Neural Computing and Applications (2020) 32:8585–8597
https://doi.org/10.1007/s00521-019-04365-9

ORIGINAL ARTICLE

Human action recognition with bag of visual words using different machine learning methods and hyperparameter optimization

Muhammet Fatih Aslan[1] · Akif Durdu[2] · Kadir Sabanci[1]

Received: 20 October 2018 / Accepted: 19 July 2019 / Published online: 26 July 2019

Abstract

Human activity recognition (HAR) has quite a wide range of applications. Due to its widespread use, new studies have been developed to improve the HAR performance. In this study, HAR is carried out using the commonly preferred KTH and Weizmann dataset, as well as a dataset which we created. Speeded up robust features (SURF) are used to extract features from these datasets. These features are reinforced with bag of visual words (BoVW). Different from the studies in the literature that use similar methods, SURF descriptors are extracted from binary images as well as grayscale images. Moreover, four different machine learning (ML) methods such as k-nearest neighbors, decision tree, support vector machine and naive Bayes are used for classification of BoVW features. Hyperparameter optimization is used to set the hyperparameters of these ML methods. As a result, ML methods are compared with each other through a comparison with the activity recognition performances of binary and grayscale image features. The results show that if the contrast of the environment decreases when a human enters the frame, the SURF of the binary image are more effective than the SURF of the gray image for HAR.

Keywords Human activity recognition · Image processing · Speeded up robust features · Bag of visual words · Machine learning · k-Nearest neighbors · Decision tree · Support vector machine · Naive Bayes · Hyperparameter optimization

1 Introduction

Human activity recognition (HAR) is the identification of a person's current activity using information from a variety of motion sensors or cameras. HAR is a difficult field of computer vision and pattern recognition. On the other hand, it is quite a common research area owing to the fact that it has a lot of applications such as video surveillance, health care, sports analysis, entertainment systems, tactical scenarios, elderly care, intelligent houses and human–machine interaction (HMI). HMI especially has a wide range among them [1, 2].

HAR applications are closely related to HMI. A person with suspicious movements can be identified by recognizing human movements with HMI applications. Thus, the information related to the psychological state of this person can be obtained. HMI implementations have accelerated with Industry 4.0. In addition, in HMI-based robotic applications, robots can now perform daily human activities such as cooking, cleaning and washing clothes, without error. In order to teach robots such actions, human activities need to be recognized by the robot [3, 4].

HAR is still popular, despite being an active field for more than a decade. The interaction of people with mobile devices also improves the applications for HAR. Moreover, owing to the difficulties of real-world problems, different studies and researches are being carried out for the HAR.

✉ Muhammet Fatih Aslan
mfatihaslan@kmu.edu.tr

Akif Durdu
adurdu@ktun.edu.tr

Kadir Sabanci
kadirsabanci@kmu.edu.tr

[1] Department of Electrical and Electronics Engineering, Engineering Faculty, Karamanoglu Mehmetbey University, Karaman, Turkey

[2] Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Science, Konya Technical University, Konya, Turkey

These difficulties may be due to factors such as background similarities, partial occlusion, scale changes, point of view, illumination change, appearance, dress variety and camera movement. Each of these difficulties negatively affects the success rate. Various studies show that the performance of HAR applications depends on an appropriate feature extraction method. Action representation and classification stages after the feature extraction are also important [5, 6].

In feature extraction, these features being scale-invariant and rotation-invariant are generally preferred. Scale-invariant feature transform (SIFT) [7] algorithm is a popular feature extraction method as it meets these requirements. SIFT is a local shape descriptor to define local gradient information. With the gradients aligned in the main direction, SIFT becomes rotation-invariant. For the condition of scale-invariant, different Gaussian scale spaces are taken into account when calculating a vector [8]. SIFT is also robust to illumination changes in some cases [9]. In this way, reliable and stable keypoints are detected. SIFT feature descriptors provide unique information for an image and are therefore suitable for image matching. The features are matched one by one between two images, and Euclidean distance is used to match these features. The speeded up robust features (SURF) algorithm [10] is also used to extract local features like SIFT. Panchal et al. [11] and Karami et al. [12] asserted that SIFT detected more features than SURF, but was slow. The SURF algorithm has emerged as an alternative to the SIFT method in terms of speed.

In order to get the most out of the features, a classification method that will strongly distinguish these features is required. For a successful recognition task, an appropriate machine learning (ML) method should be used in addition to robust feature extraction. Also, for a successful classification, the direct use of the features is, at times, not adequate. For example, SURF keypoints are both complex and numerous. Each keypoint represents the features of a pixel. Therefore, the large size of the object to be analyzed means that there can be many local features. If these features are logically combined, they are represented by fewer new features. This will make the system run faster and more accurately. One of the powerful methods to achieve this is the bag-of-visual-words (BoVW) [13] algorithm.

The BoVW model is one of the important concepts in computer vision. The BoVW model was inspired by the bag-of-words (BoW) [14] algorithm. The BoW model is widely preferred in document classification methods. The word frequency in a document is employed as a feature in training of a classifier. The same idea is implemented with BoVW using images as data. In the BoVW model based on image analysis, a visual representation of a word is used. In the BoVW method, firstly, the extracted features (e.g., SURF, SIFT, spatiotemporal interest point (STIP) [15], etc.) are clustered. Each cluster represents a visual word. Afterward, histograms of the visual words frequency in images to be classified are obtained. As a result of the BoVW model, the image is expressed as a histogram of the number of visual words [16]. These histograms are used as a feature for classification.

Since there are many ML methods in the literature, it is necessary to use the ML method which gives the best result according to the application. The robust methods commonly used in the literature are k-nearest neighbor (k-NN), decision tree (DT), support vector machine (SVM) and naive Bayes (NB). k-NN [17] is an unsupervised learning method that is easy to use and understand. DT [18] is a supervised learning method that classifies large data in a similar way to a tree structure (nodes, branches and leafs). SVM [19] is a supervised learning method that classifies features by creating hyperplanes. Finally, NB [20] is a supervised learning method which classifies the data statistically, according to the Bayes method.

To find the most suitable ML method, it is also necessary to configure the methods, since the result of an ML method depends on the parameters contained in that method. To make the most of the ML method used, most of the ML algorithms must be configured before training, regardless of whether they are supervised or unsupervised. Hence, they are improved using hyperparameter optimization (HO) [21]. This optimization, called also a hyperparameter search, not only increases the performance of the training process but also increases the quality (e.g., prediction accuracy) of the ML method. Numerous algorithms, such as grid search, swarm optimization algorithm and Bayesian optimization, can be used for hyperparameter search [22].

In this study, KTH (Royal Institute of Technology (KTH; Swedish: Kungliga Tekniska Högskolan)) [23] and Weizmann [24] dataset is used for HAR. Based on these commonly used datasets, the proposed method is evaluated. In addition, we are used our own dataset to expand the evaluation. In practice, speeded up robust features (SURF) are used as a feature extraction method. Then, these features have been reinforced with the BoVW algorithm. For this BoVW approach, k-means [25], the clustering algorithm, has been used. With this algorithm, k-centers are created. Each of these centers represents a visual word. Finally, histograms are created based on the frequency of the visual words. These steps are common to many studies which use SURF.

The main contribution of this study is the presentation of a different approach for the image used in feature extraction. Especially in previous studies using the BoVW approach, the features were generally extracted from preprocessed grayscale images. However, in this study, for the purpose of recognizing simple activities, besides the grayscale image features, the features of the binary images are also extracted, and the results are compared. In addition, the BoVW algorithm, which is usually used in conjunction with SVM, is also classified using k-NN, DT and NB methods. Prior to the classification process, the ML parameters are optimized with Bayesian optimization, and therefore, the system performance is enhanced. The classification accuracies obtained as a result of the experiments performed show the effectiveness of the proposed system.

