Real-Time Human Action Recognition with Multimodal Dataset: A Study Review

Kapil Joshi, Ritesh Rastogi, Pooja Joshi, Harishchander Anandaram, Ashulekha Gupta, and Yasmin Makki Mohialden

Abstract Due to difficulties including a cluttered background, partial occlusion, and variations on dimensions, angle, illumination, or look, identifying human endeavors using video clips or still images is a challenging process. A numerous mechanism for recognizing activity is necessary for numerous applications, such as robotics, human–computer interaction, and video surveillance for characterizing human behavior. We outline a classification of human endeavor approaches and go through their benefits and drawbacks. In specifically, we classify categorization of human activity approaches into the two broad categories based on whether or not they make use of information from several modalities. This study covered a depth motion map-based approach to human recognizing an action. A motion map in depth created by adding up to the fullest differences with respect to the two following projections maps for each projection view over the course of the entire depth video series. The suggested approach is demonstrated to be computationally effective, enabling real-time operation. Results of the recognition using the dataset for Microsoft Research Action3D show that our method outperforms other methods.

Keywords Human action recognition · Depth motion map · Multimodal

K. Joshi (✉) · P. Joshi
Department of CSE, Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, India
e-mail: kapilengg0509@gmail.com

R. Rastogi
Department of IT, Noida Institute of Engineering and Technology, Greater Noida, India

H. Anandaram
Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India
e-mail: a_harishchander@cb.amrita.edu

A. Gupta
Department of Management Studies, Graphic Era (Deemed to Be University), Dehradun, India

Y. M. Mohialden
Computer Science Department, Mustansiriyah University, Baghdad, Iraq
e-mail: ymmiraq2009@uomustansiriyah.edu.iq

Y. Singh et al. (eds.), *Proceedings of International Conference on Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 1011,
https://doi.org/10.1007/978-981-99-0601-7_32

1 Introduction

Computer vision research is currently focused on human activity recognition. Action recognition has already been attempted using video sequences recorded by cameras. Recognizing human behaviors frequently uses spatiotemporal characteristics, e.g. [1]. Real-time depth data gathering is now possible because of advancements in image technology. Depth maps can also deliver 3D details for distinguish behavior that is challenging to characterize utilizing standard images and are less susceptible to changes with lighting conditions than traditional photos. Figure 1 shows two illustrations of the actions of golf swing and a kick forward, each with nine depth maps. Numerous studies on human action detection utilizing depth images have been conducted but since introduction inexpensive depth sensors, in particular, ASUS Xtion with Microsoft Kinect, e.g., [2]. Observed in, additional information is provided to complete action recognition by the 3D multiple objects of skeleton of a person that are calculated using depth photographs.

There are two key queries regarding various classification techniques which action? (specifically, the issue with recognizing) and "Where in the video?" (specifically, the localization issue). The kinetic states of such a person must be known when trying to recognize human activity because then the computer can do so effectively.

Examining actions from still photos or video clips is the aim of human activity recognition. This fact serves as the driving force behind human activity identification systems' quest to accurately classify data input into the relevant activity category. Different types of human behavior are six categories, depending in terms of

a) Golf Swing

b) Forward Kick

Fig. 1 Actions of a golf swing and a kick forward are examples of depth map sequences [3]

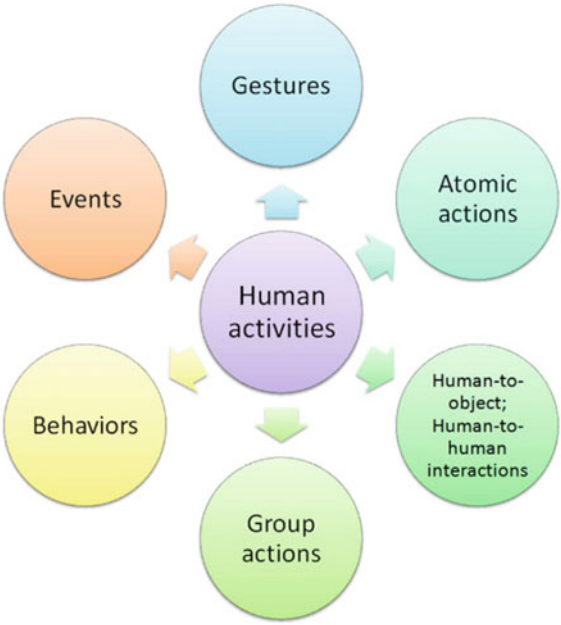

Fig. 2 Decomposition of human activities

complexity: gestures, atomic acts, human-to-human or object group actions, interactions, behaviors, and things. According to their complexity, human activities are divided up in Fig. 2.

