# Action Recognition for the Robotics and Manufacturing Automation Using 3-D Binary Micro-block Difference

**Viacheslav Voronin** ( voroninslava@gmail.com )
Moskovskij gosudarstvennyj tehnologiceskij universitet STANKIN https://orcid.org/0000-0001-8114-6383
**Marina Zhdanova**
Moscow State University of Technology
**Evgenii Semenishchev**
Moscow State University of Technology
**Aleksander Zelenskii**
Moscow State University of Technology
**Yigang Cen**
Institute of Information Science, Beijing Jiaotong University
**Sos Agaian**
City University of New York

**Research Article**

**Keywords:** action recognition, human activity, descriptor, machine vision systems, human-robot collaboration

**Posted Date:** June 4th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-189925/v1



**Version of Record:** A version of this preprint was published at The International Journal of Advanced Manufacturing Technology on July 9th, 2021. See the published version at https://doi.org/10.1007/s00170-021-07613-2.

**Noname manuscript No.**
(will be inserted by the editor)

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

**Viacheslav Voronin[a], Marina Zhdanova[a], Evgenii Semenishchev[a], Aleksander Zelenskii[a], Yigang Cen[b], Sos Agaian[c]**
**[a]Moscow State University of Technology «STANKIN», Moscow, Russian Federation**
**[b]Institute of Information Science, Beijing Jiaotong University, Beijing, China**
**[c]Department of Computer Science, City University of New York, New York, USA**

Received: date / Accepted: date

**Abstract** The main stage in the development of an algorithm for recognizing human actions is the construction of an informative and distinctive descriptor. As part of the development of a robot control system based on the recognition of human actions, this stage can be decisive. The use of technical vision elements in real conditions introduces a number of difficulties: an inhomogeneous background, uncontrolled working environment, irregular lighting, partial occlusion of the observed object, speed of actions, etc. In this paper, we propose an algorithm for recognizing human actions on complexly structured images based on a 3-D binary descriptor of micro-block difference. The developed algorithm is based on the fusion of multimodal information obtained by depth sensors and cameras of the visible range. The complementarity of information obtained in various ways 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 inaccessible when working with separate sources. In addition to the main descriptor, the paper proposes to use the analysis of the human skeleton. These data will reduce the recognition error and will focus the attention of the proposed method on smaller actions performed by a person's hands or wrist. The experimental results showed the effectiveness of the proposed algorithm on known data sets.

**Keywords** action recognition, human activity, descriptor, machine vision systems, human-robot collaboration

Viacheslav Voronin
Tel: +79885343459
Fax: +79885343459
E-mail: voroninslava@gmail.com

voronin_sl@mail.ru

## 1 Introduction

The concept of human-robot collaboration (HRC) is one of the leading trends in technology of Industry 4.0 development. The HRC model combines a human and a robot in one workspace to perform common tasks. Human-robot cooperation is characterized by the following advantages: production processes become faster, more efficient, and more cost-effective, the level of automation increases; workload decreases to the operator; product quality is growing; flexibility and mobility of production processes are provided. The implementation of such an interaction system requires, firstly, ensuring the safety of the operator who has direct contact with the moving parts of the robot, and secondly, designing an effective interaction interface that will fully use human skills.
Traditionally, robots are programmed to automatically perform various repetitive operations using text editors, direct controllers, for example, using learning consoles, touch screen interfaces. But these devices have a complex set of

tools to provide a wide range of robot functionality. 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 constantly interacts with a robot, a more efficient and native interface is required. The presence of intuitive ways to interact with robots and their programs is one of the key factors for the development and implementation of automated robotic technologies in modern production processes.
Vision-based remote-control systems satisfy the stated requirements. An important task in the implementation of such a control system is the development of a stable algorithm for recognizing human actions and the working environment, as well as the design of intuitive gesture commands. The main disadvantages of existing algorithms for recognizing human actions when implemented in real conditions are non-uniform background, uncontrolled working environment, irregular lighting, partial occlusion of the observed object, speed of actions, etc. In this paper a new algorithm for recognizing human actions is present for the implementation of a contactless interface of human-robot interaction, based on the construction of a descriptor of 3-D binary micro-block difference, which provides invariance with respect to change of scale and brightness. A multichannel input data stream minimizes the influence of external factors on the quality of video content: poor lighting, loss of information during data transmission, noise, etc.
The paper is organized in the following manner: Section II presents the human activity recognition method background information. Section III defines an algorithm of 3-D binary micro-blocks difference using a fused RGB image and depth data, extraction of the frequency spectrum, and human skeleton construction. Section IV presents some experimental results. Finally, Section V 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 important part of systems for recognizing human actions. Features of shape, movement and textural properties play an important role in the construction of a vector-features. In the literature, there are two main directions of recognition of actions for tasks of interaction with a robot. Some of these approaches are based on building a global descriptor for the entire frame (or video) or building local descriptors for keypoints [1-4]. Such methods take into account information about the work area, the background of the frame, and can analyze objects in the hands of the operator. The advantage of such approaches is the ability to assess the scene in general and to highlight in it any significant objects that will be taken into account in the classification. But this also is superfluous information, which can affect the quality of recognition.
The second group includes methods based on the construction of special joint points of the human body and, on their basis, the construction of the human skeleton [5-10]. Information about the relative position of joints and body parts is a priori known. This information allows you to track a person when partially occlusion, and also allows you to more accurately assess the posture of the human body in the frame, and, therefore, determine the action performed on the video. But such approaches do not take into account the context information of the scene, which is a disadvantage for some tasks.
Both groups of approaches to constructing descriptors for action recognition tasks can be successfully applied to applications of contactless robot control. The methods based on the construction of a human skeleton, work directly with the object of observation (operator), thereby minimizing superfluous background information. Such methods make it possible to avoid occlusions, loss of information when the object is obstructed by foreign objects, since the structure of the human body is known a priori. But it is the contextual information of the environment, of the objects with which a person interacts, that can make a decisive role in the classification of actions, which is implemented by the methods of another group. Thus, for each specific task, taking into account the characteristics of the environment and types of actions, one method from the presented groups may be suitable.

## 3 Algorithm of 3-D binary micro-blocks difference

This paper presents a new recognition actions algorithm on a sequence of frames, based on fusion images of the visible spectrum and depth data and encoding three-dimensional patches within the video sequence. The flowchart of the proposed recognition algorithm is shown in Figure 1. The proposed technique uses the difference between three-dimensional micro-block built inside the patches of the video sequence. Thus, it encodes the discriminatory information that the frame contains.