2 Related works

A number of studies have been conducted using different approaches for HAR. Plötz and Guan [26] stated that HAR applications are important but are subject to significant limitations due to problems such as noise, ambiguity and missing data. Therefore, it was emphasized that deep learning should be used instead of traditional ML methods in applications such as segmentation and classification. However, in deep learning, there is a requirement for a big data and large computational power for complex calculations used in training. For this reason, deep learning is not more advantageous for datasets containing limited data such as KTH and Weizmann. For example, in a study conducted by Baccouche et al. [27], deep learning was implemented for HAR and 94.39% accuracy was achieved with the KTH dataset. However, in our study, 95.33% accuracy was achieved using ML methods.

The increase in the success rates of HAR applications depends on the different ways of image processing, action representation, feature extraction and classification, or the combination of these methods with different methods. Owing to these reasons, different methods are applied to existing datasets. In contrast to other region-based approaches, Rahman et al. [28] performed the HAR using negative space. Dynamic time warping (DTW) was used for classification. As a result of the study with the KTH database, the accuracy rate was obtained as 94.67%. With regard to the problem of how these human actions will be represented after the action is determined, Zhang et al. [29] proposed a novel motion-based representation called motion context (MC). With this representation, the distribution of the motion words (MWs) over relative locations in a local region around the reference point was captured, and the human actions in motion images (MIs) were modeled. In that study, which used two different training strategies, the best success rate for the KTH was obtained as 91.33%, with leave-one-out (LOO) strategy. Singh et al. [30] practiced a different feature extraction approach using directionality-based feature vectors. For this, the silhouette was revealed with adaptive segmentation. Recognition was made by calculating directional vectors (DVs) on the silhouettes. This study was carried out with different datasets (UoS-HID, UoT-DB, UoA-DB, etc.). In addition, in that study, multiple accuracy values ranging from 85% to 99% were obtained for many different situations (length of frames, fps, activity type). Bian et al. [31] proposed a transfer topic model (TTM) which is a different recognition method which consists of cross-domain BoW representation and regularized target domain topic estimation. This study was specifically developed for scenarios where the target domain has limited data. Experiments on the KTH and Weizmann databases were compared with three studies. Accuracy was found between 68% and 78% according to the weighting parameter (λ) value in TTM. In another study, novel type 2 fuzzy topic models (T2 FTM) to recognize human actions were derived by Cao and Liu [32]. Unlike other topic models, this study used type 2 fuzzy sets to encode the uncertainty of each topic. T2 FTM performs better than other state-of-the-art topic models. Experiments were performed on the KTH, Weizmann, UCF and Hollywood2. The most accurate values obtained with the KTH and Weizmann dataset were 92% and 99.6%, respectively. In another study suggesting a new approach to feature extraction, Uddin et al. [33] used depth and optical flow information of human silhouettes for HAR. In that study, the spatiotemporal approach was used along with the hidden Markov model (HMM) and showed a successful performance. A different novel hidden Markov model-based approach for HAR using 3D positions of body joints was put forward by Ding et al. [34]. In [33, 34], datasets different from the KTH were used. While a 97.6% accuracy rate was obtained in [34], a 98% accuracy rate was obtained in [33].

Besides these studies, in recognition and classification applications, the BoW-SVM method [35–40] is a popular method. However, direct use of these methods is not sufficient for the HAR. A novel method for the HAR based on hybrid features was suggested by Vo and Ly [41]. The SURF were used with BoW. In the study, the BoW features were combined with the histogram of oriented gradients (HOG) and the histogram of optical flow (HOF). As a result of classification using SVM, the KTH dataset was classified with a 95.2% accuracy. Finally, Liu et al. [37] presented the partwise BoW (PBoW) representation to outperform the standard BoW-SVM method in the HAR applications. The HAR task was formulated as a joint multitask learning (MTL) problem by transfer learning. In that study, KTH was used and the highest accuracy rate was obtained as 93.4%. In addition to these studies, there are many different studies [42–61] related to the HAR, because, although the HAR task is quite common, it still needs to be further developed.

In studies which use local features such as SURF and SIFT, the features are generally extracted from gray images. For example, in a study related to the HAR, Sun et al. [51] have extracted SIFT features from the original KTH and Weizmann images. In addition to the SIFT properties, holistic properties were also used. As a result, 94% success was achieved with the KTH. In a different study conducted by Liu et al. [62], a novel video descriptor by combining local spatiotemporal features and global positional distribution information of interest points was proposed. As a result of the classification using SVM, KTH data were classified with 94.92% accuracy. Similarly, Moussa et al. [63] extracted SIFT features from gray space images. As novelty, they limited the number of keypoints and applied a normalization for BoVW. They achieved 97.89% accuracy for KTH and 96.66% for Weizmann.

Similar to the grayscale images, binary images are often used in image recognition applications, as binarization is quite advantageous in applications such as medical image processing, document image analysis and face recognition. Singh and Singh [64] performed face recognition according to the features extracted from the binary images. The binary image of the whole face was used as a feature for artificial neural network (ANN). As a result of the study, the face recognition rate was 97.5%. Pandey et al. [65] developed optical character recognition (OCR) using binary document images. Convolutional layers were used for feature extraction. Perner et al. [66] developed a human epithelial (HEp-2) cell classification system. The cell regions were represented by a number of features derived from binary images. These features were then subdivided into six classes using the DT algorithm. Based on these previous studies, it can be concluded that binary image features can also be used for the HAR.

In this study, BoVW method was used for HAR. However, unlike the above studies, SURF keypoints were also obtained from binary images. As a result of the study, a comparison was made with respect to the performances of the gray image and the binary image features. Four different ML algorithms were enhanced using HO to obtain optimal results. The results were compared with the previous study results.

3 Datasets

These datasets used in the application include simple activities. Recognition performance of simple activities is important for the recognition of complex activities, because a simple activity can be considered as part of a complex task. For example, “walking” is a part of the complex activity of “approaching an object to hold it” [67]. In practice, three different datasets were used to show the effectiveness of the proposed method. These are KTH, Weizmann and our data.

The KTH dataset, which is frequently used in the literature, is image-based and includes six types of single action such as walking, jogging, running, boxing, hand waving and hand clapping (see Fig. 1). These activities are performed several times by 25 people in four different scenarios. This database contains 2391 sequences. All images have a static background and are recorded using a camera with a frame rate of 25 fps.

The image-based Weizmann dataset, which is widely preferred in applications, also includes single action similar to KTH. Ten different actions (bend, jack, jump, P-jump, run, side, skip, walk, wave1 and wave2) (see Fig. 2) performed by nine different people were recorded in a static background using a camera with a frame rate of 25 fps.

