ISSN 2395-1621

Human Activity Recognition: A Detailed Study

[#1]Mr. Bhushan Nanche, [#2]Hiren Jayantilal Dand, [#3]Dr. Bhagyashree Tingare

[1]Research Scholar, Computer Engineering Department, JJTU, Jhunjhunu, Rajasthan.
[2]Guide, Computer Engineering Department, JJTU, Jhunjhunu, Rajasthan
[3]CO–Guide, AI & DS Department, DYPCOE, Akurdi, Pune, Maharashtra

ABSTRACT

Recognizing human activities from video sequences or still images is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and appearance. Many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. In this work, we provide a detailed review of recent and state-of-the-art research advances in the field of human activity classification. We propose a categorization of human activity methodologies and discuss their advantages and limitations. In particular, we divide human activity classification methods into two large categories according to whether they use data from different modalities or not. Then, each of these categories is further analyzed into sub-categories, which reflect how they model human activities and what type of activities they are interested in. Moreover, we provide a comprehensive analysis of the existing, publicly available human activity classification datasets and examine the requirements for an ideal human activity recognition dataset. Finally, we report the characteristics of future research directions and present some open issues on human activity recognition.

Keywords: Deep Belief Neural Network; Dragonfly optimization; fuzzy classifier; activity recognition; Gaussian Mixture Model Clustering.

ARTICLE INFO

Article History

Received: 15th December 2022

Received in revised form :

15th December 2022

Accepted:20th December 2022

Published online :

25th December 2022

I. INTRODUCTION

Nowadays, human activity recognition is a vital area in computer vision research. The applications of activity recognition include patient monitoring systems, video surveillance systems [1], and systems, which involve interactions among electronic devices and persons, like human–computer interfaces [2]. Human activity recognition from still images or video sequences is a difficult task owing to problems like appearance, lighting, viewpoint, changes in scale, partial occlusion, and background clutter. The activity recognition is actively concentrated and the digital cameras are employed for capturing the regular activities of the humans [3, 4]. Due to the employment of the digital camera in monitoring the daily activities, the video sources [5] are spread widely on the internet, and the solution for solving the issues related to the classification of the action classes is attained. However, the processing is time-consuming if the manual power is employed for the action classification [6]. The process of action recognition finds a valuable application in the field of entertainment, sports, smart home, and healthcare. The activities are categorized such that the irregular movements like walking running, jogging, and so on, which are already available in the considerable movement range, are traced as the normal behaviors [7]. The human activities are categorized into four types, namely group activities, interactions, actions, and gestures. Group activities completed by groups contain multiple persons. Interactions involve two or more persons. Actions are performed by a single person, which include multiple gestures, like reading, walking, and so on. Gestures are simple movements of a body part of a person [2]. The main aim of the human activity recognition [8–10] is to automate the analysis of ongoing activities using an unknown video [11, 12]. The video is read using the human activity recognition system, and it is segmented to possess a single execution such that the individual segment represents the single activity and the activity recognition functions to classify the activity in the video [13]. The sequence of images is used for the human activity inference that is named as the Space-Time Volume (STV) [14]. The Spacetime approaches consider the video input as a 3-D (XYT) video volume and assume an action to a particular class of The human activity recognition [8, 16] is classified as model-based recognition system and the model free-based recognition system [15, 17]. On analyzing the model-based approaches for the human activity recognition, it is clear that there exists a trade-off between the retrieving information and the computational cost and robustness of the method that makes the model-free methods highly advantageous [4]. The advantages of the model-free methods are due to the posture, global motion, and the local

motionbased features. These approaches depend on the cross-view, templates of activity or the texture of the activities [6]. The common methods of activity recognition are to develop the templates of activities based on the temporal templates of motion energy images and motion history images. The pattern extraction methods like the Local Binary Patterns (LBP) extracts the dynamic texture patterns for recognizing the activity of the humans [18]. The improved version of the LBP is the uniform LBP that possesses the high discriminative capacity to recognize the objects. The usage of the uniform patterns not only reduces the length of the LBP, but also improves the performance of classifiers [19]. Many application developers and researchers have developed an activity recognition system. Machine learning [20] based techniques have been extensively implemented for the sensor-based activity recognition. Anyhow, activity recognition becomes more challenging and cost-intensive. The paper introduces a method for performing the activity recognition in videos. The proposed method is called as Fuzzy-DDBN that is the combination of the fuzzy and the DDBN, where the DDBN is the optimization process formed with the integration of the dragonfly optimization in the Deep Belief Neural network (DBN). Initially, the video is segmented to form the single object represented in a frame, and the keyframe is extracted using the similarity measure, termed as the Bhattacharya similarity coefficient.

