



Action recognition for the robotics and manufacturing automation using 3-D binary micro-block difference

Viacheslav Voronin¹ · Marina Zhdanova¹ · Evgenii Semenishchev¹ · Aleksander Zelenskii¹ · Yigang Cen² · Sos Agaian³

Received: 15 January 2021 / Accepted: 30 June 2021 / Published online: 9 July 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

Vision-based control systems play an important role in modern robotics systems. An important task in implementing such a system is developing an effective algorithm for recognizing human actions and the working environment and the design of intuitive gesture commands. This paper proposes an action recognition algorithm for robotics and manufacturing automation. The key contributions are (1) fusion of multimodal information obtained by depth sensors and cameras of the visible range, (2) modified Gabor-based and 3-D binary-based descriptor using micro-block difference, (3) efficient skeleton-based descriptor, and (4) recognition algorithm using the combined descriptor. The proposed binary micro-block difference representation of 3-D patches from video with a complex background in several scales and orientations leads to an informative description of the scene action. The experimental results showed the effectiveness of the proposed algorithm on datasets.

Keywords Action recognition · Human activity · Descriptor · Machine vision system · Human-robot collaboration

1 Introduction

The concept of human-robot collaboration (HRC) is one of the leading trends in Industry 4.0 development technology. The HRC model combines a human and a robot in one workspace to perform common tasks. The following advantages characterize human-robot cooperation: production processes become faster, more efficient, and more cost-effective, and the level of automation increases; workload decreases to the operator; product quality is growing; and flexibility and mobility of production processes are provided. Firstly, implementing such an interaction system requires ensuring the operator's safety

who has direct contact with the robot's moving parts, and secondly, designing an effective interaction interface will fully use human skills [1].

Traditionally, robots are programmed to automatically perform various repetitive operations using text editors and direct controllers, such as using learning consoles and touch screen interfaces. But these devices have a complex set of tools to provide a wide range of robot functionality. In addition, the use of such interfaces, in most cases, is not ergonomic for the operator, requires inconvenient hand movements for remote control tasks, which increases the user's workload and also requires additional training.

Such control may be acceptable for short-term intervention in the operation of a robotic system, but for collaborative systems where a human operator continually interacts with a robot, a more efficient and native interface is required [2]. The presence of intuitive ways to interact with robots and their programs is one of the key factors for developing and implementing automated robotic technologies in modern manufacturing processes.

The vision system for robots is widely used in many modern industries for various tasks [3]. To ensure contactless control of the robotic complex, the robot must understand the gesture commands. The model of recognizing actions in the

✉ Viacheslav Voronin
voroninslava@gmail.com; voronin_sl@mail.ru

¹ Moscow State University of Technology "STANKIN",
Moscow, Russian Federation

² Institute of Information Science, Beijing Jiaotong University,
Beijing, China

³ Department of Computer Science, City University of New York,
New York, USA

interaction of a person and a robot is shown in Figure 1. At the first stage, data collection occurs, which is carried out by the elements of the technical vision system. Then, based on the information received, a descriptor is formed that describes a specific time period.

Next, activities are classified according to predefined classes. As a result of the operator action command classification, a control command to the robot is generated in the control center. The collaborative robot performs work or simulates its actions based on the commands given by the human operator and the recognition results. The commands given to the robot can take various forms: from simple and concrete to quite abstract. In this work, a vocabulary of gestures of the whole body and hands is defined; each command calls a specific subroutine to control the robot.

An important task in implementing such a control system is developing a stable algorithm for recognizing human actions and the working environment and the design of intuitive gesture commands [1].

The main disadvantages of existing algorithms for recognizing human actions when implemented in natural conditions are non-uniform background, uncontrolled working environment, irregular lighting, partial occlusion of the observed object, speed of actions, etc. [3]. Therefore, this paper focuses on the problems associated with robotics and manufacturing automation using action recognition.

The primary contributions of our paper include a novel:

- 1) A new algorithm for recognizing human actions for implementing a contactless interface for human-robot interaction, based on constructing a new descriptor, which provides invariance concerning the change of scale and brightness
- 2) modified logarithmic image processing algorithm for fusion depth and color images as a Multichannel input data that minimizes external factors influence on the quality of video content

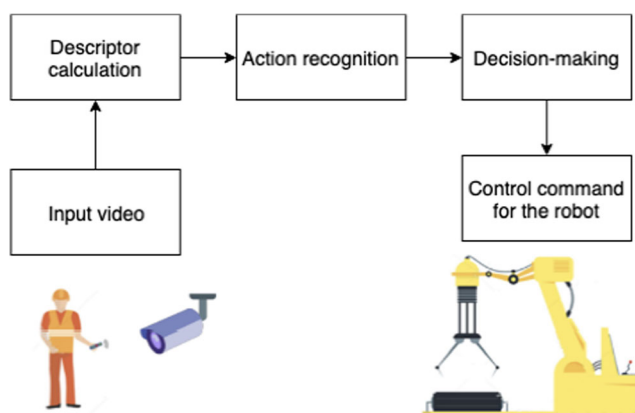


Fig. 1 Action recognition model in the human-robot concept

- 3) Combined 3-D Gabor-based, binary micro-block difference-based, and skeleton-based descriptor

The remainder of this paper is organized as follows. Section 2 presents the human activity recognition method background information. Section 3 defines an algorithm of 3-D binary micro-block difference using a fused RGB image and depth data, extraction of the frequency spectrum, and human skeleton construction. Section 4 presents some experimental results. Finally, Section 5 gives some concluding comments.