The remainder of the essay the following structure Sect. 2 gives some historical context. The specifics of the depth motion maps' characteristics are explained in Sect. 3. Human activity categorization would be described in Sect. 4. Unimodal and multimodal approaches are covered in Sects. 5 and 6. Dataset collected is considered in Sect. 7. Section 8 also includes final remarks.

2 Background

For the purpose of identifying video clips of people acting recorded by conventional space-time-based RGB cameras, techniques such spatiotemporal space-time volumes characteristics, as well as trajectory, are extensively used. In [4], to recognize human action, spatiotemporal interest points and an SVM classifier were combined. Cuboids' descriptors accustomed to express actions. Activities in a series of videos were identified using SIFT-feature trajectories that were described in an order of three degrees of abstraction. In order to accomplish action categorization, several characteristics of local motion were assembled as spatiotemporal features from a bag (BoF) [3]. As motion templates to characterize the spatial and temporal properties of human motions in movies, motion energy images (MEIs) or motion-history images (MHIs)

were only launched in [5]. When computing dense motion flow using MHI is occurred then, a hierarchical extension was provided with correct accuracy. The sensitivity for recognition to variations in illumination is a significant drawback of adopting either depending on hue or intensity approaches, restricting the robustness in recognition. Research with action recognition dependent on depth data has expanded with the introduction of RGBD sensors. Skeletal joint locations are retrieved from depth pictures for skeleton-based techniques. A customized spherical coordinate system and histograms of 3D joint positions (HOJ3D) were used to create a view-invariant posture representation. With the use of LDA, reprojected HOJ3D were used and grouped around K-situation visual words. A continuous hidden Markov model was used to simulate the sequential evolutions of such visual words. Based on Eigen joints, a Naive Bayes Nearest-Neighbor (NBNN) classification was used to identify human behavior. (i.e., variations in joint position) integrating data on offset, motion, and still posture. Due to some errors in skeletal estimate, many skeleton-based techniques have limits. Additionally, many programmers do not always have access to the skeleton information.

To discriminate between various actions, several techniques require spatiotemporal data extraction information based on complete [6] collection of a depth map's point's series. The use of an action graph in a group was 3D points which was also used to describe body positions and describe the dynamics of actions. The 3D points' sample technique, however, produced a lot of data, necessitating a time-consuming training phase. To efficiently describe the body shape as well as movement information for distinguishing actions, an extent motions' histogram with a map on directional gradients (HOG) has been used. A weighted sampling strategy was used to extract random occupancy frequency (ROP) features from depth pictures. The characteristics were demonstrated to be robust to occlusion by using a sparse coding strategy to effectively encode random occupancy sequence features during action recognition. In order to preserve spatial and geographic context statement while managing intra-class conflict variability, 4D advanced patterns were being used as features. Then, for action recognition, a straightforward technical design here on cosine distance was applied. A hybrid system for action recognition method incorporating depth and the skeleton data was employed. Local occupancy patterns and 3D joint position were employed as features then, to characterize each action and account for intra-class variances; another action let accuracy of the model was learned.

3 Depth Motion Maps as Features

The 3D structure but also shape information can be recorded using a depth map. Alemayoh et al. [7] suggested to characterize the motion of an action by imposing depth pictures across three Cartesian orthogonal planes. Because it is computationally straightforward, the same strategy is used throughout the work while the method for getting DMMs is changed. In more detail, any 3D depths are frame also used like create three map v 2D mapped projections that represent the top, side, or front

perspectives

$$\text{Where } v = \{f, s, t\} \tag{1}$$

To illustrate (x, y, z) with in a frame depth z, the number of pixels in three projected maps is denoted by the value of depth in such an orthogonal coordinate system, z, x, and y, respectively.

Separated from, the actual distinction between these two separate maps before thresholding is used in this calculation to determine the motion energy for each projected map. The depth gesture map DMM_v is created in-depth video series N frame's worth by stacking all motion energies throughout the full sequence as follows:

$$\text{DMM}_v = \sum_{i=a}^{b} \left| \text{map}_v^i - \text{map}_v^{i-1} \right| \tag{2}$$

where i shows the frame index.

4 Human Activity Categorization

Over the past two decades, the categorization of human activities has remained a difficult job in computer vision. There is a lot of potential in this field based on earlier studies on describing human behavior. According to the type of sensor data they use, we first divide the acknowledgement of human action techniques into the two broad categories: (i) unimodal and (ii) multimodal identification system approaches. According on how they represent human activities, every one of those is two types, then further broken into smaller divisions. As a result, we suggest alternative classification of human activities in hierarchy techniques, as shown in Figs. 3 and 4.