The Bhattacharya similarity coefficient retrieves the keyframes that are used for the extraction of the features. The keyframes are required to offer an improved feature extraction such that the recognition becomes effective and faster. The features, such as the motion pattern and the other local features, are extracted using the SIFT and the STI descriptors. The extracted features are fed to the proposed Fuzzy-DDBN such that the effective classification is enabled. The proposed FuzzyDDBN uses the GMM clustering to cluster the data points present in the feature vector. Accordingly, the action class is determined.

Human activity recognition plays a significant role in human-to-human interaction and interpersonal relations. Because it provides information about the identity of a person, their personality, and psychological state, it is difficult to extract. The human ability to recognize another person's activities is one of the main subjects of study of the scientific areas of computer vision and machine learning. As a result of this research, many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system.

Among various classification techniques two main questions arise: "What action?" (i.e., the recognition problem) and "Where in the video?" (i.e., the localization problem). When attempting to recognize human activities, one must determine the kinetic states of a person, so that the computer can efficiently recognize this activity. Human activities, such as "walking" and "running," arise very naturally in daily life and are relatively easy to recognize. On the other hand, more complex activities, such as "peeling an apple," are more difficult to identify. Complex activities may be decomposed into other simpler activities, which are generally easier to recognize. Usually, the detection of objects in a scene may help to better understand human activities as it may provide useful information about the ongoing event (Gupta and Davis, 2007).

Most of the work in human activity recognition assumes a figure-centric scene of uncluttered background, where the actor is free to perform an activity. The development of a fully automated human activity recognition system, capable of classifying a person's activities with low error, is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting and appearance, and frame resolution. In addition, annotating behavioral roles is time consuming and requires knowledge of the specific event. Moreover, intra- and interclass similarities make the problem amply challenging. That is, actions within the same class may be expressed by different people with different body movements, and actions between different classes may be difficult to distinguish as they may be represented by similar information. The way that humans perform an activity depends on their habits, and this makes the problem of identifying the underlying activity quite difficult to determine. Also, the construction of a visual model for learning and analyzing human movements in real time with inadequate benchmark datasets for evaluation is challenging tasks.

To overcome these problems, a task is required that consists of three components, namely: (i) background subtraction (Elgammal et al., 2002; Mumtaz et al., 2014), in which the system attempts to separate the parts of the image that are invariant over time (background) from the objects that are moving or changing (foreground); (ii) human tracking, in which the system locates human motion over time (Liu et al., 2010; Wang et al., 2013; Yan et al., 2014); and (iii) human action and object detection (Pirsiavash and Ramanan, 2012; Gan et al., 2015; Jainy et al., 2015), in which the system is able to localize a human activity in an image.

The goal of human activity recognition is to examine activities from video sequences or still images. Motivated by this fact, human activity recognition systems aim to correctly classify input data into its underlying activity category. Depending on their complexity, human activities are categorized into: (i) gestures; (ii) atomic actions; (iii) human-to-object or human-to-human interactions; (iv) group actions; (v) behaviors; and (vi) events. Figure 1 visualizes the decomposition of human activities according to their complexity.

Figure 1. Decomposition of human activities.

I. Human Activity Categorization

The human activity categorization problem has remained a challenging task in computer vision for more than two decades. Previous works on characterizing human behavior have shown great potential in this area. First, we categorize the human activity recognition methods into two main categories:
(i) unimodal and (ii) multimodal activity recognition methods according to the nature of sensor data they employ. Then, each of these two categories is further analyzed into sub-categories depending on how they model human activities. Thus, we propose a hierarchical classification of the human activity recognition methods, which is depicted in Figure 2.