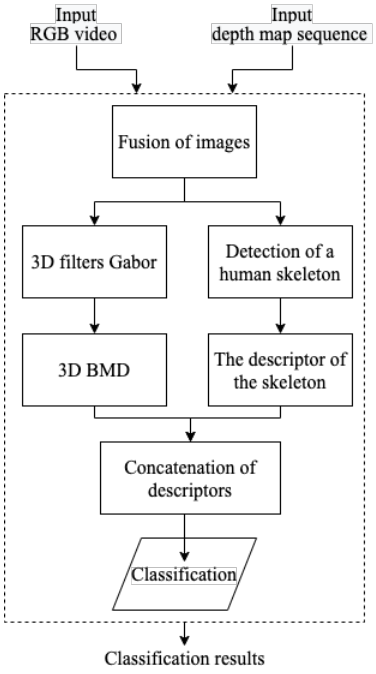


**Fig. 1** Flowchart of the recognition human actions algorithm

The algorithm includes the following main steps: 1) fusion input images of the visible spectrum and depth data; 2a) convolution of fused data with 3D space-time Gabor filters and construction of a descriptor by the 3D binary micro-block difference algorithm; 2b) extraction of the human skeleton and construction of a descriptor based on the geometric features of the location of the articular (special) points of the skeleton; 3) descriptors concatenation obtained at step 2a and 2b into a single feature vector; 4) classification. This structure has some variations and does not enforce to strictly follow the implementation rules. The variety and complexity of recognition tasks, the specificity of the features of the actions performed, does not allow the implementation of a single universal approach to their solution. The presented algorithm allows users to bypass some implementation stages if, in their opinion, they are not rational, require significant time or computational resources. So, for example, it's 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 stage of data combining is skipped.

### 3.1 Fusing images

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. Linear operations usually do not produce results consistent with physical phenomena.
Logarithmic image processing (LIP) replaces linear arithmetic (addition, subtraction, and multiplication) with nonlinear ones, which more accurately characterizes the nonlinearity of computer image arithmetic. In works [11-13] it is presented that the LIP model is consistent with the unified model of vision by Xie and Stockham [14] in the sense that it satisfies Weber's law, namely: the intensity of sensation of something is directly proportional to the logarithm of the stimulus intensity [15], and saturation characteristics of the human visual system.
The logarithmic imaging model has been used successfully for image enhancement, edge detection, and image reconstruction. But 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 [16].
In [17], a parameterized model of logarithmic image processing (PLIP) is presented, which allows minimizing the above disadvantages, and which is used in the new recognition algorithm to merge images of the visible spectrum and images of depth.
The most important advantage of using depth sensors and visible cameras together is the complementary nature of the different modalities that provide information about the depth and visible spectrum of a scene. The complementarity

of information obtained in various ways allows increasing the reliability and stability of recognition [18], provides the construction of a more informative descriptor [19, 20].

### 3.2 Extraction of the frequency spectrum features

One of the stages of the proposed recognition human actions algorithm on a video sequence is convolution with three-dimensional space-time 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 since it represents the signal as a sum of individual frequency components. And the use of three-dimensional Gabor filters [21], differing in orientation and center frequencies, effectively extracts the information about the structure of motion and the shape of scene objects.
This procedure of convolution frequency spectrum with three-dimensional space-time Gabor filters is presented in [17], illustrated in Figure 2 and proceeds as Algorithm 1.

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

**Input:** Video set (size: M×N×T).
1. 3D Discrete Fourier Transform.
2. Generation of a 3D filter bank with different orientations and scales.
3. Convolution of the frequency spectrum with three-dimensional space-time Gabor filters.
4. Inverse 3D Discrete Fourier Transform.
5. Dimension reduction.

**Output:** Resulting data array ($H_1$, $H_2$,…,$H_i$)

Description and formulas for extraction the frequency spectrum using a three-dimensional Fourier transform, generating and convolution bank of 3-D Gabor filters, the inverse three-dimensional Fourier transform is described in more detail in works [17, 20, 22].

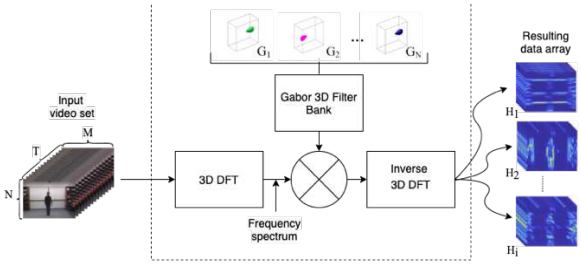

Fig. 2 Diagram of convolution with three-dimensional space-time Gabor filters [17]

At the output of the presented algorithm, an array is formed in the form of a clip (the size coincides with the input data) containing components of a certain structure and direction of movement. The number of output clips corresponds to the number of filters in the 3D Filter Bank. For the subsequent processing of the received data, the dimension is reduced by 4 components (by the number of output video clips).

### 3.3 3D binary micro-block difference

The presented method of recognizing human actions is based on the idea that three-dimensional patches (areas) of a video have a characteristic structure and, if they are captured effectively, discriminative information about the action taking place in the video can be obtained. The block-diagram of the three-dimensional algorithm for the binary micro-block difference is illustrated in Figure 3 and proceeds as Algorithm 2.

**Algorithm 2:** 3D binary micro-block difference

**Input:** Video set.
1. Splitting the video sequence into video clips of three frames.
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 three-dimensional patch.
6. The vectors of each three-dimensional patch of the video sequence are concatenated into a single vector to describe the texture properties of the

entire video sequence.
7. Step 2-5 is 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:** Descriptor described of the entire video sequence.

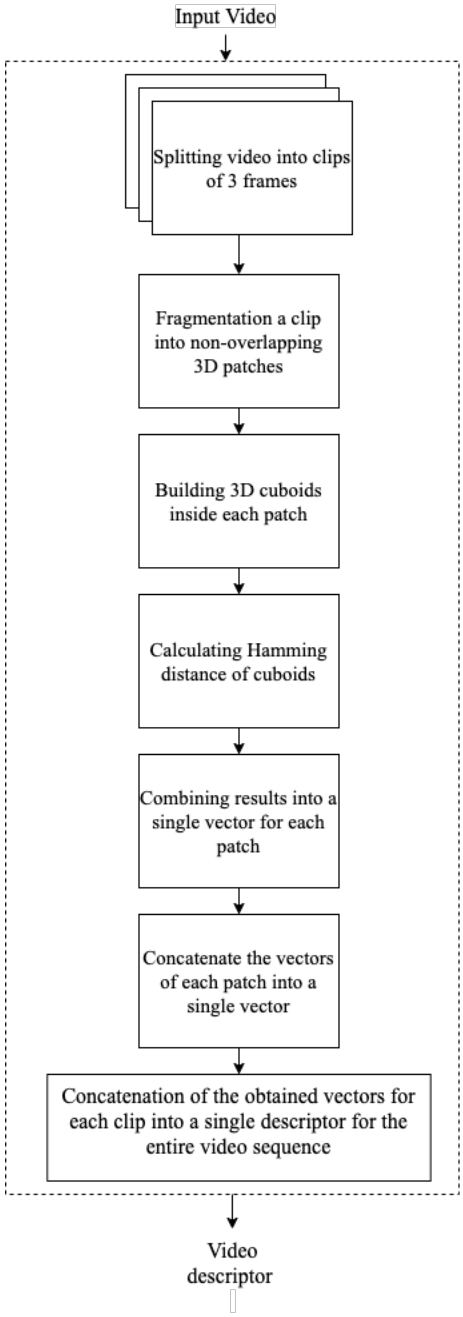


**Fig. 3** Diagram of 3D binary micro-block difference algorithm

At the first stage, the video sequence is divided into intersecting sequences of three frames, as illustrated in Figure 4.
Further, each video sequence is divided into three-dimensional non-overlapping patches, the size of each patch is 16×16×3. 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 at random, but they are fixed for cuboids of different sizes.