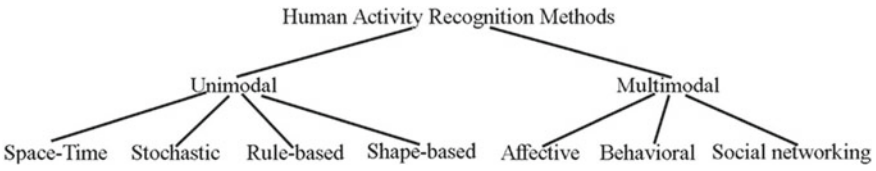

Fig. 3 Proposed hierarchical categorization of human activity recognition methods

Fig. 4 Representative frames of the main human action classes for various datasets [8]

5 Unimodal-Based Methods

Utilizing data from a single modality, single-modal identification of human action algorithms cites examples of human activity. The majority of current methods classifies the underlying activities' label using various classification models and show human activity as either a series of images elements collected from still images or video. For identifying human activities based upon motion features, unimodal techniques are appropriate. On the other hand, it can be difficult to identify the underlying class just from motion. How to maintain is the biggest challenge that the continuity of motion throughout duration of an action takes place uniformly or unevenly throughout a video sequence. Some approaches employ brief motion velocities; others track the optically flow features to employ the whole length on motion curves.

The four basic categories we use to categorize unimodal methods are (i) space-time, (ii) stochastic, (iii) rule-based, but also (iv) methods based on shapes. Depending on just the sort of representation each approach employs, every one of those sub-categories describes particular characteristics of the strategies for recognizing human activities (Fig. 5).

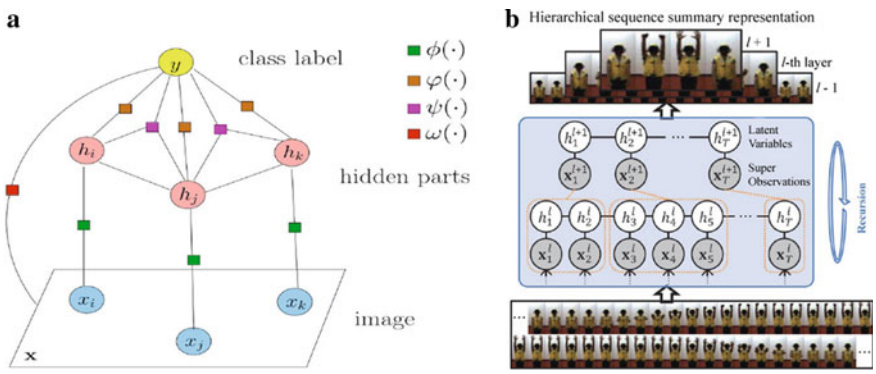

Fig. 5 Representative stochastic approaches for action recognition [9]

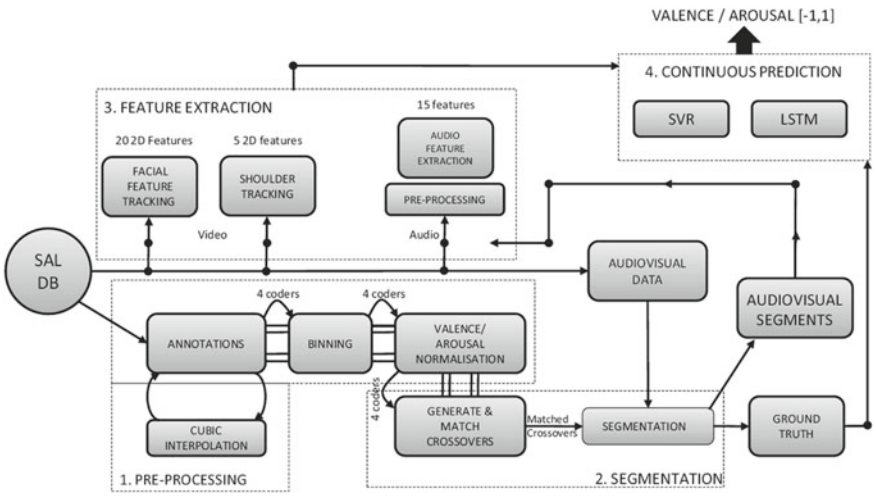

Fig. 6 Flow chart of multimodal emotion recognition [9]

6 Multimodal-Based Methods

Multimodal activity recognition techniques have received a lot of interest lately. Variety components that offer in addition to helpful information can serve as a definition an event. A number of multimodal strategies are found upon in this situation, and feature fusion could be expressed by either initial fusion or lateness fusion. Directly combining characteristics into a greater attribute vector but therefore the simplest method is to learn the underlying action benefit from numerous features. Although the resultant feature vector has a significantly bigger dimension, this feature fusion strategy may improve recognition performance.