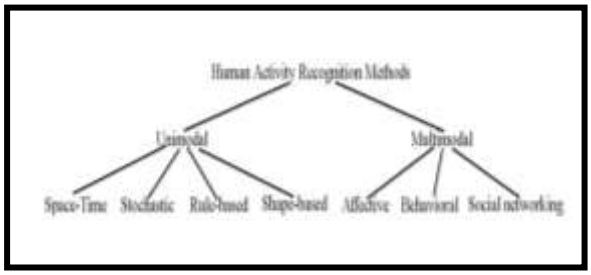

Figure 2. Proposed hierarchical categorization of human activity recognition methods**.**

Unimodal methods represent human activities from data of a single modality, such as images, and they are further categorized as: (i) space-time, (ii) stochastic, (iii) rule-based, and (iv) shape-based methods.Space-time methods involve activity recognition methods, which represent human activities as a set of spatiotemporal features (Shabani et al., 2011; Li and Zickler, 2012) or trajectories (Li et al., 2012; Vrigkas et al., 2013). Stochastic methods recognize activities by applying statistical models to represent human actions (e.g., hidden Markov models) (Lan et al., 2011; Iosifidis et al., 2012a). Rule-based methods use a set of rules to describe human activities (Morariu and Davis, 2011; Chen and Grauman, 2012). Shape-based methods efficiently represent activities with high-level reasoning by modeling the motion of human body parts (Sigal et al., 2012b; Tran et al., 2012).

Multimodal methods combine features collected from different sources (Wu et al., 2013) and are classified into three categories: (i) affective, (ii) behavioral, and (iii) social networking methods.

Affective methods represent human activities according to emotional communications and the affective state of a person (Liu et al., 2011b; Martinez et al., 2014). Behavioral methods aim to recognize behavioral attributes, non-verbal multimodal cues, such as gestures, facial expressions, and auditory cues (Song et al., 2012a; Vrigkas et al., 2014b). Finally, social networking methods model the characteristics and the behavior of humans in several layers of human-to-human interactions in social events from gestures, body motion, and speech (Patron-Perez et al., 2012; Marín-Jiménez et al., 2014).

Usually, the terms “activity” and “behavior” are used interchangeably in the literature (Castellano et al., 2007; Song et al., 2012a). In this survey, we differentiate between these two terms in the sense that the term “activity” is used to describe a sequence of actions that correspond to specific body motion. On the other hand, the term “behavior” is used to characterize both activities and events that are associated with gestures, emotional states, facial expressions, and auditory cues of a single person. Some representative frames that summarize the main human action classes are depicted in Figure 3.

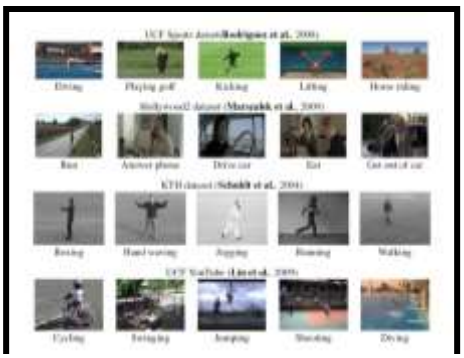

Figure 3. Representative frames of the main human action classes for various datasets.

5. Multimodal Methods

Recently, much attention has been focused on multimodal activity recognition methods. An event can be described by different types of features that provide more and useful information. In this context, several multimodal methods are based on feature fusion, which can be expressed by two different strategies: early fusion and late fusion. The easiest way to gain the benefits of multiple features is to directly concatenate features in a larger feature vector and then learn the underlying action (Sun et al., 2009). This feature fusion technique may improve recognition performance, but the new feature vector is of much larger dimension.

Multimodal cues are usually correlated in time, thus a temporal association of the underlying event and the different modalities is an important issue for understanding the data. In that context, audio-visual analysis is used in many applications not only for audio-visual synchronization (Lichtenauer et al., 2011) but also for tracking (Perez et al., 2004) and activity recognition (Wu et al., 2013). Multimodal methods are classified into three categories: (i) affective methods, (ii) behavioral methods, and (iii) methods based on social networking. Multimodal methods describe atomic actions or interactions that may correspond to affective states of a person with whom he/she communicates and depend on emotions and/or body movements.