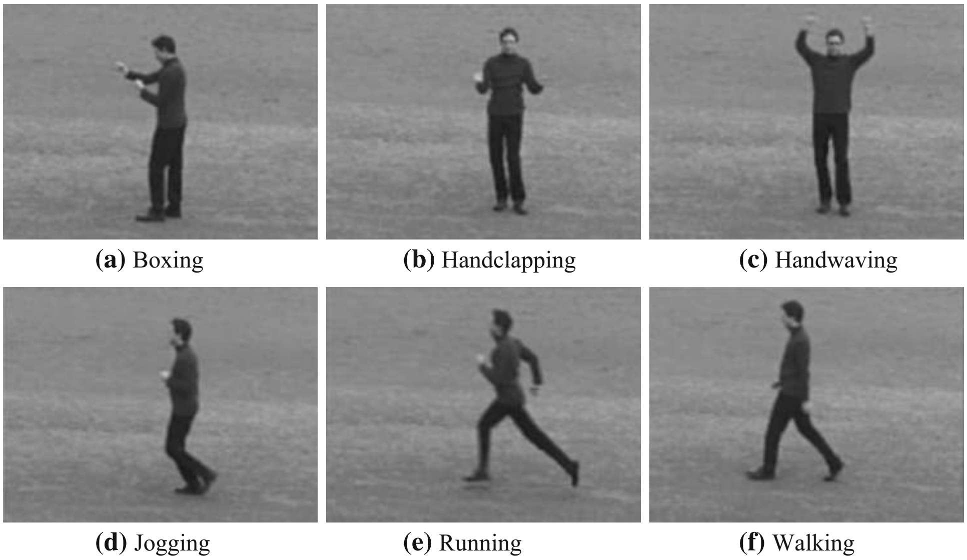
In addition to the widely used KTH and Weizmann datasets, we have recorded our own video sequences to make a more comprehensive evaluation of the proposed approach. A total of 80 videos, 10 s each, were recorded. Totally, there are eight different activities (bend, boxing, hand clapping, hand waving, running, side, skip and walking) in total and the background is static (see Fig. 3). The activities were carried out by two different people. Videos were recorded at a frame rate of 30 fps on a phone with a 12 MP camera.

4 Proposed work

In general, HAR studies consist of data acquisition from a sensor or a camera, segmentation, feature extraction and classification. Most work for data acquisition implements frame-based HAR using the camera. Similarly, this study was carried out using video frames. An overview of the proposed algorithm is shown in Fig. 4.

When Fig. 4 is examined, each original frame in the videos is recorded in binary and gray format after preprocessing. The recorded frames are then parsed into training and test data. Then, for both the training and test images, the steps of feature extraction, clustering, histogram creation and HO are performed, respectively. Once these operations are completed for the train and test images, the ML algorithm is trained using the features of the train images. This algorithm is tested by the test images. As a result of the test, the activity type in each frame is determined. This loop is done for the four ML algorithms, three datasets and two feature types (binary, gray). As a result, 24 different results are obtained.

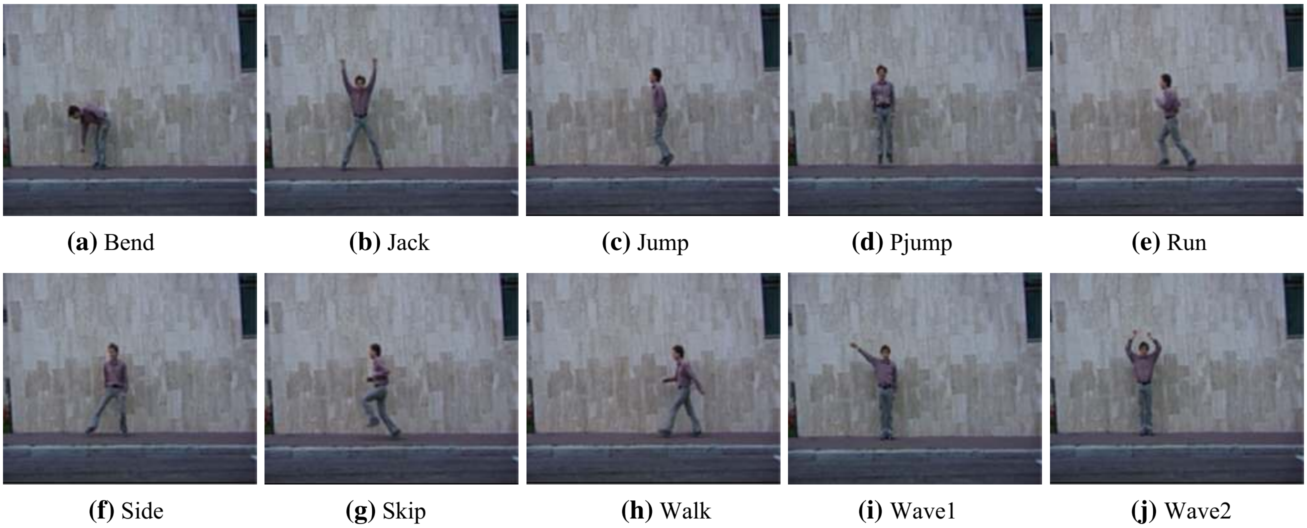
In this application, real-time HAR is achieved using image processing, feature extraction, ML methods and HO. First, the preprocessing is performed on each frame in the videos. In the preprocessing step, operations are performed to facilitate the identification of human activity. Background subtraction technique is successfully applied



(a) Boxing **(b)** Handclapping **(c)** Handwaving

(d) Jogging **(e)** Running **(f)** Walking

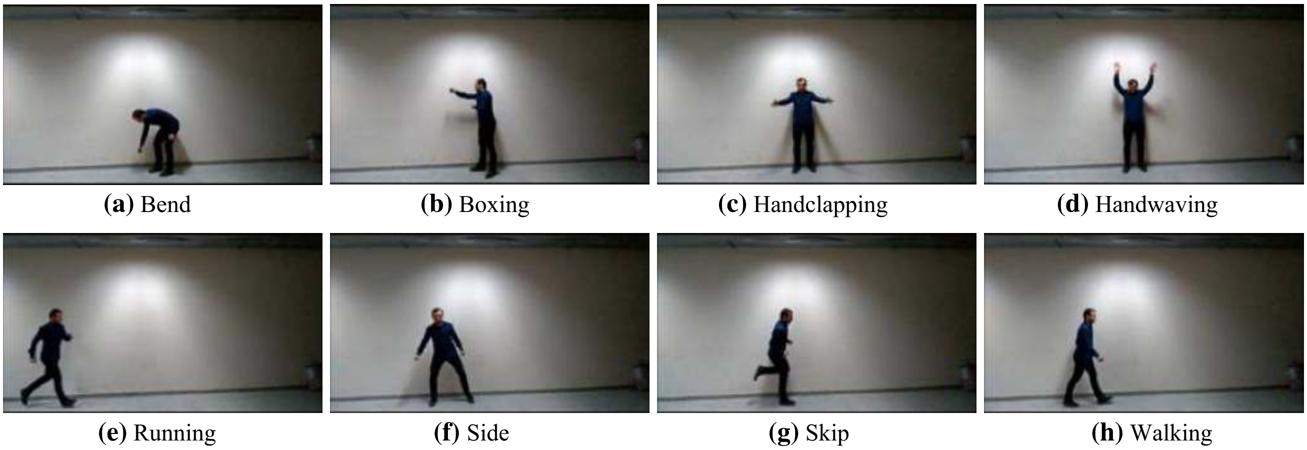
Fig. 1 KTH dataset [23, 68]



(a) Bend **(b)** Jack **(c)** Jump **(d)** Pjump **(e)** Run

(f) Side **(g)** Skip **(h)** Walk **(i)** Wave1 **(j)** Wave2

Fig. 2 Weizmann dataset [24]