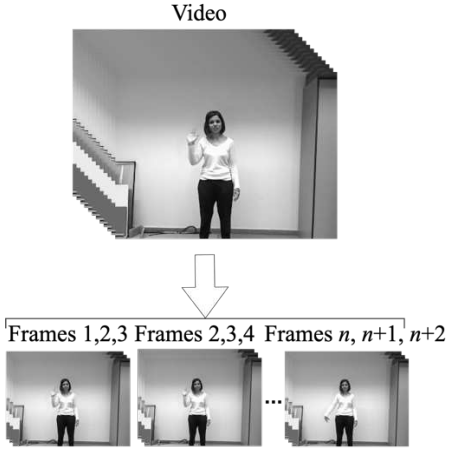


**Fig. 4** The principle of dividing a video sequence into short video sequences of three frames

Micro-block inside three-dimensional patches is built according to the principle of volumetric local binary pattern (VLBP) [22]. VLBP are an object type for describing the characteristics of an object in the space-time domain. Space-time information, in this case, is presented in such a way as to consider the sequence of frames as a volume (cuboid) and determine the neighborhood of each pixel in three-dimensional space.
LBP (Local Binary Pattern) is a widely used operator for extracting functions of two-dimensional images, which has excellent reliability in pattern recognition [24]. In the classical implementation, LBP is defined as a 3×3 window. In this window, the intensity value of the central pixel, taken as the threshold, is compared with the value of the neighboring 8 pixels. If the pixel value exceeds the threshold value, the position of this pixel is denoted 1, otherwise - 0. Thus, the result of applying the basic LBP operator to the pixel is an 8-bit binary code that describes the neighborhood of this pixel [25].
To extend the LBP to analyze the dynamic texture $V$, the dynamic texture is defined in the local region of the frame sequence. VLBP is defined in a 3×3×3 voxel. When calculating the VLBP

operator, the binary code is constructed by analogy with the LBP, but neighboring pixels located in the previous and next frames are also compared with the central pixel, as shown in Figure 5.

VLBP is calculated as follows [26]. The intensity value of the central pixel $i_{t_c,c}$ 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

$$\boldsymbol{V} = \boldsymbol{v}(\boldsymbol{i_{t_c-L,c} - i_{t_c,c}, i_{t_c-L,0} - i_{t_c,c}, \dots, i_{t_c-L,p-1} - i_{t_c,c}, i_{t_c,c}, i_{t_c,0} - i_{t_c,c}, \dots, i_{t_c,p-1} - i_{t_c,c}, i_{t_c+L,0} - i_{t_c,c}, \dots, i_{t_c+L,p-1} - i_{t_c,c}, i_{t_c+L,c} - i_{t_c,c}}) \quad (18)$$

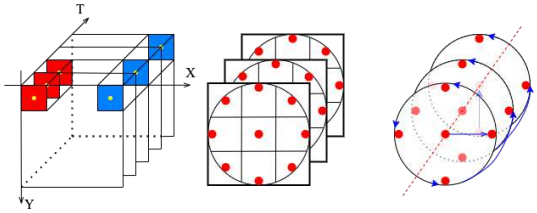


**Fig. 5** The principle of constructing a cuboid inside a space-time patch of a video sequence [26]

This texture operator captures the appearance of various patterns in the vicinity of each pixel on a $(2(P+1)+P = 3P+2)$-dimensional histogram [26].

Ehe procedure for calculating the operator of a large local binary template is shown in Figure 6.

Let us consider an example of constructing the operator of volumetric VLBP for cuboids inside a three-dimensional patch of a video sequence, in accordance with Figure 7. The figure schematically shows 3 sequential frames and pairs of cuboids are built, for which the VLBP operator is calculated, as described above. Cuboids are indicated by light gray and dark gray colors. The red lines connect the cuboids, between the binary codes of the VLBP, of which the Hamming distance will be calculated (the red lines are marked only in the first figure, with cuboids of radius R=1, in order to avoid overloading figures with cuboids of radius R=2 and R=3.

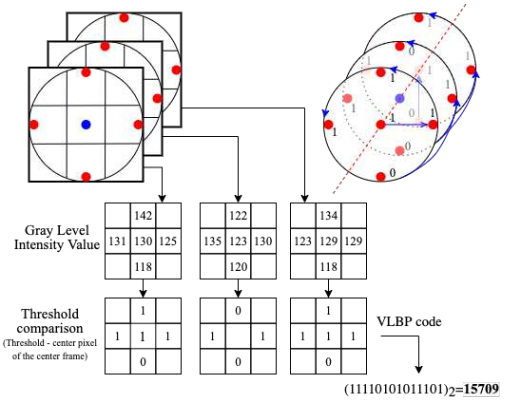


**Fig. 6** An example of the procedure for calculating the operator of a large local binary template with parameters L=1, P=4, R=1.

Let us consider an example of constructing the operator of volumetric VLBP for cuboids inside a three-dimensional patch of a video sequence, in accordance with Figure 7. The figure schematically shows 3 sequential frames and pairs of cuboids are built, for which the VLBP operator is calculated, as described above. Cuboids are indicated by light gray and dark gray colors. The red lines connect the cuboids, between the binary codes of the VLBP, of which the Hamming distance will be calculated (the red lines are marked only in the first figure, with cuboids of radius R=1, in order to avoid overloading figures with cuboids of radius R=2 and R=3).

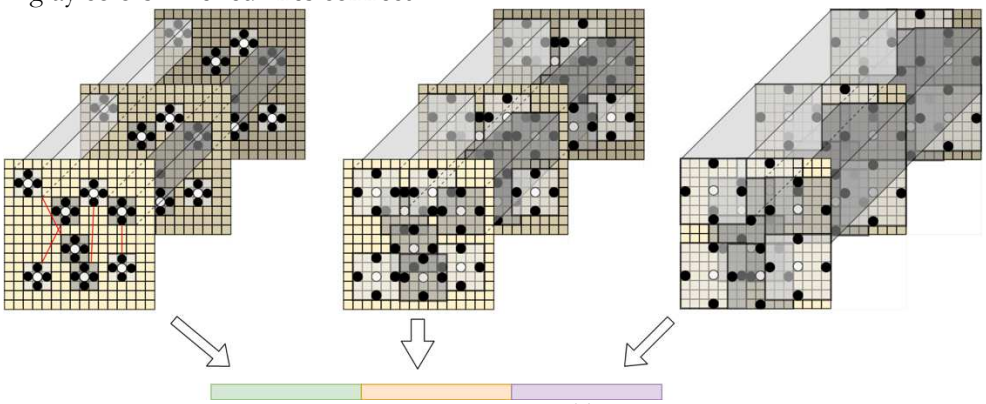

**Fig. 7** An example of constructing cuboids in a 3D patch of size 16×16×3. Three different resolutions of a cuboid with radius R=1,2,3 and the number of neighboring pixels P=4 are presented. The Hamming distance values between a pair of cuboids of different scales are combined to obtain the final feature vector

Figure 7 shows the size of a 3D 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 central coordinates of the cuboid, they remain fixed and are saved for each patch of the video sequence. 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 with respect to scale changes. As an example, in Figure 7, only 8 cuboids are presented, in practice their number can vary, in this work, 20 pairs of cuboids are built inside each image patch. Then, for each pair of cuboids (in Figure 7 connected by a red line), the Hamming distance is calculated. The values obtained for each pair of cuboids of different sizes are sequentially written into a single vector, which characterizes a three-dimensional patch in several resolution.