2 Related work

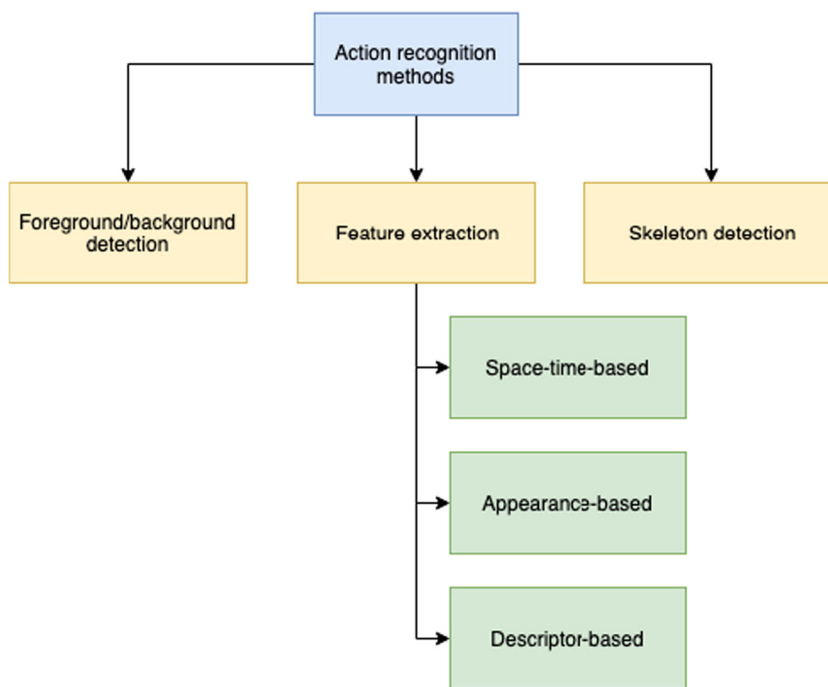
The construction of a descriptor for describing a human posture in individual frames, and the video sequence as a whole, is an integral part of systems for recognizing human actions. Features of shape, movement, and textural properties play an important role in constructing vector features. There are three main directions of recognizing actions for interaction tasks with a robot in the literature (Figure 2). These approaches build a global descriptor for the entire frame (or video) or build local descriptors for key points [4, 5]. Such methods take into account the information about the work area and the background of the frame and analyze objects in the operator's hands [3]. The advantage of such approaches is assessing the scene in general and highlighting any important things taken into account in the classification. But this also is extra information, which can affect the quality of recognition.

The skeleton-based methods are based on the construction of unique joint points of the human body and, on their basis, obtaining the human skeleton [6–8]. Information about the relative position of joints and body parts allows for tracking a person when partially occluded and enables the assessment of the human body's posture in the frame more accurately and, therefore, determines the action performed on the video. Such approaches do not consider the context information of the scene, which is a disadvantage for some tasks.

The action recognition approaches applied to contactless robot control applications. The methods based on constructing a human skeleton work directly with the observation (operator) object, thereby minimizing extra background information.

Such methods make it possible to avoid occlusions and loss of information when foreign objects obstruct the object since the human body's structure is known a priori. But it is the contextual information of the environment, of the objects with which a person interacts, that can play a decisive role in classifying actions. Thus, for each specific task, considering the characteristics of the environment and types of actions, one method from the presented groups may be suitable.

Fig. 2 The classification of the action recognition methods



3 Algorithm of 3-D binary micro-block difference

This paper presents a new action recognition algorithm using fused images of the visible spectrum and depth data and encoding 3-D patches within the video sequence. The flowchart of the proposed recognition algorithm is shown in Figure 3. The proposed technique uses the difference between 3-D micro-block built inside the patches of the video sequence. Thus, it encodes the discriminatory information that the frame contains.

This scheme has some variations and does not enforce to follow the implementation rules strictly. The variety and complexity of recognition tasks and the specificity of the actions performed do not allow implementing a single universal approach to their solution. The presented algorithm will enable users to bypass some implementation stages if, in their opinion, they are not rational, require significant time or computational resources. For example, it is possible to implement the algorithm without using depth data. In this case, only information from the visible spectrum sensor is supplied to the input, and the second step of data combining is skipped.

3.1 Image fusion

Image fusion combines two or more images obtained by different sensors of the same object into a single image that is more suitable for human visual perception and computer-aided analysis and decision-making [9]. The commonly used image fusion techniques can be categorized as pixel level, decision level, and feature level. Or, they can be classified as

spatial and frequency domain [9]. In addition, image fusion is necessary for computer vision and robotics systems in which fusion results can be used to aid further processing steps for a given task [9].

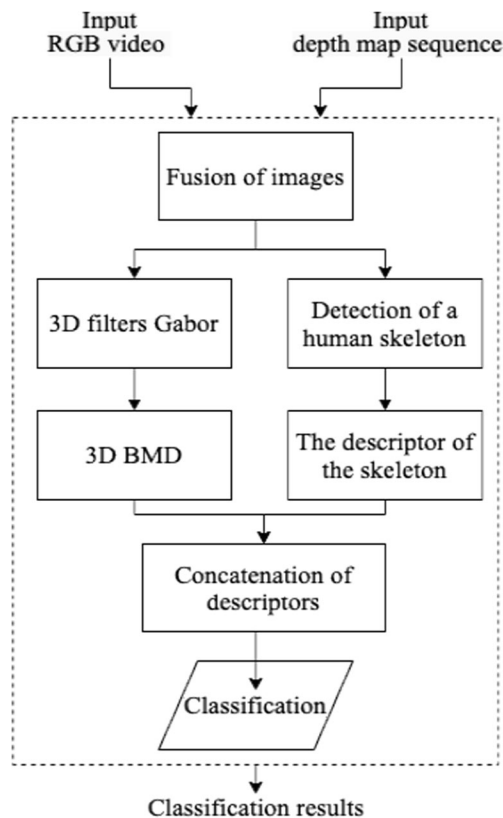


Fig. 3 Flowchart of the human action recognition algorithm

Traditionally, image processing uses linear operations to manipulate images, but since computer arithmetic is inherently a non-linear process, precision issues can arise. For example, when pixel intensities are outside the range (0, M), they are clipped, causing information loss. In addition, linear operations usually do not produce results consistent with physical phenomena.

Logarithmic image processing (LIP) replaces linear arithmetic (addition, subtraction, and multiplication) with non-linear ones, which more accurately characterizes the non-linearity of computer image arithmetic. In work [10], it is presented that the LIP model is consistent with the unified model of vision by Xie and Stockham [11] in the sense that it satisfies Weber law. The intensity of sensation of something is directly proportional to the logarithm of the human visual system stimulus intensity and saturation characteristics.