(a) Bend **(b)** Boxing **(c)** Handclapping **(d)** Handwaving

(e) Running **(f)** Side **(g)** Skip **(h)** Walking

Fig. 3 Our dataset

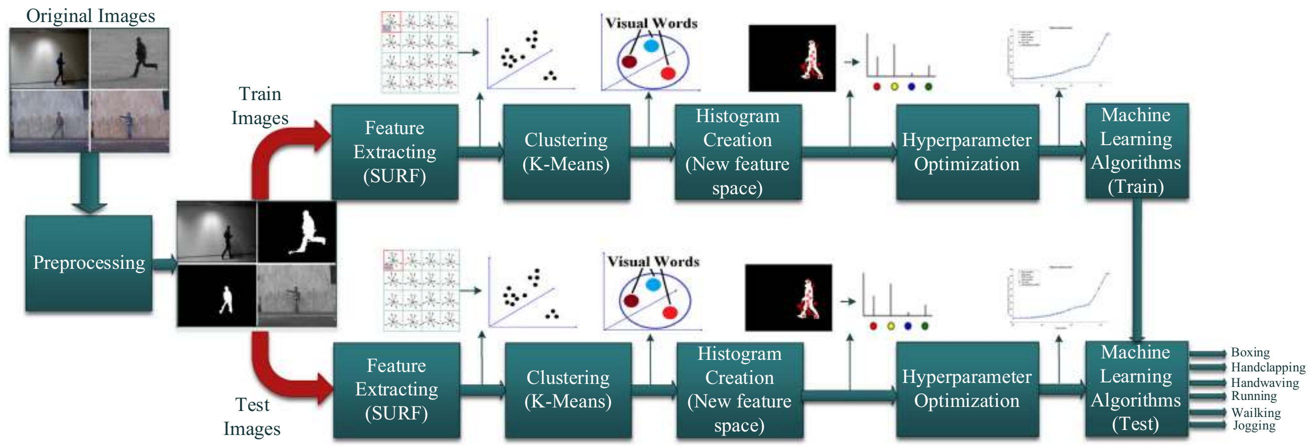

Original Images
Preprocessing
Train Images
Test Images
Feature Extracting (SURF)
Clustering (K-Means)
Visual Words
Histogram Creation (New feature space)
Hyperparameter Optimization
Machine Learning Algorithms (Train)
Feature Extracting (SURF)
Clustering (K-Means)
Visual Words
Histogram Creation (New feature space)
Hyperparameter Optimization
Machine Learning Algorithms (Test)
Boxing
Handclapping
Handwaving
Running
Wailking
Jogging

Fig. 4 BoVW-based image classification algorithm

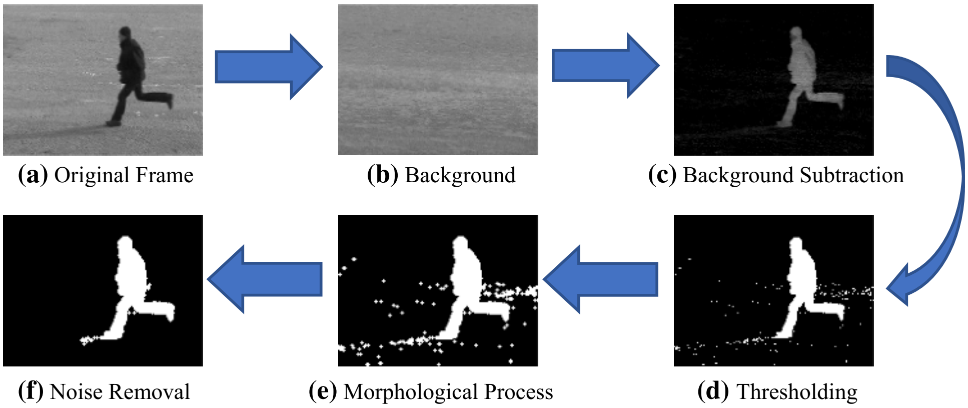
because of the fact that videos in KTH, Weizmann and our dataset are recorded in same or similar environment. In these datasets, a background image occurs in the image which does not have a person. For this, first, a frame which the person is not a part of is recorded. This frame is used as a background for the first steps. Then, every image in which there is no person is recorded as a new background. After subtracting the original image from background, the binarized image is obtained by setting the threshold value. In the following steps, morphological processing is performed to enhance the image, and finally, noise in the image is removed. These steps are shown in Fig. 5 using an example from the KTH dataset.

The steps shown in Fig. 5 are applied to all video frames. In this way, a new preprocessed dataset is created. Area values of the white pixels are also used as a limiting factor while constructing the dataset. Frames of the same class with a specific range of area are recorded separately, both in binary and in gray space. With area limitation, frames without a human are not recorded as data. In the generated dataset, each binary and gray space image is grouped in six different groups for KTH, ten for Weizmann and eight for our dataset. These datasets have different frame numbers. Therefore, for each class, 800 frames (120 × 160) from KTH data, 405 frames (144 × 180) from Weizmann and 540 frames (214 × 120) from our dataset are collected. A part of the "running" and "side" class frames of KTH and Weizmann data is shown in Fig. 6.

To identify activity in the recorded frames similar to Fig. 6, the SURF-BoVW algorithm is used. The BoVW method produces visual words from these frames. Visual words are created using SURF. The SURF keypoints are extracted from both the binary and gray images as shown in Fig. 7. The keypoints (red stars) represent the rotation-invariant, scale-invariant and the noise-resistant pixels.

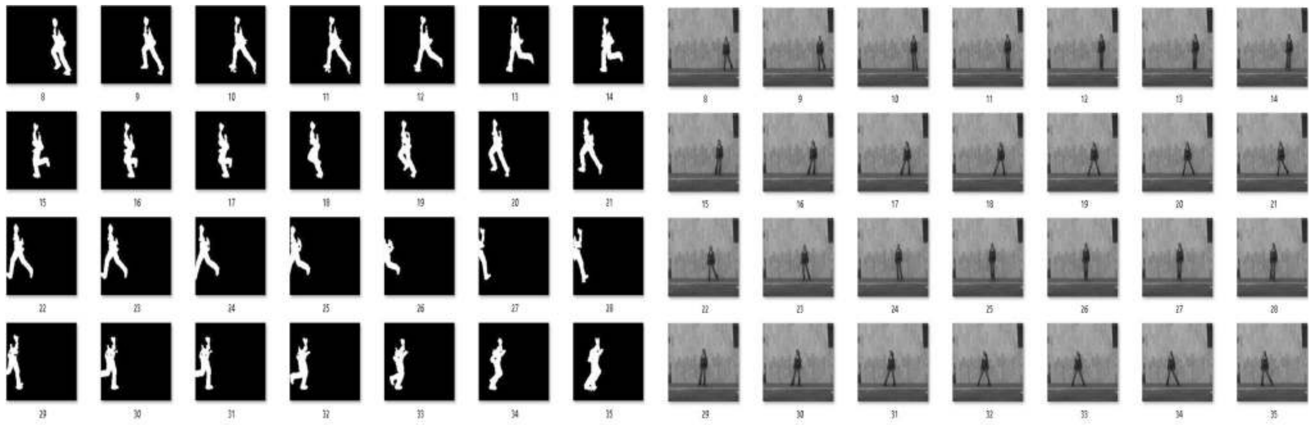
As it can be seen from the studies in Sect. 2, in HAR studies using SURF or SIFT, these features are extracted from gray or original images. However, in this study, the SURF properties of binary images are also used. In fact, the gray space image contains more gradient information than the binary image, since the gradient information is obtained by calculating the density differences between pixels. Therefore, the density change and the direction of change in a binary image do not represent that image exactly. Gradient information is particularly important for many

Fig. 5 Preprocessing of video frames



(a) Original Frame **(b)** Background **(c)** Background Subtraction

(f) Noise Removal **(e)** Morphological Process **(d)** Thresholding



(a) Binary Image dataset (KTH,Running) **(b)** Gray Image dataset (Weizmann,Side)