The Hamming distance $d_{xy}$ is the number of positions at which the corresponding characters of two words of the same length $x_k$ and $y_k$ are different:

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

As mentioned earlier, the operator of VLBP is calculated only for three frames, it is supposed to be applied sequentially to the next frames for the entire video, with the construction of a histogram for every three frames. All the obtained histograms are combined sequentially into a single resulting feature vector for multi-frame training, as shown in Figure 8.

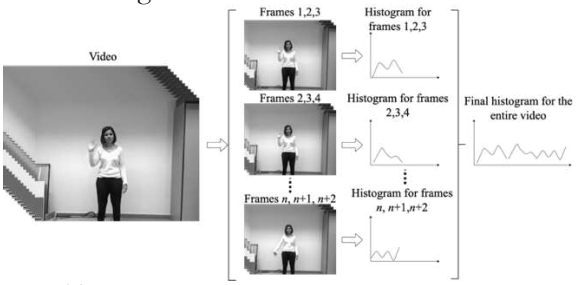


**Fig. 8** The principle of constructing a single feature vector for the video sequence under consideration.

The proposed functions extract information at three different levels from an image patch: resolution, orientation, and scale. Simultaneous extraction of micro-blocks of different sizes leads to their description in several resolutions. The random selection of sampling points facilitates obtaining information in several orientations. 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 use of the local binary template operator for calculating the histogram of a cuboid provides a number of advantages: invariance with respect to brightness, relatively low computational costs due to binary calculations.

### 3.4 Posture assessment based on human skeleton modeling

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

To build a human skeleton, a convolutional recurrent neural network we use [27], which takes video sequence frames as input and learns to build heat maps for each key (joint) point, where the location of the key point is obtained as a heat map mode [28-31].

The human body skeleton is built on the basis of 16 key points, joints, namely: 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, left shoulder (Figure 9). The action recognition method based on the construction of a human skeleton is described in detail in [17, 22, 32-33].

**Fig. 9** An example of building a skeleton for an image [17,22] from the Leeds Sports Pose Dataset test kit [34].

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 joint (Figure 9). To analyze posture and recognize human actions, a set of geometric features is used, which informatively describes the distance between the joints of the human body. 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, the length of the body carries an important informative load [17].
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. Thus, the descriptor is formed from the following data:
- X, Y coordinates of all joint points $j_n$;
- body length from ankle to crown;
- distance from ankle to hip;
- distance from shoulder to wrist;
- the distance between the wrists of the right and left hand;
- the center of gravity of the body.
In order to prepare the data, the coordinates are normalized relative to the body length and relative to the center of gravity.
The formulas describing the human body model correspond [17].
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 [17].

### 3.5 Classification

At the final stage, descriptors are fed to the classifier to categorize the actions performed on the video sequence. 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. In this paper, it is proposed to classify descriptors separately using a multiclass SVM, followed by a decision to categorize the action that occurs in the video clip [17].

## 4 Experimental results

We have tested our method on the UCF101 data. It is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories.

The results of calculations the accuracy is shown in Table 1. To assess the effectiveness of the proposed method, a comparison is made with a method [26].

The analysis of the obtained results indicates that the efficiency of the developed method is quite high and that the use of fused depth and RGB data leads to significantly increase the accuracy. The proposed method has the following advantages over currently existing techniques: it leads to a non–iterative, computationally attractive algorithm that optimizes the use of (global) spatial/temporal and dynamics information and has a moderate computational time.

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

| Methods | Accuracy |
|---|---|
| Method [26] | 89,1% |
| Proposed | 93,2% |

## 5 Conclusions

We propose an algorithm for recognizing human actions on complexly structured images based on a 3-D binary descriptor of micro-block difference. The developed algorithm is based on the fusion of multimodal information obtained by depth sensors and cameras of the visible range. In addition to the main descriptor, the paper proposes to use the analysis of the human skeleton. Such a representation of three-dimensional blocks (patches) of a video sequence by capturing sub-volumes, inside each patch, in several scales and orientations, leads to an informative description of the scene and the actions taking place in it. Experimental results showed the effectiveness of the proposed algorithm on known data sets.

## 6 Declarations

**Funding**
The reported study was funded by Educational Organizations in 2020–2022 Project under Grant NoFSFS-2020-0031.

**Conflicts of interest/Competing interests**
The authors declare that they have no conflict of interest.

**Availability of data and materials**
Not applicable.

**Code availability**
Not applicable.

**Authors' contributions**
All authors contributed in the process of critical literature review. All authors contributed in writing and revising the manuscripts.

**Ethical 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.

## References

[1] Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int. J. of computer vision, 60 (2), 91- 110.

[2] Wang H, Kläser A, Schmid C, Liu CL (2013) Dense trajectories and motion boundary descriptors for action recognition. Int. J. of computer vision, 103(1), 60-79.

[3] Solmaz B, Assari SM, Shah M (2013) Classifying web videos using a global video descriptor. Machine vision and applications, 24(7), 1473-1485.

[4] Ji XF, Wu QQ, Ju ZJ, Wang YY (2017) Study of human action recognition based on improved spatio-temporal features. Springer Berlin Heidelberg, 233-250.

[5] Shahroudy A, Liu J, Ng TT,Wang G (2016) Ntu rgb+ d: A large scale dataset for 3d human activity analysis. Proc. CVPR, 1010-1019.

[6] Zhu W, Lan C, Xing J, Zeng W, Li Y, Shen L, Xie X (2016) Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. Proc. Thirtieth AAAI Conference on Artificial Intelligence.

[7] Zhang S, Liu X, Xiao J (2017) On geometric features for skeleton-based action recognition using multilayer lstm networks. Proc. WACV, 148-157.

[8] Li C, Zhong Q, Xie D, Pu S (2017) Skeleton-based action recognition with convolutional neural networks. Proc. ICMEW, 597-600.

[9] Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3d action recognition. Proc. CVPR, 3288-3297.

[10] Kim TS, Reiter A (2017) Interpretable 3d human action analysis with temporal convolutional networks. Proc. CVPR, 1623-1631.

[11] Brailean JC, Little D, Giger ML, Chen CT, Sullivan BJ (1991) Quantitative performance evaluation of the EM algorithm applied to radiographic images. Biomedical Image Processing II, vol. 1450 of Proc. SPIE, 40–46.

[12] Brailean JC, Little D, Giger ML, Chen CT, Sullivan BJ, et al. (1992) Application of the EM algorithm to radiographic images, Medical Physics, 19(5), 1175-1182.

[13] Brailean JC, Sullivan BJ, Chen CT, Giger M L (1991) Evaluating the EM algorithm for image processing using a human visual fidelity criterion. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '91), vol. 4, pp. 2957–2960.

[14] Xie Z, Stockham TG (1989) Toward the unification of three visual laws and two visual models in brightness perception. IEEE Trans. Syst. Man Cyber., vol. 19, 379-387.