II. Conclusion

In this survey, we carried out a comprehensive study of state-of-the-art methods of human activity recognition and proposed a hierarchical taxonomy for classifying these methods. We surveyed different approaches, which were classified into two broad categories (unimodal and multimodal) according to the source channel each of these approaches employ to recognize human activities. We discussed unimodal approaches and provided an internal categorization of these methods, which were developed for

analyzing gesture, atomic actions, and more complex activities, either directly or employing activity decomposition into simpler actions. We also presented multimodal approaches for the analysis of human social behaviors and interactions. We discussed the different levels of representation of feature modalities and reported the limitations and advantages for each representation. A comprehensive review of existing human activity classification benchmarks was also presented and we examined the challenges of data acquisition to the problem of understanding human activity. Finally, we provided the characteristics of building an ideal human activity recognition system.

Most of the existing studies in this field failed to efficiently describe human activities in a concise and informative way as they introduce limitations concerning computational issues. The gap of a complete representation of human activities and the corresponding data collection and annotation is still a challenging and unbridged problem. In particular, we may conclude that despite the tremendous increase of human understanding methods, many problems still remain open, including modeling of human poses, handling occlusions, and annotating data.

III. REFRENCES

[1] Sudhakar R and Letitia S 2015 Motion estimation scheme for video coding using hybrid discrete cosine transform and modified unsymmetrical-cross multi hexagon-grid search algorithm. Middle-East J. Sci. Res. 23(5): 848–855

[2] Aggarwal J K and Ryoo M S 2011 Human activity analysis: a review. ACM Comput. Surv. 43(3): 16

[3] Liu L, Wang S, Sud G, Huang Z and Liu M 2017 Towards complex activity recognition using a Bayesian networkbased probabilistic generative framework. Pattern Recognit. 68: 295–309

[4] He X, Cai D, Shao Y, Bao H and Han J 2011 Laplacian regularized gaussian mixture model for data clustering. IEEE Trans. Knowl. Data Eng. 23(9): 1406–1418

[5] Daga B S and Ghatol A A 2016 Detection of objects and activities in videos using spatial relations and ontology based approach in video database system. Int. J. Adv. Eng. Technol. 9(6): 640–650

[6] Maid A and Bhorge S B 2015 Automated human action recognition using machine learning. Int. J. Adv. Eng. Res. Dev. 2(6)

[7] Ryoo M S and Matthies L 2016 First-person activity recognition: feature, temporal structure, and prediction. Int. J. Comput. Vis. 119(3): 307–328

[8] Chinimilli P T, Redkar S and Zhang W 2017 Human activity recognition using inertial measurement units and smart shoes. In: 2017 American Control Conference (ACC), pp. 1462–1467

[9] Zhuang N, Yusufu T, Ye J and Hua KA 2017 Group activity recognition with differential recurrent convolutional neural networks. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 526–531

[10] Zhang Y, Li X, Zhang J, Chen S, Zhou M, Farneth R A, Marsic I and Burd R S 2017 Poster abstract: CAR—a deep learning structure for concurrent activity recognition. In: 2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), pp. 299–300

[11] Ramasso E, Panagiotakis C, Pellerin D and Rombaut M 2008 Human action recognition in videos based on the Transferable Belief Model. Pattern Anal. Appl. 11(1): 1–19

[12] Savvaki S, Tsagkatakis G, Panousopoulou A and Tsakalides P 2017 Matrix and tensor completion on a human activity recognition framework. IEEE J. Biomed. Health Inform. 21(6): 1554–1561

[13] Wang Z, Wu D, Gravinac R, Fortinoc G, Jiangd Y and Tange K 2017 Kernel fusion based extreme learning machine for cross-location activity recognition. Inf. Fusion 37: 1–9

[14] Li J, Wu R, Zhao J and Ma Y 2017 Convolutional neural networks (CNN) for indoor human activity recognition using Ubisense system. In: 2017 29th Chinese Control and Decision Conference (CCDC), pp. 2068–2072

[15] Fu Y, Zhang T and Wang W 2017 Sparse coding-based space-time video representation for action recognition. Multimed. Tools Appl. 76(10): 12645–12658