Fig. 6 Generated running and side class data

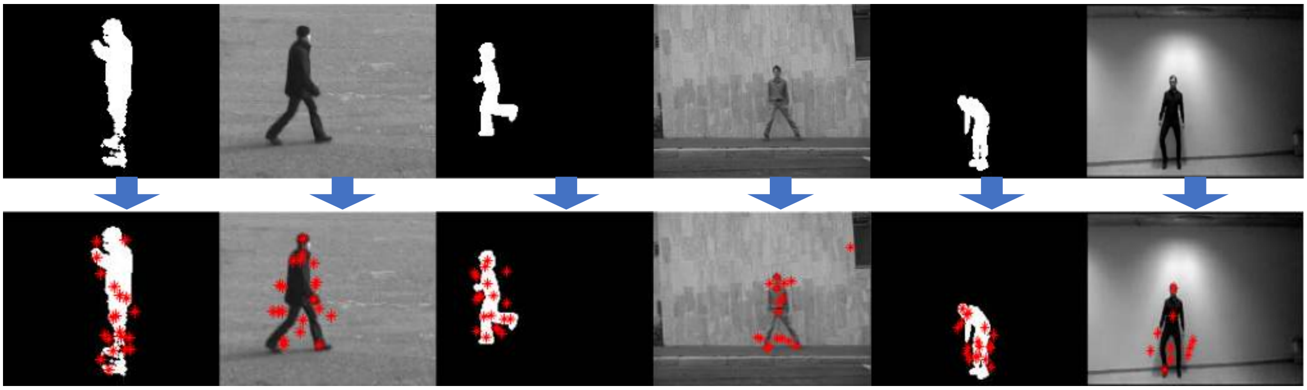


Fig. 7 Extracting SURF from binary and gray images for each data

recognition applications, especially image matching applications. However, in HAR, it is not vitally necessary for a feature to match another feature in a different image. What is more important is that the features represent that activity as a whole. Hence, a foreground information with edge features rather than what the object in the image contains may represent that activity in question. In addition, the features in the binary image are less complex and do not contain background keypoints. The SURF in the binary image generally include scale-invariant and rotation-invariant edge features. These features may be sufficient for HAR.

The next step after the SURF are extracted is the clustering phase. The k-means is used as a clustering algorithm. The number of k is set to 500. In total, there are 4800 images for KTH, 4050 images for Weizmann and 4320 images for our dataset. 80% of these images are used for testing and 20% for training. The number of keypoints to be used for the KTH, Weizmann and our dataset is 3686400, 4292350 and 4478976, respectively. The different number of keypoints depends on the foreground of the image and the number of frames used for training. For a faster and more accurate classification, these features are divided into 500 clusters. Now, the images can be expressed using 500 words. With this clustering, visual word distribution in each image is found. According to this distribution, histograms of each image are created. The histograms now represent a feature vector.

Histograms only provide frequency information. That is, information about the location of the extracted features is not found in the histograms. A histogram generated according to a gray frame is shown in Fig. 8. This graph shows the frequency of the SURF extracted from the frame according to clusters. In this way, the classification is made by using these histogram values obtained from each frame. That is, for ML approaches, the inputs contain 500 frequency information of each frame, while the outputs include the activity index.

After histograms are obtained for each binary and gray image shown in Fig. 8, the classification of histograms can be performed using ML methods. In this study, for each ML method used, appropriate parameters (hyperparameters) are determined using HO. The types of hyperparameters vary according to the ML method. Hyperparameters

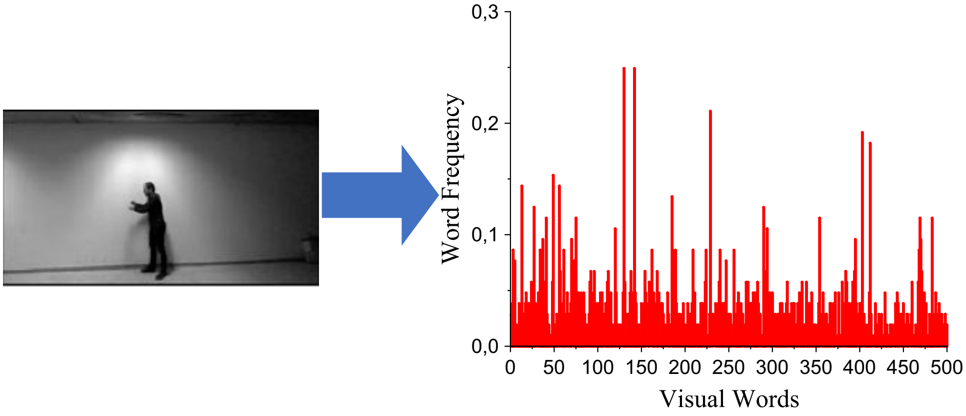


Fig. 8 Frequency of the visual words in an image

are the number of neighbors and distance for k-NN, minimum leaf size for the DT, coding method, box constraints (C), kernel scale for SVM, distribution type and kernel width (in case the distribution type is kernel) for NB. These parameters have been affected by training processes. The values of these calculated parameters using HO are shown in Table 1. After that, training and testing are carried out using these values. In the training process, the number of iterations is used as stop criteria for all ML methods.

In SVM, as a result of HO, the most optimal coding is provided by the one-versus-one (OVO) approach. According to the OVO approach, multi-class problems are transformed into binary classes. These problems can be thus solved with binary classifiers. Eventually, the results are combined to provide a solution to the multi-class problem [69].

5 Results and comparison

Table 2 shows the test accuracy rates and the training time obtained by ML algorithms according to the binary and gray image features. The train time of the NB is generally much higher than the other methods as the distribution type is kernel. When Table 2 is examined, it is seen that the best result values are obtained using k-NN and SVM. Also, the confusion matrices obtained as a result of these ML algorithms are shown in Fig. 9.

The results of the study shown in Table 2 are compared with previous studies. When Table 3 is examined, it is seen that the proposed method is quite successful for KTH and Weizmann data.

6 Discussion

HAR accuracy values shown in Table 2 depend on the characteristics of the human silhouette that represent the foreground. Thus, the features extracted from the images in

Table 1 Hyperparameters of ML algorithms obtained using HO based on binary and gray image features

ML alg.	Hyperparameters	KTH binary images	Weizmann binary images	Our binary images	KTH gray images	Weizmann gray images	Our gray images
k-NN	Number of neighbors	1	1	1	1	1	1
	Distance	Euclidean	Cosine	Correlation	Cityblock	Correlation	Cityblock
DT	Minimum leaf size	1	1	3	3	1	9
SVM	Coding method	OVO	OVO	OVO	OVO	OVO	OVO
	Box constraints (C)	56.472	5.859	0.0364	0.1638	0.2589	0.0010
	Kernel scale	0.0039	0.0011	0.0027	0.0062	0.0054	0.0038
NB	Distribution type	Kernel	Kernel	Kernel	Kernel	Kernel	Kernel
	Kernel width	0.00119	0.00041	0.00034	0.00095	0.00087	0.00193

Table 2 Comparison of ML algorithm results for binary and gray image features

	k-NN	DT	SVM	NB
Binary image features				
KTH				
Train time (s)	**5.57**	8.73	6.04	343.48
Accuracy (%)	**95.33**	86.16	95.17	86.33
Weizmann				
Train time (s)	3.99	11.73	**11.63**	333.47
Accuracy (%)	86.91	70.61	**90.24**	78.27
Our				
Train time (s)	4.32	7.44	**6.77**	344.60
Accuracy (%)	95.94	86.92	**96.52**	90.62
Gray image features				
KTH				
Train time (s)	**8.37**	9.51	14.54	321.27
Accuracy (%)	**96.14**	85.72	95.17	86.04
Weizmann				
Train time (s)	3.88	18.72	**46.39**	260.71
Accuracy (%)	87.78	69.63	**91.11**	73.70
Our				
Train time (s)	7.89	15.38	**6.95**	313.59
Accuracy (%)	89.58	80.09	**92.71**	80.67

Bold values indicate the methods that have best performance

which the foreground is strongly separated from the background represent the foreground more accurately. To determine such datasets, the contrast values of both the background and the original frame image are calculated using the gray-level co-occurrence matrix (GLCM) [70], a texture analysis method. These values are shown in Fig. 10. The values obtained for the original frames are the average values of the frames in which the human exists. For datasets with different backgrounds, the average value is calculated.