The logarithmic imaging model has been used successfully for image enhancement, edge detection, and image reconstruction. However, the non-linear arithmetic operations also have several limitations. For example, when two visually “good” images are added together, the output image may not retain its representative properties in terms of overall brightness, becoming unnaturally too dark or too bright [12].

In [9, 13], a parameterized logarithmic image processing (PLIP) is presented, which allows minimizing the above disadvantages used in the new recognition algorithm to merge images of the visible spectrum and images of depth. A math-

ematical analysis shows that (a) the LIP model and standard mathematical operators are extreme cases of the PLIP model operators and (b) the PLIP framework is more suitable for processing images, including image fusion applications. Moreover, the PLIP model operators also can take on cases in between LIP and standard operators based on the visual requirements of the input images [9]. The most important advantage of using depth sensors and visible cameras is the complementary nature of the different modalities that provide information about the depth and visible spectrum. This information obtained in various ways allows increasing the reliability and stability of recognition and provides a more informative descriptor [14, 15].

3.2 Extraction of the frequency spectrum features

One of the algorithm steps is convolution input video with 3-D Gabor filters. This procedure is applied to the frequency spectrum calculated for a video clip, which contains information about both the scene and the movement. It represents the signal as a sum of individual frequency components. The use of 3-D Gabor filters [16] effectively extracts information about the structure of motion and the shape of scene objects.

This convolution frequency spectrum procedure with 3-D Gabor filters is presented in [13]. It is illustrated in Figure 4 and proceeds as Algorithm 1.

Algorithm 1: Convolution of a 3-D Gabor Filter Bank

Input: Video set (size: $M \times N \times T$, where M is the number of rows, N is the number of columns, T is the number of frames).

1. 3-D Discrete Fourier Transform.
 2. Generation of a 3-D filter bank with different orientations and scales.
 3. Convolution of the frequency spectrum with 3-D Gabor filters.
 4. Inverse 3-D Discrete Fourier Transform.
 5. Dimension reduction.
-

Output: Resulting data array (H_1, H_2, \dots, H_i) .

The frequency spectrum extraction using a 3-D Fourier transform is described in more detail in works [13, 17].

At the output of the presented algorithm, an array is formed in a clip (the size coincides with the input data) containing components of a particular structure and direction of movement. The number of output clips corresponds to the number of filters in the 3-D filter bank. For the subsequent processing of the received data, the dimension is reduced by four components (by the number of output video clips).

3.3 3-D binary micro-block difference

The presented method of recognizing human actions is based on the idea that 3-D patches of a video have a characteristic structure. The block diagram of the 3-D algorithm for the binary micro-block difference is illustrated in Figure 5 and proceeds as Algorithm 2.

The video sequence is divided into intersecting sequences of three frames at the first stage, as illustrated in Figure 6.

Further, each video sequence is divided into three-dimensional non-overlapping patches; each patch's size is $16 \times 16 \times 3$. Within each patch, cuboids are built, the number

of which can vary. The coordinates of the central pixels for the construction of cuboids are chosen randomly, but they are fixed for cuboids of different sizes.

Micro-block inside 3-D patches is built according to the volumetric local binary pattern (VLBP) [18]. VLBP is an object type for describing the characteristics of an object in the space-time domain. In this case, the space-time information is considered the sequence of frames as a volume (cuboid) and determines each pixel’s neighborhood in 3-D space. The LBP is a widely used operator for extracting two-dimensional image functions, which has excellent reliability in pattern recognition [19]. In the classical implementation, LBP is defined as a 3×3 window. In this window, the central pixel’s intensity value, taken as the threshold, is compared with the value of the neighboring 8 pixels. Thus, applying the basic LBP operator to the pixel is an 8-bit binary code that describes the neighborhood of this pixel. In the dynamic image, the LBP is defined in the frame sequence’s local region. VLBP is defined in a 3×3×3 voxel. When calculating the VLBP operator, the binary code is constructed by analogy with the LBP. Still, neighboring pixels in the previous and next frames are also compared with the central pixel, as shown in Figure 7.

The procedure for calculating the operator of a large local binary template is shown in Figure 8. Let us consider an example of constructing the operator of volumetric VLBP for cuboids inside a 3-D patch of a video sequence, following Figure 8. The figure schematically shows 3 sequential frames, and pairs of cuboids are built, for which the VLBP operator is calculated, as described above. Light gray and dark gray colors indicate cuboids. The red lines connect the cuboids between the binary codes of the VLBP where the Hamming distance will be calculated (the red lines are marked only in the first figure, with cuboids of radius R=1, to avoid overloading figures with cuboids of radius R=2 and R=3).

Figure 9 shows the size of a 3-D patch 16×16×3; cuboids are 3×3×3, 5×5×3, and 7×7×3. The coordinates of the central pixels (C) are selected at random—marked in white. After choosing the cuboid’s central coordinates, they remain fixed and saved for each video sequence patch. Neighboring pixels (P=4) are located equidistant from the central one with a radius R=1, 2, 3—schematically marked in black color. VLBP is calculated for cuboids of different scales, which provides invariance concerning scale changes. As an example, in Figure 9, only 8 cuboids are presented. Their number can

Algorithm 2: 3-D binary micro-block difference

Input: Video set.

1. Splitting the video sequence into video clips.
2. Each clip is divided into three-dimensional non-overlapping patches (patch size is 16×16×3).
3. Inside each patch, cuboids of different sizes (3×3×3, 5×5×3, 7×7×3) are built.
4. Calculate the Hamming distance between a randomly selected pair of cuboids within each patch of a video clip.
5. The values of Hamming distances between cuboids of different sizes for each patch are sequentially written into a separate vector, characterizing a 3-D patch.
6. The vectors of each 3-D patch of the video sequence are concatenated into a single vector to describe the entire video sequence's texture properties.
7. Steps 2-5 are repeated sequentially for every three frames of the entire video until the last 3 frames are reached.
8. The vectors of each clip are concatenated into a single vector to describe the entire video sequence.
9. Dimension reduction.

Output: The vector (descriptor).