[15] Ginkin GG (1948) Handbook of radio engineering. GEI, Moscow-Leningrad.

[16] Panetta K, Wharton E, Agaian S (2007) Parameterization of logarithmic image processing models. IEEE Tran. Systems, Man, and Cybernetics, Part A: Systems and Humans.

[17] 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

[18] 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.Applied Sciences. 4(3): 331-350.

[19] Zhdanova MM et al. (2019) Model for combining images obtained from sensors of different nature. Proc. Dynamics of technical systems "DTS-2019", 81-84.

[20] 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.

[21] Berkan Solmaz, Shayan Modiri Assari, Mubarak Shah (2012) Classifying Web Videos using a Global Video Descriptor.

[22] 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.

[23] Baumann F, et al. (2016) Recognizing human actions using novel space-time volume binary patterns. Neurocomputing. 173: 54-63.

[24] Ojala T, Pietikainen M, Harwood D (1994) Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. Pattern Recognition, vol. 1.

[25] Maenpaa T (2003) The Local Binary Pattern Approach to Texture Analysis. Extensions and Applications. Oulu University Press.

[26] Zhao G, Pietikäinen M (2006) Dynamic texture recognition using volume local binary patterns. Springer, Berlin, Heidelberg, 165-177.

[27] Belagiannis V, Zisserman A (2017) Recurrent human pose estimation. 12th IEEE International Conference on Automatic Face & Gesture Recognition. 468-475.

[28] Jain A, Tompson J, LeCun Y, Bregler C (2014) Modeep: A deep learning framework using motion features for human pose estimation. In Proc. ACCV, 302–315.

[29] Pfister T, Charles J, Zisserman A (2015) Flowing convnets for human pose estimation in videos. In Proc. ICCV.

[30] Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C (2015) Efficient object localization using convolutional networks. Proc. CVPR, 648–656.

[31] 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.

[32] Pismenskova M, Balabaeva O, Voronin V, Fedosov V (2017) Classification of a two-dimensional pose using a human skeleton. MATEC Web of Conferences, 132, 05016.

[33] Zelensky AA, Pismenskova MM (2018) Method of recognizing human actions on complex-structured images and a background in the form of stochastic textures. Vestnik MGTU Stankin, 3: 116-120.

[34] Johnson S, Everingham M (2010) Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In BMVC 2(4): 5.