When the values in Fig. 10 are examined, the contrast is increased in the KTH and Weizmann dataset with the foreground (human). The rate of increase is higher for KTH. However, the contrast is reduced due to the foreground in our dataset. There is a relationship between this situation and the accuracy values shown in Table 2. Although the grayscale image features for KTH and Weizmann are more accurate, the results of the binary image features for our data are more accurate. If the foreground in an image is prominent, the grayscale features represent the foreground better. Otherwise, if the foreground does not increase the contrast of frame, the use of the local features of binary images for HAR will be more accurate. Because, in the binary image, the background is removed, the foreground is made clear, and therefore, the contrast is increased. Although the features obtained from this binary image are insufficient in terms of the gradient, it gives better results for the HAR than the grayscale image. Moreover, considering the train times in Table 2, these results indicate that binary images are more advantageous in terms of train time.

It should be noted that the threshold value must be sensitively adjusted in order to make an accurate segmentation in the binary image. When SURF are obtained from grayscale images, the threshold setting is not required. However, in order to obtain a correct result from the binary image, the appropriate threshold must be determined. For example, in this application, the threshold value for KTH and Weizmann is 0.1, while for our data, this value is 0.4.

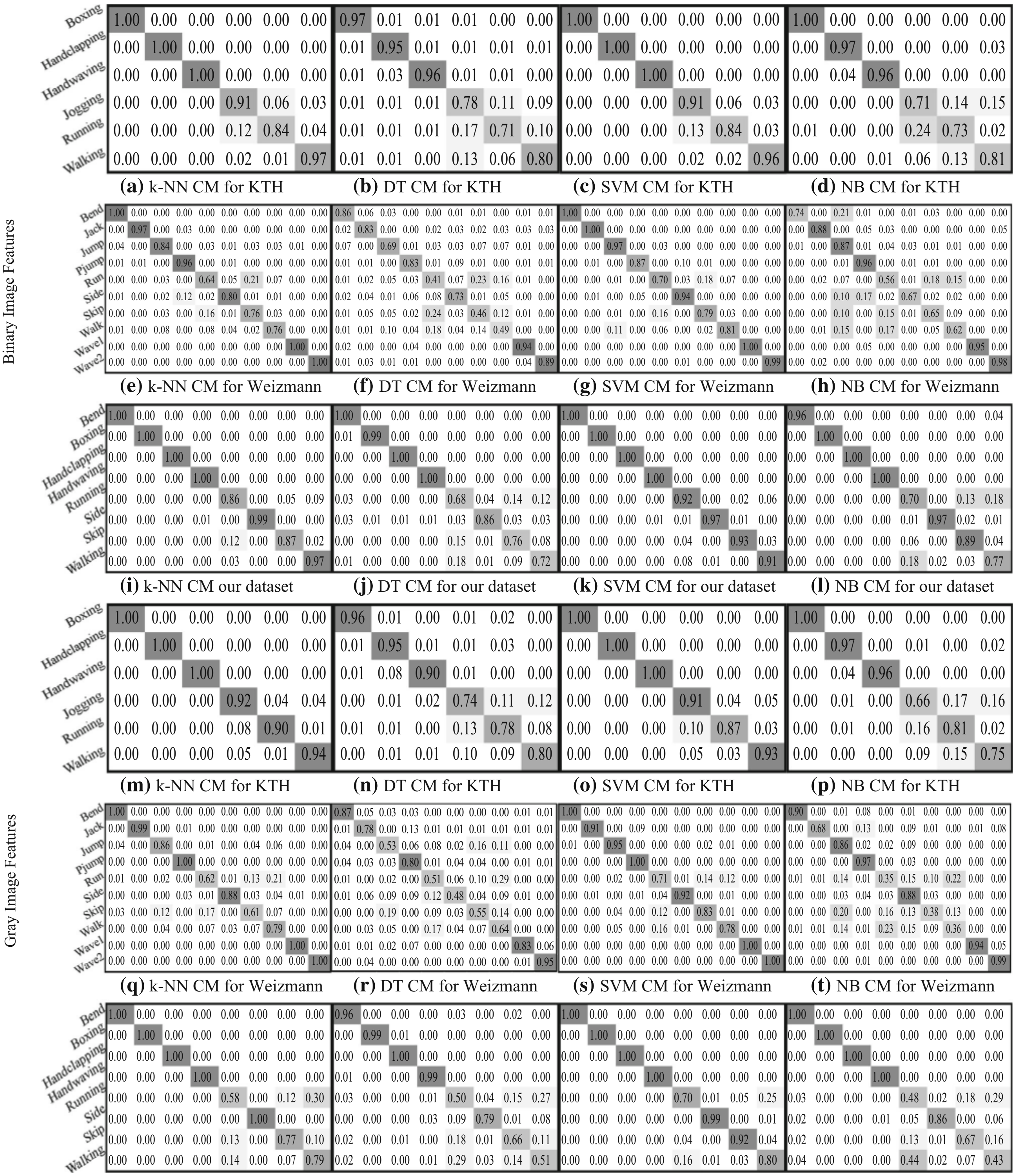
At present, it is easy to reach a lot of data owing to the development in artificial intelligence applications. In particular, large-size data are required for training in deep learning practices like the convolutional neural network (CNN). However, while using small datasets such as KTH and Weizmann, they tend to suffer from the overfitting problem. Moreover, if large-size data are used in deep learning, the training time is a problem. In this study, it is seen that the training time of binary image features is shorter than gray image features. This shows that the binary frame local features are more advantageous in terms of training time, especially in deep learning applications. This inference will be taken into consideration in future works.

7 Conclusion

The aim of is to recognize human movements. In practice, first the frames in KTH, Weizmann and our dataset are preprocessed. Then the SURF are extracted from 4800, 4050 and 4320 frames for KTH, Weizmann and our dataset, respectively. As a result, 3686400, 4292350 and 4478976 features are obtained, respectively. Since there are too many keypoints, they are reduced by clustering (k-means) with the help of the BoVW method. Thus, the images are expressed with fewer features. This provides advantage in terms of speed and accuracy.

In the BoVW algorithm, each cluster represents a word. After clustering, the word frequency in each frame is obtained. Histograms are created according to the frequency of visual words in an input frame. As a result of the BoVW algorithm, each frame is now represented by a histogram. Since the cluster number (k) was set to 500, 500 different types of frequency information are extracted from a frame. Before the ML methods for classification are applied, the hyperparameters are determined using HO.