A temporal relationship between the underlying activity and the various modalities is crucial in understanding the data since multimodal cues are typically connected in time. In that situation, audiovisual analysis serves a variety of purposes beyond just synchronizing audio and video, however, for monitoring identification of activities. Three groups of multimodal techniques are distinguished: (i) effective techniques, (ii) behavioral techniques, and (iii) social networking-based techniques. Multimodal approaches define atomic interactions or activities that may be related to the effective states of either a communicator's counterpart and depend on feelings and/or physical movements (Fig. 6).

7 Performance of Collected Dataset

Dataset is satisfied with high class variability (intra-class) and high class similarity. The following values are shown in Table 1.

Table 1 Table captions should be placed above the tables

Test data	Precision	Recall
Call	0.389	0.392
Running	1.23	1.45
Stop	1.5	1.4
Hello	0.944	0.833
Pointing	1.0	0.49
Others	0.764	0.765

Table 2 Scaling with tested data and random data

Scale	Test data	New person data
1–10	Rappel	–
10–20	Precision	Precision
20–30	Accuracy	–
30–40	–	–
40–50	–	–

In both Tables 1 and 2, we calculated the precision and recall value of tested data where some data [10] on precision, rappel and accuracy with latest relevant data. We also categorized the age scale between 1 and 10, and last range was 40–50 for monitoring the activity of human. Some results are better in age from 25 to 40, i.e., middle age. We use dataset in further study if we consider any image [11–14] pattern [15–19].

8 Conclusion

Real-time-based model can be predicted with human activity recognition, so in this paper, we conducted a thorough analysis of contemporary techniques for identifying human activity and developed a hierarchical taxonomy of grouping these techniques. According to channel of origin, many of these methods are used to identify activities of humans, and we surveyed many methodologies and divided them into two major categories (unimodal and multimodal). The motion properties of an action sequence were captured through using depth motion maps created from three projection perspectives. In future work, motion monitoring, image classification, and video classification may be useful for exascale computing with fast computing technique.

References

1. Chen C, Liu K, Kehtarnavaz N (2016) Real-time human action recognition based on depth motion maps. J Real-Time Image Proc 12(1):155–163
2. Cheng X et al (2022) Real-time human activity recognition using conditionally parametrized convolutions on mobile and wearable devices. IEEE Sensors J 22(6):5889–5901
3. Park J, Lim W-S, Kim D-W, Lee J (2022) Multi-temporal Sam pling module for real-time human activity recognition. IEEE Access
4. Mazzia V et al (2022) Action transformer: a self-attention model for short-time pose-based human action recognition. Pattern Recog 124:108487
5. Andrade-Ambriz YA, Yair A et al (2022) Human activity recognition using temporal convolutional neural network architecture. Expert Syst Appl 191:116287
6. Sun X et al (2022) Capsganet: deep neural network based on capsule and GRU for human activity recognition. IEEE Systems J
7. Alemayoh TT, Lee JH, Okamoto S (2021) New sensor data structuring for deeper feature extraction in human activity recognition. Sensors 21(8):2814
8. http://crcv.ucf.edu/data/UCF_Sports_Action.php
9. Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. Front Robot AI 2:28
10. Kumar M, Gautam P, Semwal VB (2023) Dimensionality reduction-based discriminatory classification of human activity recognition using machine learning. In: Proceedings of Third International Conference on Computing, Communications, and Cyber-Security. Springer, Singapore, pp 581–593
11. Joshi K, Diwakar M, Joshi NK, Lamba S (2021) A concise review on latest methods of image fusion. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) 14(7):2046–2056
12. Sharma T, Diwakar M, Singh P, Lamba S, Kumar P, Joshi K (2021) Emotion analysis for predicting the emotion labels using Machine Learning approaches. In: 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp. 1–6, November. IEEE
13. Diwakar M, Sharma K, Dhaundiyal R, Bawane S, Joshi K, Singh P (2021) A review on autonomous remote security and mobile surveillance using internet of things. J Phys: Conference Series 1854(1):012034, April. IOP Publishing
14. Tripathi A, Sharma R, Memoria M, Joshi K, Diwakar M, Singh P (2021) A review analysis on face recognition system with user interface system. J Phys: Conference Series 1854(1):012024. IOP Publishing
15. Wang Y et al (2021) m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE
16. Sun B, Wang S, Kong D, Wang L, Yin B (2021) Real-time human action recognition using locally aggregated kinematic-guided skeletonlet and supervised hashing-by-analysis model. IEEE Trans Cybernetics
17. Varshney N et al (2021) Rule-based multi-view human activity recognition system in real time using skeleton data from RGB-D sensor. Soft Comp, 1–17
18. Hossain T, Ahad M, Rahman A, Inoue S (2020) A method for sensor-based activity recognition in missing data scenario. Sensors 20(14):3811
19. AlShorman O, Alshorman B, Masadeh MS (2020) A review of physical human activity recognition chain using sensors. Indonesian J Elect Eng Inform (IJEEI) 8(3):560–573