# Figures

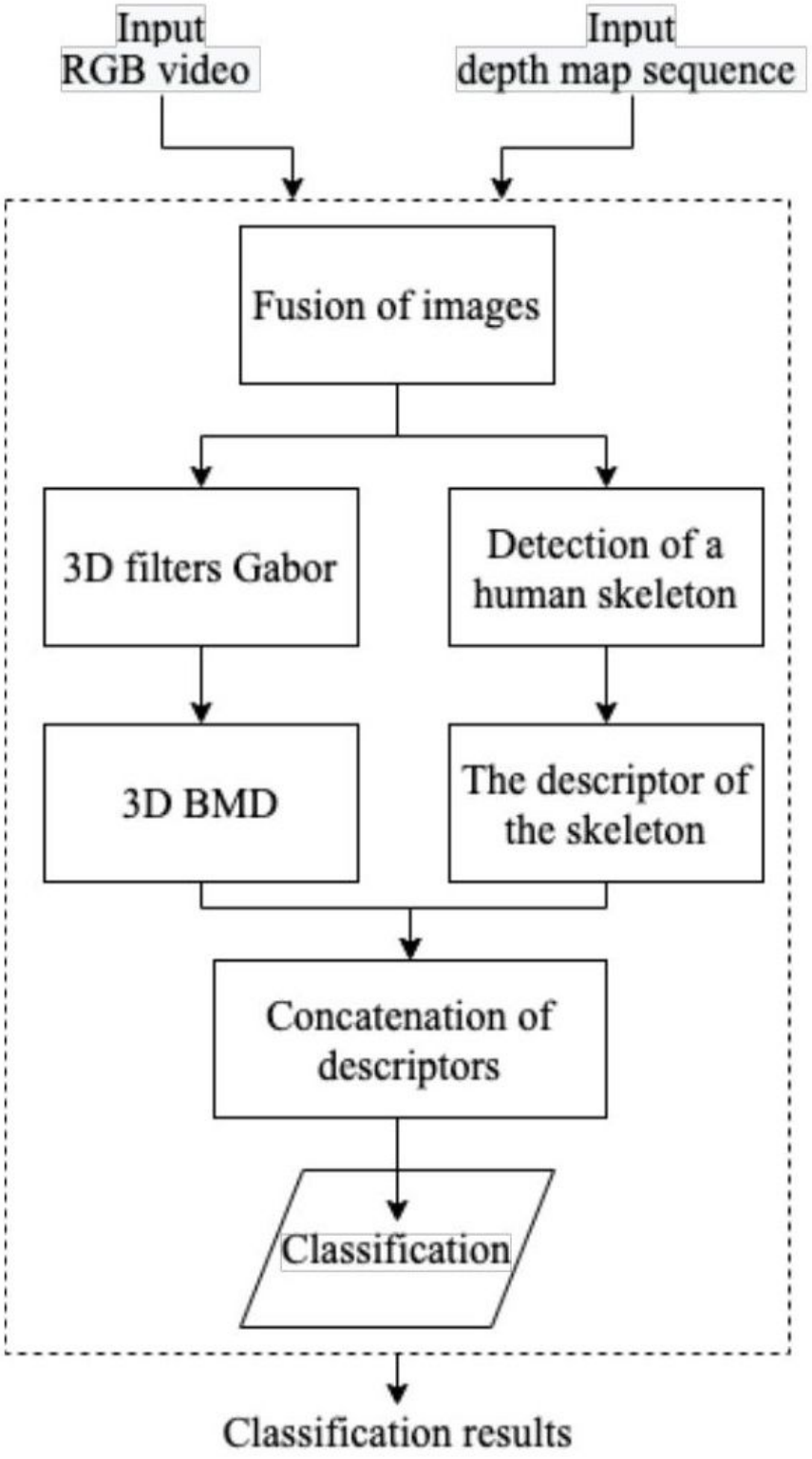


**Figure 1**

Flowchart of the recognition human actions algorithm

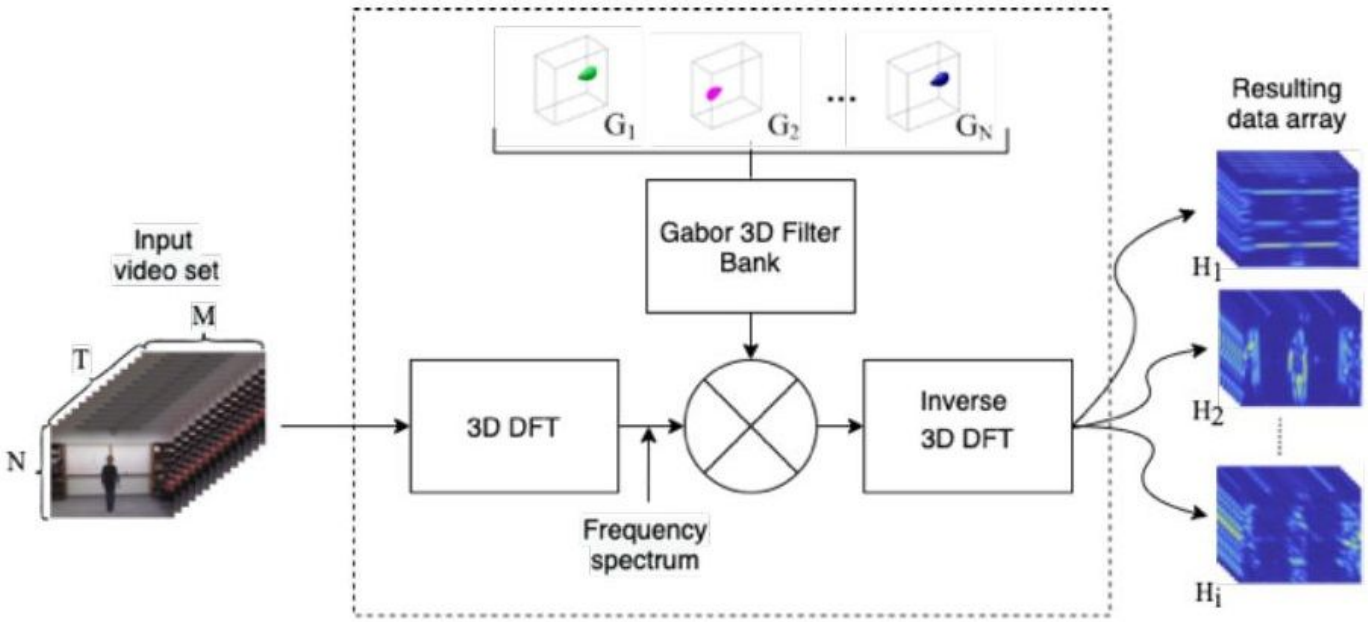


**Figure 2**

Diagram of convolution with threedimensional space-time Gabor filters [17]