[16] Doewes A, Swasono S E and Harjito B 2017 Feature selection on human activity recognition dataset using minimum redundancy maximum relevance. In: 2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), pp. 171–172

[17] Vanrell S R, Milone D H and Rufiner H L 2017 Assessment of homomorphic analysis for human activity recognition from acceleration signals. IEEE J. Biomed. Health Inform. 22(4): 1001–1010

[18] Ni Q, Pan Q, Du H, Cao C and Zhai Y 2017 A novel cluster head selection algorithm based on fuzzy clustering and particle swarm optimization. IEEE/ACM Trans. Comput. Biol. Bioinform. 14(1): 76–84

[19] Nigam S and Khare A 2016 Integration of moment invariants and uniform local binary patterns for human activity recognition in video sequences. Multimed. Tools Appl. 75(24): 17303–17332

[20] Valsalan P, Manimegalai S O and Augustine S 2017 Non invasive estimation of blood pressure using a linear regression model from the photoplethysmogram (PPG) signal. Perspectivas em Ciencia da Informacao 22(4):

[21] Gupta, A., and Davis, L. S. (2007). "Objects in action: an approach for combining action understanding and object perception," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Minneapolis, MN), 1–8.

[22] Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. S. (2002). Background and foreground modeling using nonparametric kernel density for visual surveillance. Proc. IEEE 90, 1151–1163. doi:10.1109/JPROC.2002.801448

[23] Mumtaz, A., Zhang, W., and Chan, A. B. (2014). "Joint motion segmentation and background estimation in dynamic scenes," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 368–375.

[24] Yan, X., Kakadiaris, I. A., and Shah, S. K. (2014). Modeling local behavior for predicting social interactions towards human tracking. Pattern Recognit. 47, 1626–1641. doi:10.1016/j.patcog.2013.10.019

[25] Yung, H. Y., Lee, S., Heo, Y. S., and Yun, I. D. (2015). "Random treewalk toward instantaneous 3D human pose estimation," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2467–2474.

[26] Lan, T., Sigal, L., and Mori, G. (2012a). "Social roles in hierarchical models for human activity recognition," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1354–1361.

[27] Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., and Mori, G. (2012b). Discriminative latent models for recognizing contextual group activities. IEEE Trans. Pattern Anal. Mach. Intell. 34, 1549–1562. doi:10.1109/TPAMI.2011.228

[28] Liu, J., Kuipers, B., and Savarese, S. (2011a). "Recognizing human actions by attributes," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3337–3344.

[29] Liu, N., Dellandréa, E., Tellez, B., and Chen, L. (2011b). "Associating textual features with visual ones to improve affective image classification," in Proc. International Conference on Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science, Vol. 6974 (Memphis, TN), 195–204.

[30] Wu, Q., Wang, Z., Deng, F., Chi, Z., and Feng, D. D. (2013). Realistic human action recognition with multimodal feature selection and fusion. IEEE Trans. Syst. Man Cybern. Syst. 43, 875–885. doi:10.1109/TSMCA.2012.2226575

[31] Bhagyashree Tingare, Dr. Prasadu Peddi, Dr. Prashant Kumbharkar, "Implementation of Virtual Dresing Room using Kinect along with OpenCV", International Journal of Mechanical Engineering,ISSN: 0974-5823 ,Vol. 7 (Special Issue, Jan.-Feb. 2022)

[32] Vrigkas, M., Karavasilis, V., Nikou, C., and Kakadiaris, I. A. (2013). "Action recognition by matching clustered trajectories of motion vectors," in Proc. International Conference on Computer Vision Theory and Applications (Barcelona), 112–117.

[33] Sun, X., Chen, M., and Hauptmann, A. (2009). "Action recognition via local descriptors and holistic features," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (Los Alamitos, CA), 58–65.

[34] Perez, P., Vermaak, J., and Blake, A. (2004). Data fusion for visual tracking with particles. Proc. IEEE 92, 495–513. doi:10.1109/JPROC.2003.823147

[35] Lichtenauer, J., Valstar, J. S. M., and Pantic, M. (2011). Cost-effective solution to synchronised audio-visual data capture using multiple sensors. Image Vis. Comput. 29, 666–680. doi:10.1016/j.imavis.2011.07.004