VLBP is calculated as follows [20]. The intensity value of the central pixel $i_{t,c}$ is subtracted from the intensity values of pixels P located in a circular neighborhood of radius R ($R > 0$) in the image t : $i_{t,p}$ ($t = t_c - L, t_c, t_c + L; p = 0, \dots, P - 1$), thus

$$V = v \left(\begin{matrix} i_{t_c-L,c} - i_{t_c,c}, i_{t_c-L,0} - i_{t_c,c}, \dots, i_{t_c-L,p-1} - i_{t_c,c}, i_{t_c,c} \\ c, i_{t_c,0} - i_{t_c,c}, \dots, i_{t_c,p-1} - i_{t_c,c}, i_{t_c+L,0} - i_{t_c,c}, \dots, i_{t_c+L,p-1} - i_{t_c,c}, i_{t_c+L,c} - i_{t_c,c} \end{matrix} \right)$$

This texture operator captures the appearance of various patterns in each pixel vicinity on a $(2(P + 1) + P = 3P + 2)$ -dimensional histogram [21].

vary; 20 pairs of cuboids are built inside each image patch in this work. Then, the Hamming distance is calculated for each cuboid pair (in Figure 8 connected by a red line). The values obtained for each pair of cuboids of various sizes are sequentially written into a single vector, which characterizes a three-dimensional patch in several resolutions.

The Hamming distance is calculated as the number of different characters at the same positions of two words:

$$d_{xy} = \sum_{k=1}^n |x_k - y_k|$$

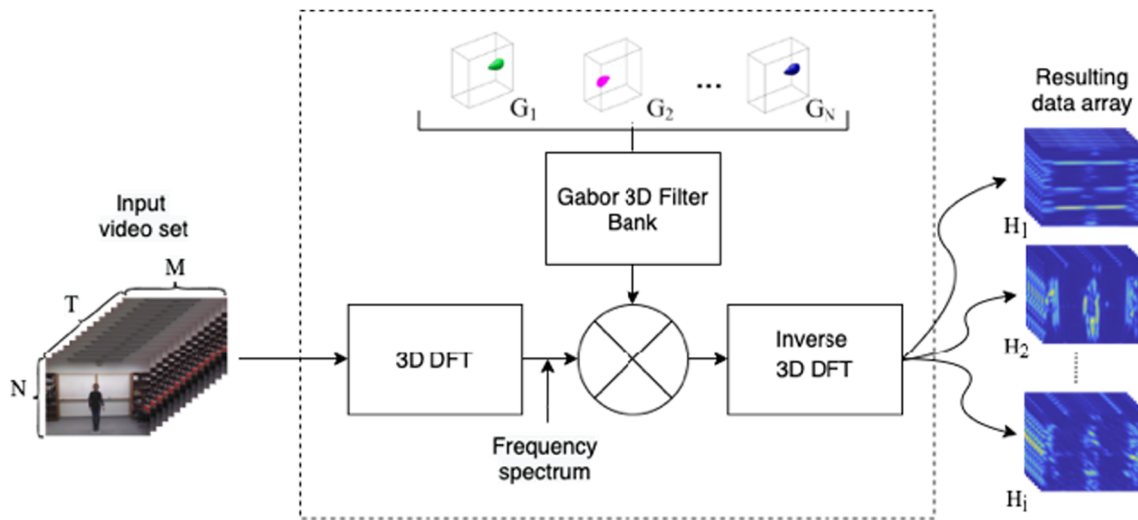


Fig. 4 Diagram of convolution with 3-D Gabor filters

As mentioned earlier, the operator of VLBP is calculated only for three frames; it is supposed to be applied sequentially to the following frames for the entire video, with the construction of a histogram for every three frames. The obtained histograms are combined sequentially into a single resulting feature vector for multi-frame training, as shown in Figure 10.

The proposed functions extract information at three different levels from an image patch: resolution, orientation, and scale. The random selection of sampling points facilitates obtaining information in several orientations. In addition, the sampling points are aligned at different angles to help capture variations in different patch orientations. Scale invariance is also achieved by changing the distance between points.

The local binary template operator used to calculate a cuboid histogram provides several advantages: invariance concerning brightness and relatively low computational costs due to binary calculations.

3.4 Pose estimation based on human skeleton modeling

To increase the action recognition system productivity and efficiency propose to use the human skeleton analysis. The dynamics of the skeletons of the human body carry important information for recognizing human actions. This data will reduce the recognition error and focus the proposed method's attention on smaller actions performed by a person's hands or wrist.

A convolutional uses the recurrent neural network [22], which takes video sequence frames as input. This network learns to create heat maps for each key (joint) point, where the location of the key point is obtained as a heat map mode [23].

The human body skeleton is based on 16 key points, joints, namely the right ankle, right knee, right thigh, left thigh, left knee, left ankle, torso, neck, chin, crown of the head, right wrist, right elbow, right shoulder, left wrist, left elbow, and left shoulder (Figure 11). The action recognition method based on constructing a human skeleton is described in detail [13, 17].

For each frame of the video sequence, a human skeleton is built, and the coordinates of the singular points j_n are allocated, where n is the number of the joints (Figure 11). A set of geometric features is used to recognize human actions, which informatively describes the distance between the human body's joints. For example, when a person sits down, the distance between the key points of the thigh and ankle decreases and vice versa; when a person stands, the distance between the joints of the thigh and ankle increases; the distance between the shoulder joint and the wrist also varies when performing various actions. In this way, geometric features and the body's length carry a significant informative load [18].

When constructing an informative descriptor describing a human action, the coordinates of 16 joints of the human body, detected frame by frame, are used, and geometric features are built on their basis [24]. Thus, the descriptor is formed from the following data:

- X, Y coordinates of all joint points j_n
- Body length from the ankle to the crown
- Distance from the ankle to hip
- Distance from the shoulder to wrist
- The distance between the wrists of the right and left hand
- The center of gravity of the body

The coordinates are normalized relative to the body length and close to the center of gravity to prepare the data. The formulas describing the human body model correspond [18].