The above explanation roughly summarizes the work. Like previous related works, this study is presented a different approach for the HAR. It is suggested that the SURF



Binary Image Features

Boxing	1.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	1.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	1.00	0.00	0.00	0.00
Jogging	0.00	0.00	0.00	0.91	0.06	0.03
Running	0.00	0.00	0.00	0.12	0.84	0.04
Walking	0.00	0.00	0.00	0.02	0.01	0.97

(a) k-NN CM for KTH

Boxing	0.97	0.01	0.01	0.01	0.00	0.00
Handclapping	0.01	0.95	0.01	0.01	0.01	0.01
Handwaving	0.01	0.03	0.96	0.01	0.01	0.00
Jogging	0.01	0.01	0.01	0.78	0.11	0.09
Running	0.01	0.01	0.01	0.17	0.71	0.10
Walking	0.01	0.01	0.00	0.13	0.06	0.80

(b) DT CM for KTH

Boxing	1.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	1.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	1.00	0.00	0.00	0.00
Jogging	0.00	0.00	0.00	0.91	0.06	0.03
Running	0.00	0.00	0.00	0.13	0.84	0.03
Walking	0.00	0.00	0.00	0.02	0.02	0.96

(c) SVM CM for KTH

Boxing	1.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.97	0.00	0.00	0.00	0.03
Handwaving	0.00	0.04	0.96	0.00	0.00	0.00
Jogging	0.00	0.00	0.00	0.71	0.14	0.15
Running	0.01	0.00	0.00	0.24	0.73	0.02
Walking	0.00	0.00	0.01	0.06	0.13	0.81

(d) NB CM for KTH

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jack	0.00	0.97	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.04	0.00	0.84	0.00	0.03	0.01	0.03	0.03	0.01	0.00
Pjump	0.01	0.01	0.00	0.96	0.00	0.01	0.00	0.00	0.00	0.00
Run	0.00	0.00	0.03	0.00	0.64	0.05	0.21	0.07	0.00	0.00
Side	0.01	0.00	0.02	0.12	0.02	0.80	0.01	0.01	0.00	0.00
Skip	0.00	0.00	0.03	0.00	0.16	0.01	0.76	0.03	0.00	0.00
Walk	0.01	0.00	0.08	0.00	0.08	0.04	0.02	0.76	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(e) k-NN CM for Weizmann

Bend	0.86	0.06	0.03	0.00	0.00	0.01	0.01	0.00	0.01	0.01
Jack	0.02	0.83	0.00	0.00	0.02	0.03	0.02	0.03	0.03	0.03
Jump	0.07	0.00	0.69	0.01	0.03	0.03	0.07	0.07	0.01	0.00
Pjump	0.01	0.01	0.00	0.83	0.01	0.09	0.01	0.00	0.01	0.01
Run	0.01	0.02	0.05	0.03	0.41	0.07	0.23	0.16	0.01	0.00
Side	0.02	0.04	0.01	0.06	0.08	0.73	0.01	0.05	0.00	0.00
Skip	0.01	0.05	0.05	0.02	0.24	0.03	0.46	0.12	0.01	0.00
Walk	0.01	0.01	0.10	0.04	0.18	0.04	0.14	0.49	0.00	0.00
Wave1	0.02	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.94	0.00
Wave2	0.01	0.03	0.01	0.01	0.00	0.00	0.00	0.00	0.04	0.89

(f) DT CM for Weizmann

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jack	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.00	0.00	0.97	0.00	0.03	0.00	0.00	0.00	0.00	0.00
Pjump	0.00	0.01	0.00	0.87	0.00	0.10	0.01	0.00	0.00	0.00
Run	0.00	0.00	0.01	0.00	0.70	0.03	0.18	0.07	0.00	0.00
Side	0.00	0.01	0.00	0.05	0.00	0.94	0.00	0.00	0.00	0.00
Skip	0.00	0.00	0.01	0.00	0.16	0.00	0.79	0.03	0.00	0.00
Walk	0.00	0.00	0.11	0.00	0.06	0.00	0.02	0.81	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.99

(g) SVM CM for Weizmann

Bend	0.74	0.00	0.21	0.01	0.00	0.01	0.03	0.00	0.00	0.00
Jack	0.00	0.88	0.00	0.05	0.03	0.00	0.00	0.00	0.00	0.05
Jump	0.01	0.00	0.87	0.01	0.04	0.03	0.01	0.01	0.00	0.00
Pjump	0.00	0.00	0.01	0.96	0.00	0.01	0.01	0.00	0.00	0.00
Run	0.00	0.02	0.07	0.00	0.56	0.01	0.18	0.15	0.00	0.00
Side	0.00	0.00	0.10	0.17	0.02	0.67	0.02	0.02	0.00	0.00
Skip	0.00	0.00	0.10	0.00	0.15	0.01	0.65	0.09	0.00	0.00
Walk	0.00	0.01	0.15	0.00	0.17	0.00	0.05	0.62	0.00	0.00
Wave1	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.95	0.00
Wave2	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98

(h) NB CM for Weizmann

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Boxing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.00	0.86	0.00	0.05	0.09
Side	0.00	0.00	0.00	0.01	0.00	0.99	0.00	0.00
Skip	0.00	0.00	0.00	0.00	0.12	0.00	0.87	0.02
Walking	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.97

(i) k-NN CM our dataset

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Boxing	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Running	0.03	0.00	0.00	0.00	0.68	0.04	0.14	0.12
Side	0.03	0.01	0.01	0.01	0.03	0.86	0.03	0.03
Skip	0.00	0.00	0.00	0.00	0.15	0.01	0.76	0.08
Walking	0.01	0.00	0.00	0.00	0.18	0.01	0.09	0.72

(j) DT CM for our dataset

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Boxing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.00	0.92	0.00	0.02	0.06
Side	0.00	0.00	0.00	0.01	0.01	0.97	0.01	0.00
Skip	0.00	0.00	0.00	0.00	0.04	0.00	0.93	0.03
Walking	0.00	0.00	0.00	0.00	0.08	0.01	0.00	0.91

(k) SVM CM for our dataset

Bend	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.04
Boxing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.00	0.70	0.00	0.13	0.18
Side	0.00	0.00	0.00	0.00	0.01	0.97	0.02	0.01
Skip	0.00	0.00	0.00	0.00	0.06	0.00	0.89	0.04
Walking	0.00	0.00	0.00	0.00	0.18	0.02	0.03	0.77

(l) NB CM for our dataset

Gray Image Features

Boxing	1.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	1.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	1.00	0.00	0.00	0.00
Jogging	0.00	0.00	0.00	0.92	0.04	0.04
Running	0.00	0.00	0.00	0.08	0.90	0.01
Walking	0.00	0.00	0.00	0.05	0.01	0.94

(m) k-NN CM for KTH

Boxing	0.96	0.01	0.00	0.01	0.02	0.00
Handclapping	0.01	0.95	0.01	0.01	0.03	0.00
Handwaving	0.01	0.08	0.90	0.01	0.00	0.00
Jogging	0.00	0.01	0.02	0.74	0.11	0.12
Running	0.01	0.01	0.00	0.13	0.78	0.08
Walking	0.00	0.01	0.01	0.10	0.09	0.80

(n) DT CM for KTH

Boxing	1.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	1.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	1.00	0.00	0.00	0.00
Jogging	0.00	0.00	0.00	0.91	0.04	0.05
Running	0.00	0.00	0.00	0.10	0.87	0.03
Walking	0.00	0.00	0.00	0.05	0.03	0.93

(o) SVM CM for KTH

Boxing	1.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.97	0.00	0.01	0.00	0.02
Handwaving	0.00	0.04	0.96	0.00	0.00	0.00
Jogging	0.00	0.01	0.00	0.66	0.17	0.16
Running	0.00	0.01	0.00	0.16	0.81	0.02
Walking	0.00	0.00	0.00	0.09	0.15	0.75