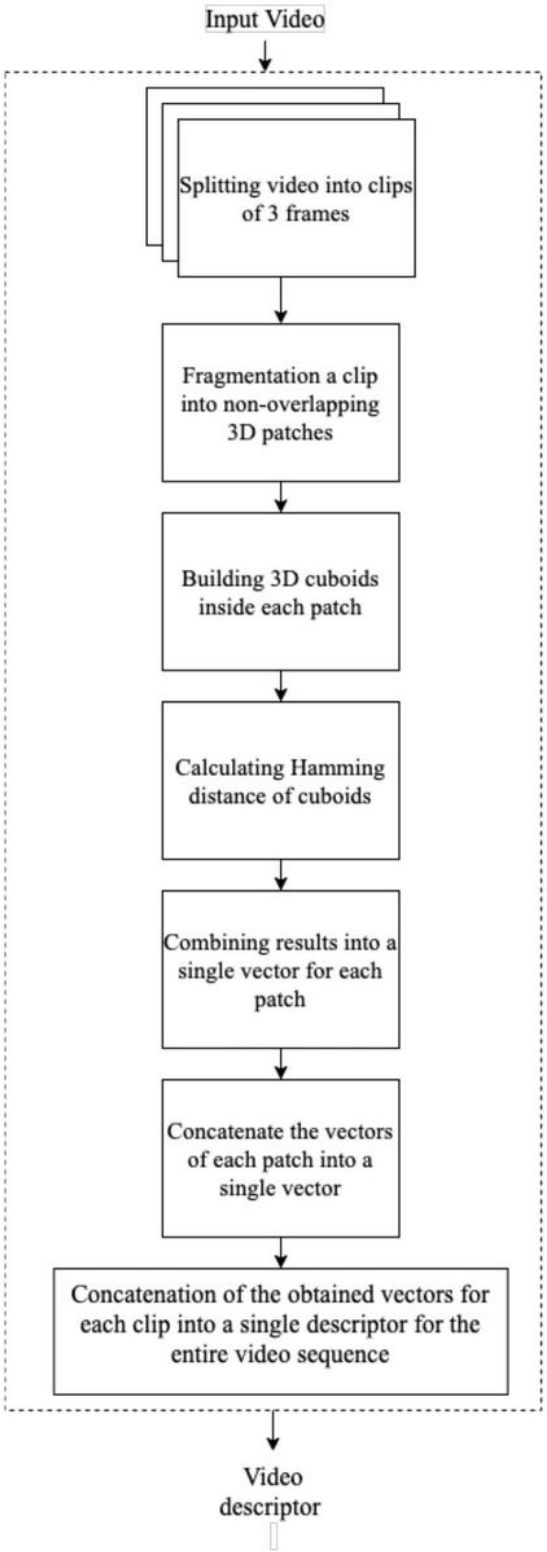


**Figure 3**

Diagram of 3D binary micro-block difference algorithm

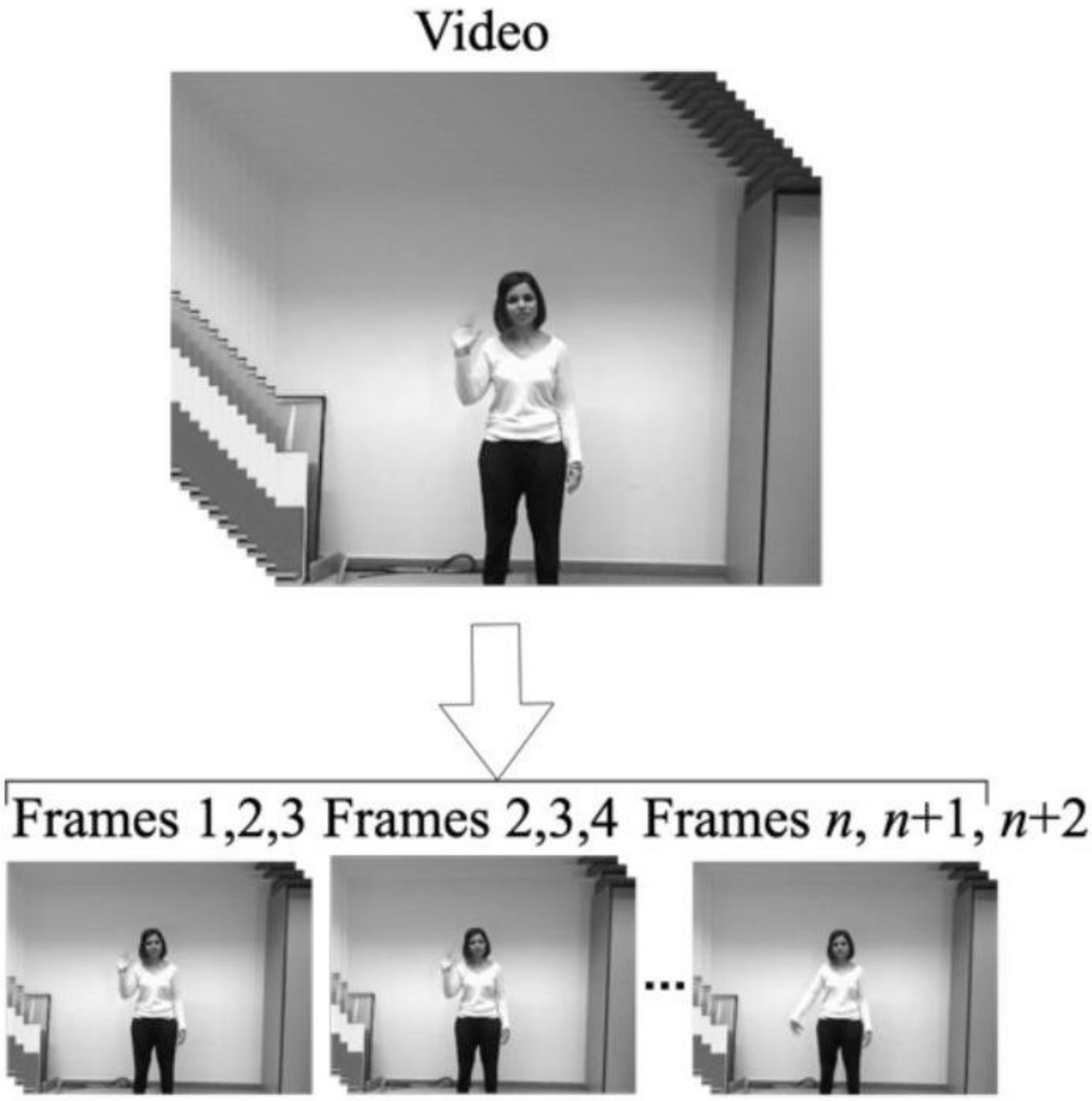


**Figure 4**

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

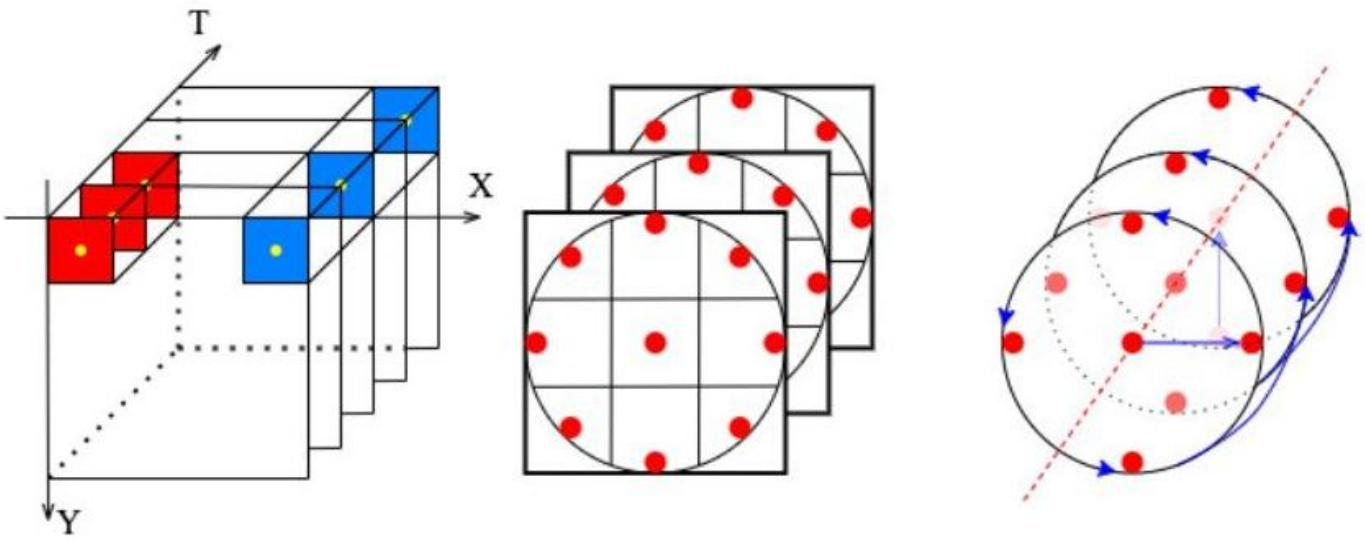


**Figure 5**

The principle of constructing a cuboid inside a spacetime patch of a video sequence [26]