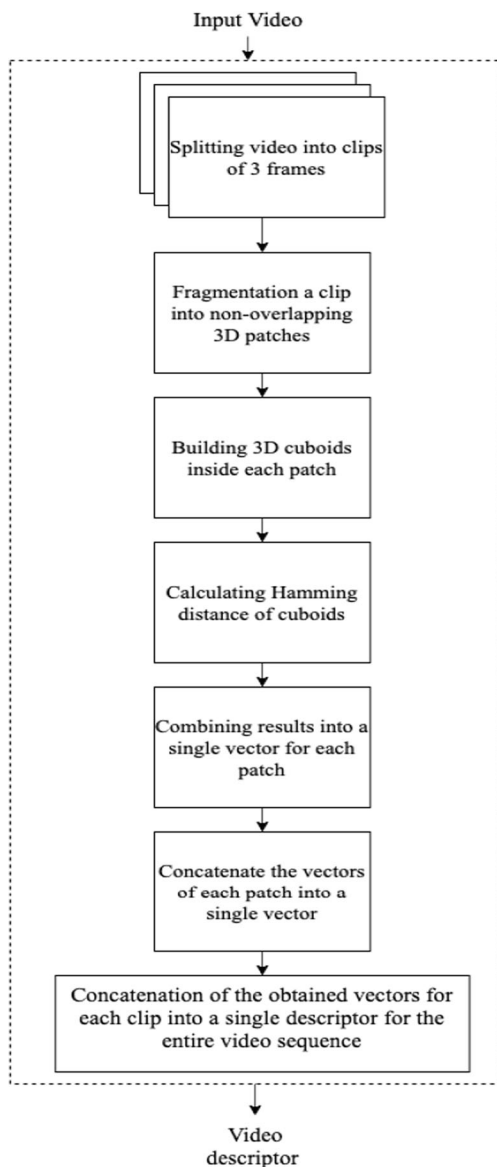


Fig. 5 Diagram of 3-D binary micro-block difference algorithm

The trajectory of movement of the skeleton and joints of the human body is resistant to changes in lighting and scene changes; they are easy to obtain thanks to high-precision depth sensors or posture estimation algorithms.

3.5 Classification

At the final stage, descriptors proceed to the classifier to categorize the video sequence actions. This stage can be characterized by two approaches: combining a skeletal descriptor and 3-D binary micro-block difference into a single vector and classifying it or classifying each of the descriptors separately with the subsequent combining of the results, assigning weights to each of them. This paper proposes to classify

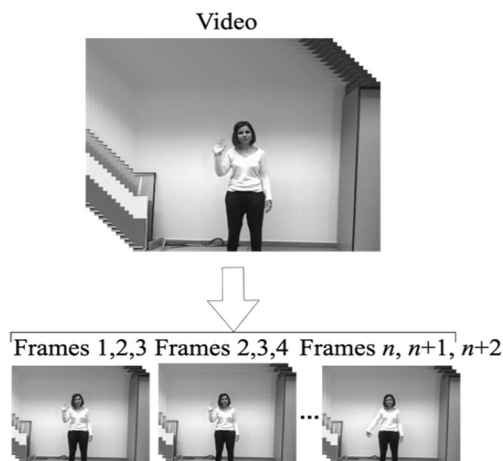


Fig. 6 The principle of dividing a video sequence into short video sequences of three frames

descriptors separately using a multiclass SVM, followed by a decision to categorize the action in the video clip.

4 Experimental results

4.1 Numerical comparison

The proposed method was tested on the UCF101 dataset [25]. It is an action recognition dataset of realistic videos, collected from YouTube, having 101 action categories. This dataset is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. It includes the actions: boxing, clean and jerk, jumping jack, tai chi, body weight squats, jump rope, bench press, push-ups, hula hoop, juggling balls, and yoyo.

To estimate the effectiveness of the proposed method, the classification accuracy is used:

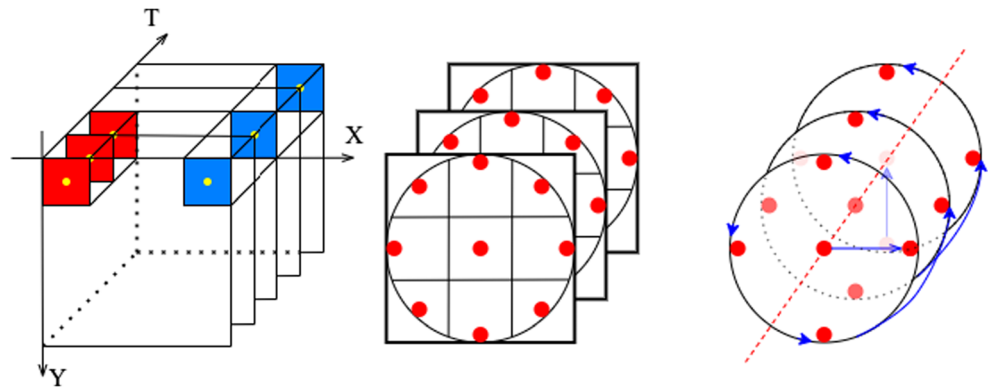
$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \cdot 100\%$$

It is the ratio of the number of correct predictions to the total number of input samples.

The results of calculations the accuracy is shown in Table 1. The effectiveness of the proposed algorithm is compared with methods (iDT+HSV [26], Two-stream+LSTM [27], LTC [21], R-STAN-50 [28], Dynamic Image Networks + IDT [29], P3D [30], MV-CNN [31], ActionFlowNet) [32].

As shown in Table 1, the proposed method significantly outperforms “ActionFlowNet,” “MV-CNN,” and “P3D.” Furthermore, the analysis of the obtained results indicates that the efficiency of the developed method is relatively high in UCF101 (93.2%), even with the rich background context appearance in the dataset.

