workshop on Visual Surveillance 116 and Performance Evaluation of Tracking and Surveillance, 2005 (pp. 65–72). IEEE

[21] Aggarwal, J. K., & Xia, L. (2014). Human activity recognition from 3d data: A review. Pattern Recognition Letters, 48, 70-80.

[22] Ben-Arie, J., Wang, Z., Pandit, P., & Rajaram, S. (2002). Human activity recognition using multidimensional indexing (Vol. 24, pp. 1091–1104). IEEE.

[23] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., … Blake, A. (2013). Real-time human pose recognition in parts from single depth images. Communications of the ACM, 56(1), 116–124.

[24] Ding, W., Liu, K., Cheng, F., & Zhang, J. (2015). STFC: Spatio-temporal feature chain for skeleton-based human action recognition. Journal of Visual Communication and Image Representation, 26, 329–337.

[25] Masood, S. Z., Ellis, C., Nagaraja, A., Tappen, M. F., Laviola, J. J., & Sukthankar, R. (2011). Measuring and reducing observational latency when recognizing actions. In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (pp. 422–429). IEEE.

[26] Zhang, C., & Tian, Y. (2012). RGB-D Camera-based Daily Living Activity Recognition. Journal of Computer Vision and Image Processing, 2(4), 12.

[27] Sempena, S., Maulidevi, N. U., & Aryan, P. R. (2011, July). Human action recognition using dynamic time warping. In Electrical Engineering and Informatics (ICEEI), 2011 International Conference on (pp. 1-5). IEEE.

[28] Xia, L., Chen, C. C., & Aggarwal, J. K. (2012, June). View invariant human action recognition using histograms of 3d joints. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on (pp. 20-27). IEEE.

[29] Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3D points. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 9–14). IEEE.

[30] Kumari, S., & Mitra, S. K. (2011). Human Action Recognition Using DFT. In 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (pp. 239–242). IEEE.

[31] Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994). Systems and Experiment Performance of Optical Flow Techniques. International Journal of Computer Vision, 12(1), 43–77.

[32] Chaudhry, R., Ravichandran, A., Hager, G., & Vidal, R. (2009, June). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In computer vision and pattern recognition, 2009. CVPR 2009. IEEE Conference on (pp. 1932-1939). IEEE.

[33] Fanello, S. R., Gori, I., Metta, G., & Odone, F. (2013). Keep It Simple And Sparse : RealTime Action Recognition. Journal of Machine Learning Research, 14(1), 2617–2640.

[34] Roh, M. C., Shin, H. K., & Lee, S. W. (2010). View-independent human action recognition with Volume Motion Template on single stereo camera. Pattern Recognition Letters, 31(7), 639–647.

[35] Sempena, S., Maulidevi, N. U., & Aryan, P. R. (2011, July). Human action recognition using dynamic time warping. In Electrical Engineering and Informatics (ICEEI), 2011 International Conference on (pp. 1-5). IEEE.

[36] Yao, A., Gall, J., & Van Gool, L. (2010). A hough transform-based voting framework for action recognition. In 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2061–2068). IEEE.

[37] Murphy, K. P. (2002). Dynamic bayesian networks: representation, inference and learning. University of California, Berkeley.

[38] Subramanya, A., Raj, A., Bilmes, J., & Fox, D. (2006). Hierarchical models for activity recognition. In 2006 IEEE 8th Workshop on Multimedia Signal Processing (pp. 233–237). IEEE

[39] Gupta, A., & Davis, L. S. (2007, June). Objects in action: An approach for combining action understanding and object perception. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on (pp. 1-8). IEEE.

[40] Osmani, V., Balasubramaniam, S., & Botvich, D. (2007, September). A bayesian network and rule-base approach towards activity inference. In Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th (pp. 254-258). IEEE.

[41] Ren, S., & Sun, Y. (2013, January). Human-object-object-interaction affordance. In Robot Vision (WORV), 2013 IEEE Workshop on (pp. 1-6). IEEE.

[42] Inomata, T., Naya, F., Kuwahara, N., Hattori, F., & Kogure, K. (2009, May). Activity recognition from interactions with objects using dynamic bayesian network. In Proceedings of the 3rd ACM International Workshop on Context-Awareness for SelfManaging Systems (pp. 39-42).

[43] Luo, Y., Wu, T. Der, & Hwang, J. N. (2003). Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks. Computer Vision and Image Understanding, 92(2-3), 196–216.

[44] Brand, M., Oliver, N., & Pentland, A. (1997). Coupled hiddenMarkov models for complex action recognition. In 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997 (pp. 994–999). IEEE.

[45] Xu, D., Chen, Y.-L., Lin, C., Kong, X., & Wu, X. (2012). Real-time dynamic gesture recognition system based on depth perception for robot navigation. In 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO) (pp. 689–694).

[46] Kurakin, A., Zhang, Z., & Liu, Z. (2012). A real time system for dynamic hand gesture recognition with a depth sensor. In 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO) (pp. 1975–1979). IEEE.

[47] Chua, T. W., Leman, K., & Pham, N. T. (2011). Human action recognition via sum-rule fusion of fuzzy K-Nearest Neighbor classifiers. In IEEE International Conference on Fuzzy Systems (pp. 484–489). IEEE.

[48] Yang, X., & Tian, Y. (2012). EigenJoints-based Action Recognition Using Naïve-BayesNearest-Neighbor. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 14–19). IEEE.