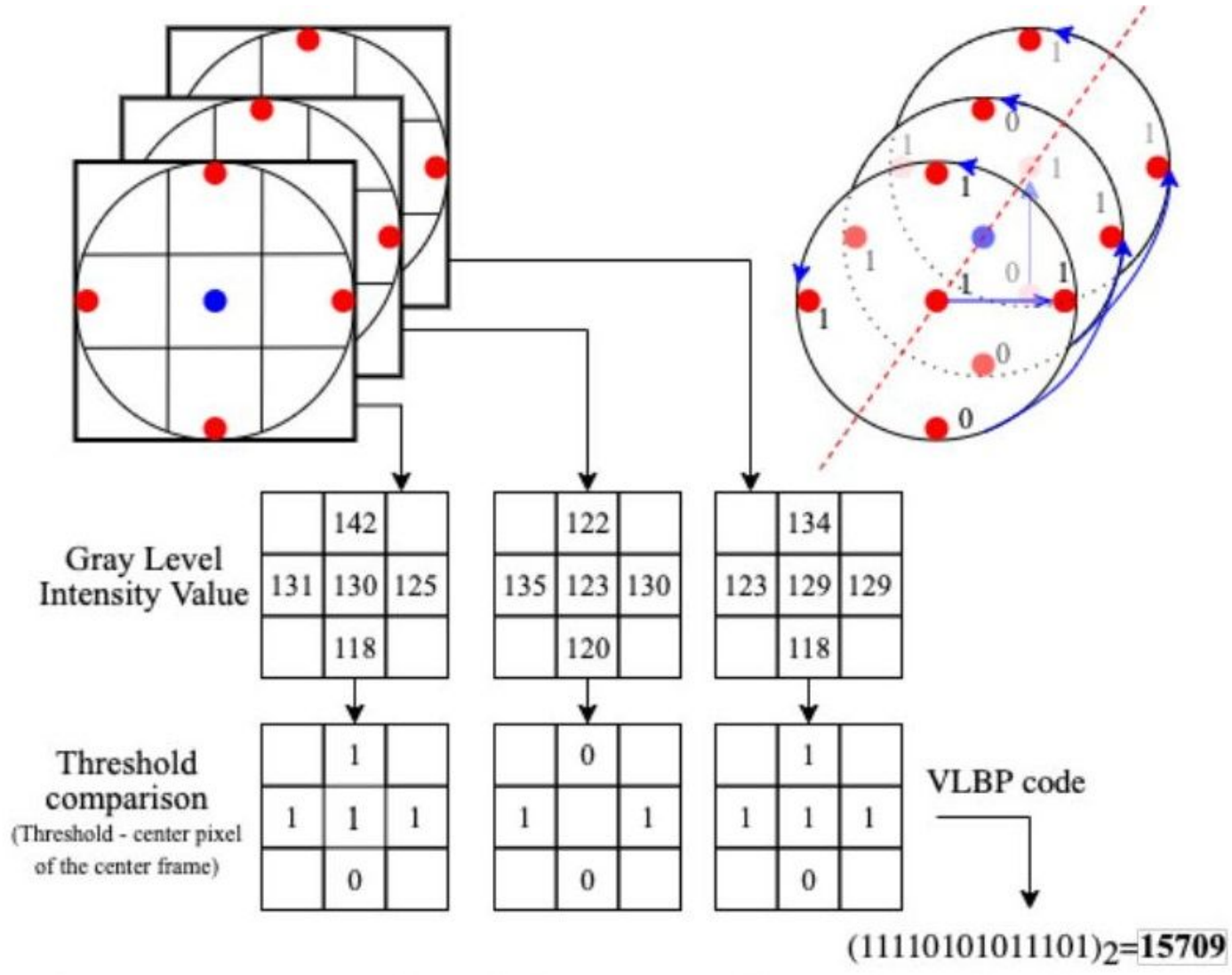

**Figure 6**

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

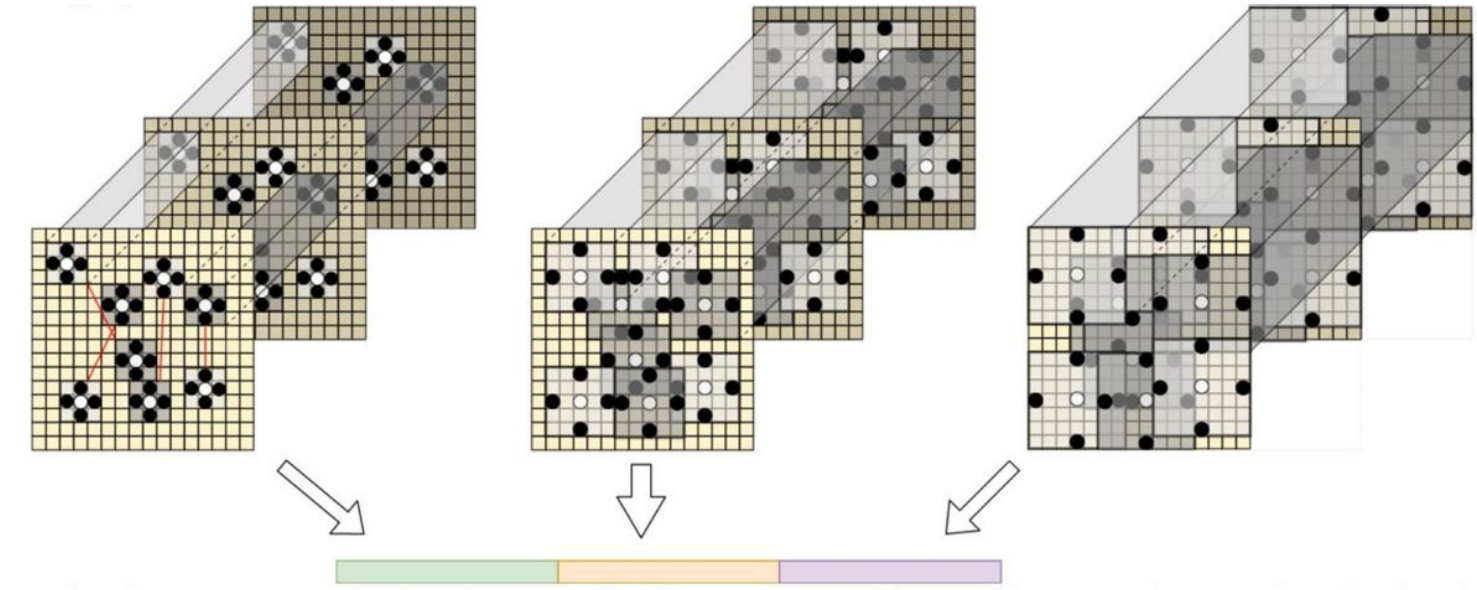

**Figure 7**

An example of constructing cuboids in a 3D patch of size 16×16×3. 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

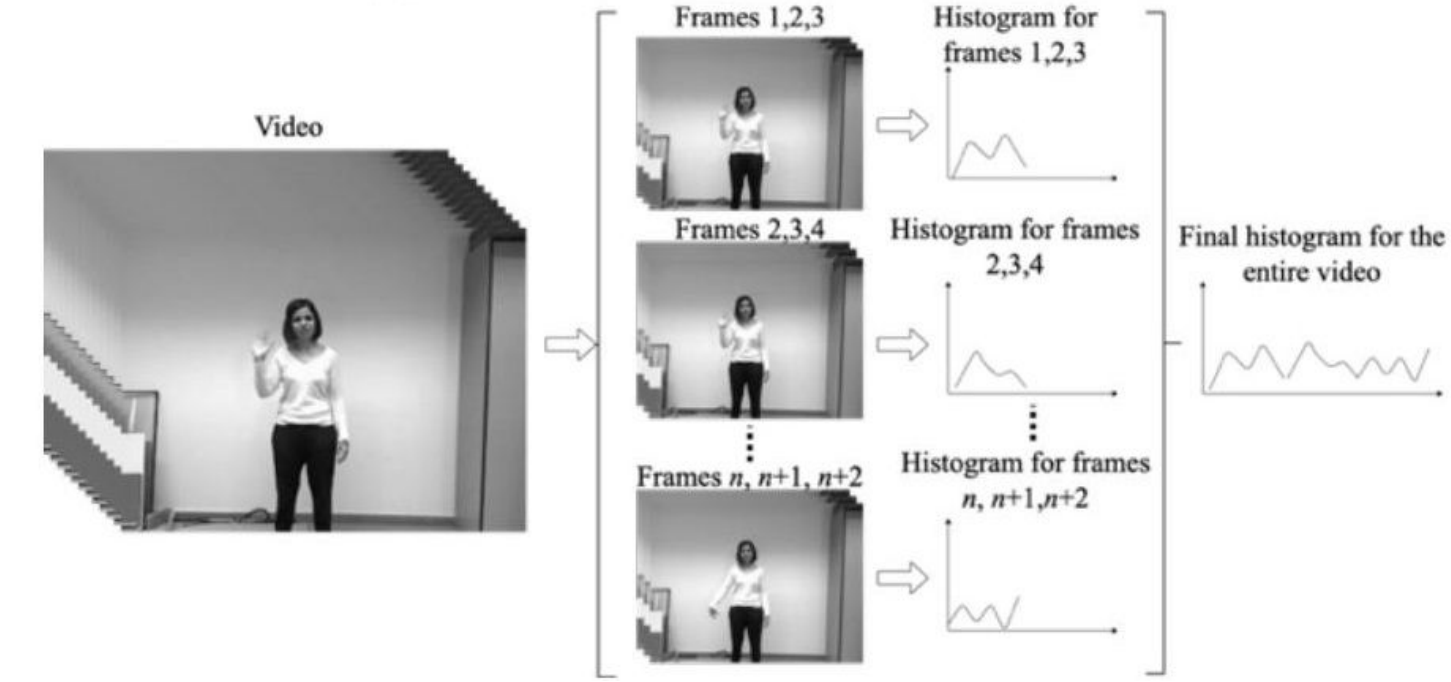


**Figure 8**

The principle of constructing a single feature vector for the video sequence under consideration.


**Figure 9**

An example of building a skeleton for an image [17,22] from the Leeds Sports Pose Dataset test kit [34].