Fig. 7 The principle of constructing a cuboid inside a space-time patch of a video sequence



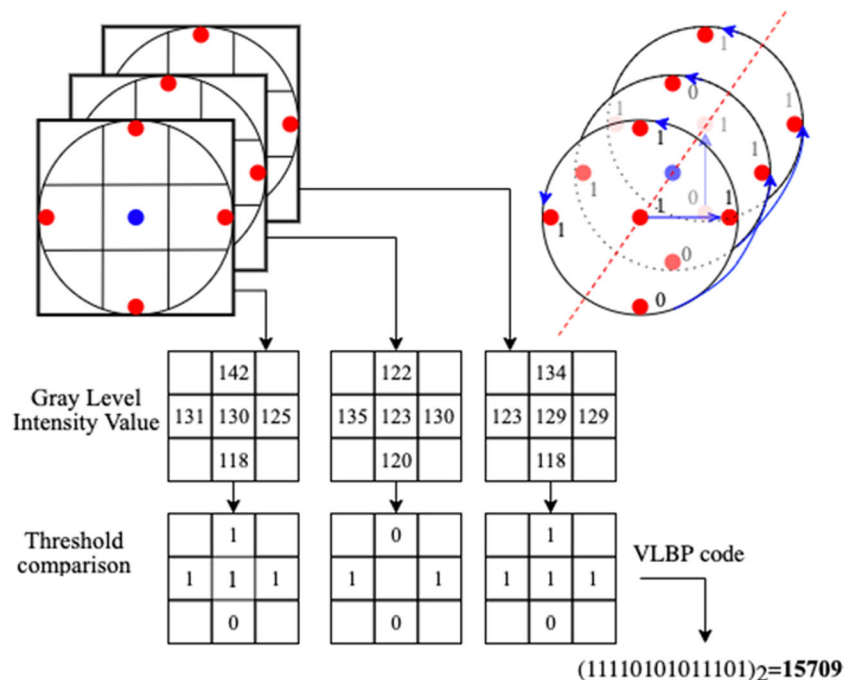
The proposed method increases the recognition efficiency by combining information about the background and the human skeleton, which avoids uncertainty in decision-making when tracking key points and contextual information about the background. The developed 3-D binary descriptor of the micro-block difference for describing human actions, providing invariance of recognition features for large-scale and brightness transformations of halftone images, and binary calculations significantly reduce computational costs. The proposed algorithm for recognizing human actions on complexly structured images combines a 3-D binary descriptor and the geometric features of the human skeleton, which makes it possible to increase the efficiency of the presented approach.

The algorithm input data is characterized by multichannel, namely, the stream of frames of the video sequence of the visible spectrum. The stream of frames received from the depth sensor is fed to the input of the algorithm. This

multichannel allows minimizing the influence of external factors on the quality of video content: poor lighting, loss of information during data transmission, noise, etc. Combining data of both modalities ensures the complementary nature of the final video stream, which may contain information that is not available when working with separate sources. The complementarity of information makes it possible to increase the reliability and stability of recognition and provides a more informative descriptor.

It is impractical to use each stage of the presented algorithm in some cases since it is necessary to find a balance between computing resources, time costs, and classification quality. This feature provides the universality of the algorithm since the diversity and complexity of recognition tasks and the specificity of the features of the actions performed do not allow the implementation of one universal approach to their solution.

Fig. 8 An example of the procedure for calculating the operator of a large local binary template with parameters $L=1$, $P=4$, $R=1$.



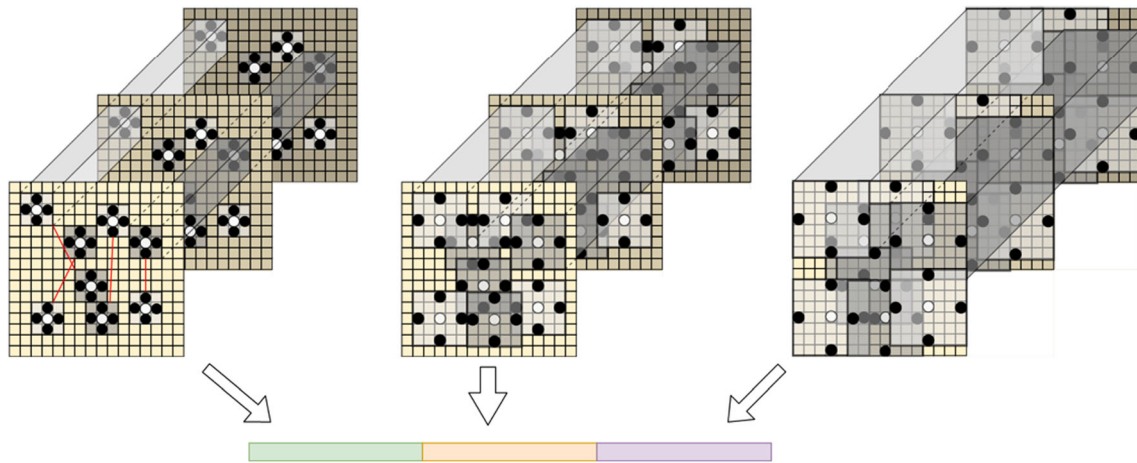


Fig. 9 An example of constructing cuboids in a 3-D patch of size $16 \times 16 \times 3$. Three different resolutions of a cuboid with radius $R=1, 2, 3$ and the number of neighboring pixels $P=4$ are presented. The Hamming

distance values between a pair of cuboids of different scales are combined to obtain the final feature vector

4.2 Action recognition for the robotics and manufacturing automation application

This section considered the application of action recognition for robotics and manufacturing automation. The developed control system for the robotic complex is an interface for human-robot interaction, which receives commands from the operator in the form of gestures. The system uses the elements of technical vision to contactless receive input data from the user. The block scheme of the robotic complex control system is shown in Fig. 12.

Control system for a robotic complex based on an industrial robot with 6 degrees of freedom, a data processing system, a TCP/IP server, and a robot controller. The data processing system consists of software as the central control element of the user interface, which receives input signals from the visual information sensors and generates input data for the decoder

program in the controller. The controller receives a signal from the TCP server data processing unit software and starts the robot to perform actions.

As command gestures, for the developed method of recognizing human actions for a collaborative robotic, the following are used: swipe the hand to the right, swipe the hand to the left, and swipe the hand from the right to the left and back (Hello). Table 2 shows examples of the described gesture commands generated from the UTD Multimodal Human Action Dataset (UTD-MHAD) [33]. The UTD-MHAD dataset consists of 27 different actions. However, this study uses only 3 actions that correspond to ergonomic indicators and correspond to the intuitive nature for controlling the robot.

The resulting confusion matrix is shown in Figure 13. Interestingly, for proposed command actions, the worst recognition occurs for the “hello.” The action “swipe to the left” and “swipe to the right” are similar and shows the high accuracy.

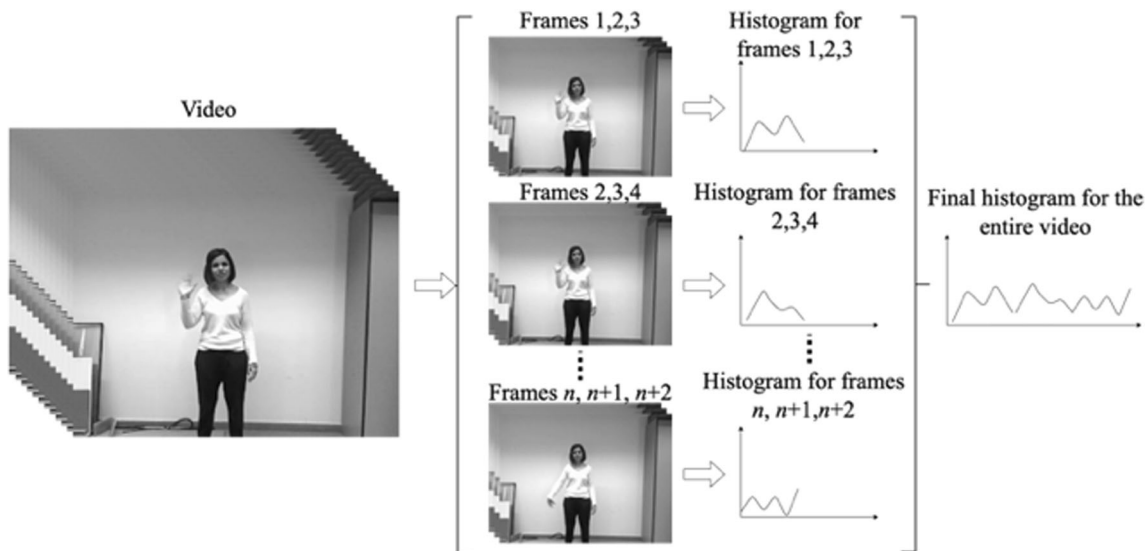


Fig. 10 The principle of constructing a single feature vector for the video sequence



Fig. 11 An example of building a skeleton

The gesture recognition program, in manipulation mode, analyzes the position of the user’s three-dimensional joints with a frequency of 30 Hz. These include the neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, and left wrist. The body joints used in the recognition process in the positioning mode are shown in Figure 11. Algorithms constantly analyze the x (horizontal), y (vertical), and z (depth) values of each joint to find conditions for matching any of the gestures. If all the conditions are met, then the action will be recognized, and the control signal will be sent to the robot.

The process of perception of gestures by the robot should be continuous throughout the entire period of control. The user should be able to automatically change the path of movement of the robot at the required moment. The robot movements must have an

Table 1 The classification accuracy of the action recognition on the test videos

Methods	Accuracy
iDT+HSV [25]	87.9 %
Two-stream+LSTM [26]	88.6 %
LTC [26]	91.7 %
R-STAN-50 [27]	91.5 %
Dynamic Image Networks + IDT [28]	89.15 %
P3D [29]	88.65 %
MV-CNN [30]	86.45 %
ActionFlowNet [31]	83.95 %
Proposed	93.2 %

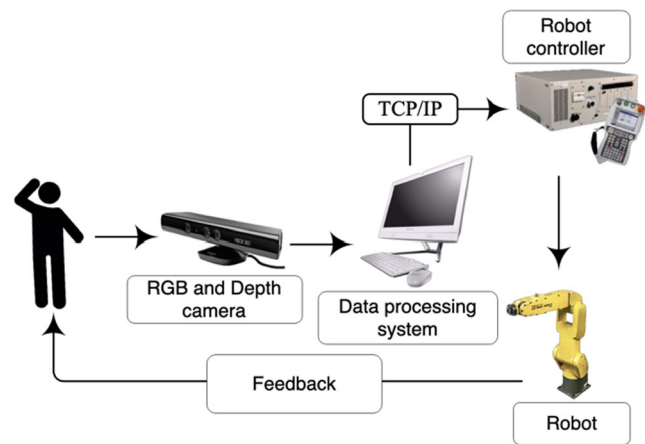


Fig. 12 The control system for robotic system

instant response, which will ensure the user is in contact with the robot and that the robot’s actions match the input gestures. Otherwise, any delay in feedback will prevent the operator from syncing with the robot, and contact will be lost. Thus, the use of dynamic gesture recognition is prohibited because it takes half a second to complete the dynamic gesture procedure, so the system’s response is not instantaneous. Therefore, it cannot

Table 2 The actions for robot control

The actions	The command for the robot	Image
swipe to the left	command mode	
swipe to the right	positioning mode	
enable remote control mode	enable remote control mode	
hello	hello	

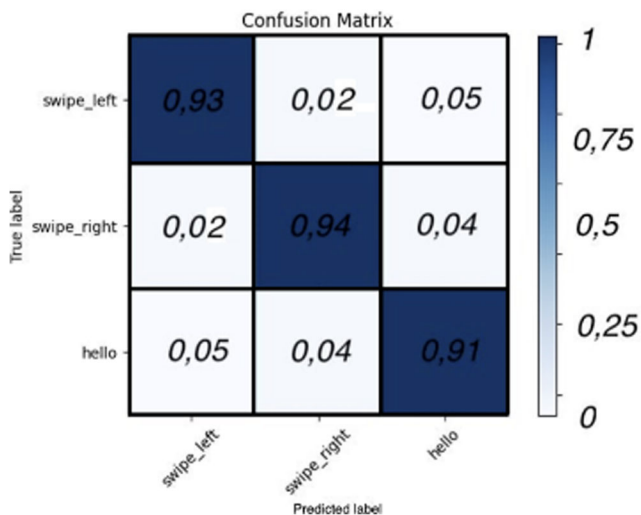


Fig. 13 Confusion matrix of action recognition for robot control

provide continuous input to control the movement of the robot. The activation method used in the developed system is based on positional threshold values of the hands, so the robot reacts as soon as the user's hands reach the threshold values.

5 Conclusions

The proposed human actions recognition algorithm on complexly structured images based on a 3-D binary micro-block difference descriptor for fused multimodal information is obtained by depth sensors and cameras of the visible range. Also, we use the analysis of the human skeleton. Such representation of 3-D blocks of a video sequence by capturing sub-volumes, inside each patch, in several scales and orientations leads to an informative description of the scene action. Considered the application of action recognition for robotics and manufacturing automation as an interface for human-robot interaction, which receives commands from the operator in the form of gestures. Experimental results showed the effectiveness of the proposed algorithm on known data sets.

The limitation of the proposed approach is that its effectiveness is reduced on video with nonstationary camera motions and moving partly occluded objects, which can be a point of interest for further research.

Author contribution All authors contributed to the process of critical literature review. In addition, all authors contributed to writing and revising the manuscripts.

Funding The reported study was funded by Educational Organizations in 2020–2022 Project under Grant No FSFS-2020-0031 and in part by RFBR and NSFC according to the research project No 20-57- 53012.

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Ethics approval The manuscript in part or in full has not been submitted or published anywhere. The manuscript will not be submitted elsewhere until the editorial process is completed.

Consent to participate Not applicable.

Consent for publication The author transfers to Springer the non-exclusive publication rights.

Conflict of interest The authors declare no competing interests.

References

- Ogenyi U, Liu J, Yang C, Ju Z, Liu H (2021) Physical human–robot collaboration: robotic systems, learning methods, collaborative strategies, sensors, and actuators. *IEEE Transactions on Cybernetics* 51(4):1888–1901
- Heo Y, Kim D, Lee W, Kim H, Park J, Chung W (2019) Collision detection for industrial collaborative robots: a deep learning approach. *IEEE Robotics and Automation Letters* 4(2):740–746
- Nascimento H, Mujica M, Benoussaad M (2021) Collision avoidance interaction between human and a hidden robot based on kinect and robot data fusion. *IEEE Robotics and Automation Letters* 6(1): 88–94
- Solmaz B, Assari SM, Shah M (2013) Classifying web videos using a global video descriptor. *Mach Vis Appl* 24(7):1473–1485
- Ji XF, Wu QQ, Ju ZJ, Wang YY (2017) Study of human action recognition based on improved spatio-temporal features. Springer, Berlin Heidelberg, pp 233–250
- Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+ d: A large scale dataset for 3-D human activity analysis. *Proc CVPR*:1010–1019
- Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3-D action recognition. *Proc CVPR*:3288–3297
- Kim TS, Reiter A (2017) Interpretable 3-D human action analysis with temporal convolutional networks. *Proc. CVPR*, 1623–1631
- Brailean JC, Little D, Giger ML, Chen CT, Sullivan BJ et al (1992) Application of the EM algorithm to radiographic images. *Med Phys* 19(5):1175–1182
- Nercessian S, Panetta K, Agaian S (2011) Multiresolution decomposition schemes using the parameterized logarithmic image processing model with application to image fusion, computer science, *EURASIP J. Adv. Signal Process*
- Xie Z, Stockham TG (1989) Toward the unification of three visual laws and two visual models in brightness perception. *IEEE Trans Syst Man Cyber* 19:379–387
- Panetta K, Wharton E, Agaian S (2007) Parameterization of logarithmic image processing models. *IEEE Tran. Systems, Man, and Cybernetics, Part A: Systems and Humans*
- Zhdanova M, Voronin V, Semenishchev E, Ilyukhin Y, Zelensky A (2020) Human activity recognition for efficient human-robot collaboration. *Proc. International Society for Optics and Photonics*, 115430K
- Serrano-Cuerda J, Fernández-Caballero A, López M (2014) Selection of a visible-light vs. thermal infrared sensor in dynamic environments based on confidence measures. *Appl Sci* 4(3):331–350

15. Voronin V, Zhdanova M, Semenishchev E, Zelensky A, Tokareva O (2020) Fusion of color and depth information for human actions recognition. Proc. International Society for Optics and Photonics, 114231C
16. Berkan Solmaz, Shayan Modiri Assari, Mubarak Shah (2012) Classifying web videos using a global video descriptor
17. Zelensky A, Zhdanova M, Voronin V, Alepko A, Gapon N, Egiazarian KO, Balabaeva O (2019) Control system of collaborative robotic based on the methods of contactless recognition of human actions. EPJ Web of Conferences 224:04006
18. Baumann F, Ehlers A, Rosenhahn B, Liao J (2016) Recognizing human actions using novel space-time volume binary patterns. Neurocomputing. 173:54–63
19. Ojala T, Pietikainen M, Harwood D (1994) Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. Pattern Recognition, vol. 1
20. Zhao G, Pietikäinen M (2006) Dynamic texture recognition using volume local binary patterns. Springer, 165-177
21. Varol G, Laptev I, Schmid C (2018) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1510–1517
22. Belagiannis V, Zisserman A (2017) Recurrent human pose estimation. 12th IEEE International Conference on Automatic Face & Gesture Recognition. 468-475
23. Tompson J, Jain A, LeCun Y, Bregler C (2014) Joint training of a convolutional network and a graphical model for human pose estimation. In NIPS, 1799–1807
24. Johnson S, Everingham M (2010) Clustered pose and nonlinear appearance models for human pose estimation. In BMVC 2(4):5
25. Soomro K, Zamir AR and Shah M (2012) UCF101: A dataset of 101 human action classes from videos in the wild. In CRCV-TR-12-01
26. Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput Vis Image Underst 150:109–125
27. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In CVRP
28. Liu Q, Che X, Bie M (2019) R-STAN: residual spatial-temporal attention network for action recognition. IEEE Access 7:82246–82255
29. Bilen H, Fernando B, Gavves E, Vedaldi A, Gould S (2016) Dynamic image networks for action recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3034-3042
30. Zhaofan Q, Ting Y, Tao M (2017) Learning spatio-temporal representation with pseudo-3D residual networks. ICCV
31. Zhang B, Wang L, Wang Z, Qiao Y, Wang H (2016) Real-time action recognition with enhanced motion vector CNNs. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2718-2726
32. Ng J, Choi J, Neumann J, Davis L (2018) ActionFlowNet: learning motion representation for action recognition. IEEE Winter Conference on Applications of Computer Vision (WACV), 1616-1624
33. Chen C, Jafari R, Kehtarnavaz N (2015) UTD-MHAD: a multi-modal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. Proceedings of IEEE International Conference on Image Processing

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.