(p) NB CM for KTH

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jack	0.00	0.99	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.04	0.00	0.86	0.00	0.01	0.00	0.04	0.06	0.00	0.00
Pjump	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Run	0.01	0.00	0.02	0.00	0.62	0.01	0.13	0.21	0.00	0.00
Side	0.00	0.00	0.00	0.03	0.01	0.88	0.03	0.04	0.01	0.00
Skip	0.03	0.00	0.12	0.00	0.17	0.00	0.61	0.07	0.00	0.00
Walk	0.00	0.00	0.04	0.00	0.07	0.03	0.07	0.79	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(q) k-NN CM for Weizmann

Bend	0.87	0.05	0.03	0.03	0.00	0.00	0.00	0.00	0.01	0.01
Jack	0.01	0.78	0.00	0.13	0.01	0.01	0.01	0.01	0.01	0.01
Jump	0.04	0.00	0.53	0.06	0.08	0.02	0.16	0.11	0.00	0.00
Pjump	0.04	0.03	0.03	0.80	0.01	0.04	0.04	0.00	0.01	0.01
Run	0.01	0.00	0.02	0.00	0.51	0.06	0.10	0.29	0.00	0.00
Side	0.01	0.06	0.09	0.09	0.12	0.48	0.04	0.09	0.01	0.00
Skip	0.00	0.00	0.19	0.00	0.09	0.03	0.55	0.14	0.00	0.00
Walk	0.00	0.03	0.05	0.00	0.17	0.04	0.07	0.64	0.00	0.00
Wave1	0.01	0.01	0.02	0.07	0.00	0.00	0.00	0.00	0.83	0.06
Wave2	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.95

(r) DT CM for Weizmann

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jack	0.00	0.91	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.01	0.00	0.95	0.00	0.00	0.00	0.02	0.01	0.00	0.00
Pjump	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Run	0.00	0.00	0.02	0.00	0.71	0.01	0.14	0.12	0.00	0.00
Side	0.00	0.01	0.00	0.01	0.04	0.92	0.00	0.01	0.00	0.00
Skip	0.00	0.00	0.04	0.00	0.12	0.00	0.83	0.01	0.00	0.00
Walk	0.00	0.00	0.05	0.00	0.16	0.01	0.00	0.78	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(s) SVM CM for Weizmann

Bend	0.90	0.00	0.01	0.08	0.00	0.01	0.00	0.00	0.00	0.00
Jack	0.00	0.68	0.00	0.13	0.00	0.09	0.01	0.00	0.01	0.08
Jump	0.00	0.00	0.86	0.02	0.02	0.09	0.00	0.00	0.00	0.00
Pjump	0.00	0.00	0.00	0.97	0.00	0.03	0.00	0.00	0.00	0.00
Run	0.01	0.01	0.14	0.01	0.35	0.15	0.10	0.22	0.00	0.00
Side	0.00	0.00	0.03	0.04	0.03	0.88	0.03	0.00	0.00	0.00
Skip	0.00	0.00	0.20	0.00	0.16	0.13	0.38	0.13	0.00	0.00
Walk	0.01	0.01	0.14	0.01	0.23	0.15	0.09	0.36	0.00	0.00
Wave1	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.94	0.05
Wave2	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.99

(t) NB CM for Weizmann

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Boxing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.00	0.58	0.00	0.12	0.30
Side	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Skip	0.00	0.00	0.00	0.00	0.13	0.00	0.77	0.10
Walking	0.00	0.00	0.00	0.00	0.14	0.00	0.07	0.79

(u) k-NN CM for our dataset

Bend	0.96	0.00	0.00	0.00	0.03	0.00	0.02	0.00
Boxing	0.00	0.99	0.01	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Handwaving	0.01	0.00	0.00	0.99	0.00	0.00	0.00	0.00
Running	0.03	0.00	0.00	0.01	0.50	0.04	0.15	0.27
Side	0.00	0.00	0.00	0.03	0.09	0.79	0.01	0.08
Skip	0.02	0.00	0.01	0.00	0.18	0.01	0.66	0.11
Walking	0.02	0.00	0.00	0.01	0.29	0.03	0.14	0.51

(v) DT CM for our dataset

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Boxing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.00	0.70	0.01	0.05	0.25
Side	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.01
Skip	0.00	0.00	0.00	0.00	0.04	0.00	0.92	0.04
Walking	0.00	0.00	0.00	0.00	0.16	0.01	0.03	0.80

(w) SVM CM for our dataset

Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Boxing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Handclapping	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Handwaving	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Running	0.03	0.00	0.00	0.00	0.48	0.02	0.18	0.29
Side	0.02	0.00	0.00	0.01	0.05	0.86	0.00	0.06
Skip	0.02	0.00	0.00	0.00	0.13	0.01	0.67	0.16
Walking	0.04	0.00	0.00	0.00	0.44	0.02	0.07	0.43

(x) NB CM for our dataset

Fig. 9 Confusion matrices according to ML methods

extracted from binary frame can be effective in HAR applications. To prove this, SURF are extracted from both the gray space and the binary images. In addition, three different datasets are used to determine the performance of the binary image features on the images with different contrasts.

When Table 2 is examined, for our data, it is seen that the SURF-based BoVW features in the gray space frame

Table 3 Comparison of our approach with previous works

Previous works	KTH	Weizmann
Baccouche et al. [27]	91.04	–
Rahman et al. [28]	94.67	–
Zhang et al. [29]	91.33	92.89
Singh et al. [30]	–	–
Bian et al. [31]	68.00–78.00	–
Cao and Liu [32]	92.00	99.6
Uddin et al. [33]	–	–
Ding et al. [34]	–	–
Liu et al. [37]	93.40	–
Vo and Ly [41]	95.20	–
Gilbert et al. [42]	94.50	–
Grushin et al. [43]	90.70	–
Jhuang et al. [44]	91.70	–
Kläser [45]	94.20	79.80
Lin et al. [46]	93.43	–
Liu et al. [47]	93.80	–
Liu and Shah [48]	94.16	–
Rodriguez [49]	81.50	–
Schindler and Van Gool [50]	92.70	–
Sun et al. [51]	94.00	–
Veeriah et al. [52]	93.96	–
Wu et al. [53]	95.10	–
Niebles et al. [55]	–	90.60
Ramage et al. [56]	–	97.20
Blank et al. [57]	–	99.63
Scovanner et al. [58]	–	82.60
Bregonzio et al. [59]	–	96.60
Dollár et al. [60]	–	85.20
Klaser et al. [61]	–	84.30
Liu et al. [62]	94.92	–
Moussa et al. [63]	97.89	96.66
Our approach		
Binary image features	95.33	90.24
Gray image features	96.14	91.11

are classified as 92.71% by SVM. The accuracy rate for the binary frame features is 96.52%. The most accurate values for the KTH are obtained using the k-NN as 96.14% and 95.33%, respectively. Similarly, the most accurate values for the Weizmann are obtained using SVM as 91.11% and 90.24%, respectively. When the results of the studies in Table 3 were examined, Cao and Liu [32] and Moussa et al. [63] achieved much better results for KTH and Weizmann. Like this study, Cao and Liu [32] also used BOW paradigm. However, in that study, higher-order uncertainties were encoded with the type 2 fuzzy topic models (T2 FTM) used instead of SVM. This made this work more successful. Moussa et al. [63] utilized the SIFT method with BoVW and performed classification using SVM. The more accurate results were substantially due to the normalization of histograms generated by BoVW.

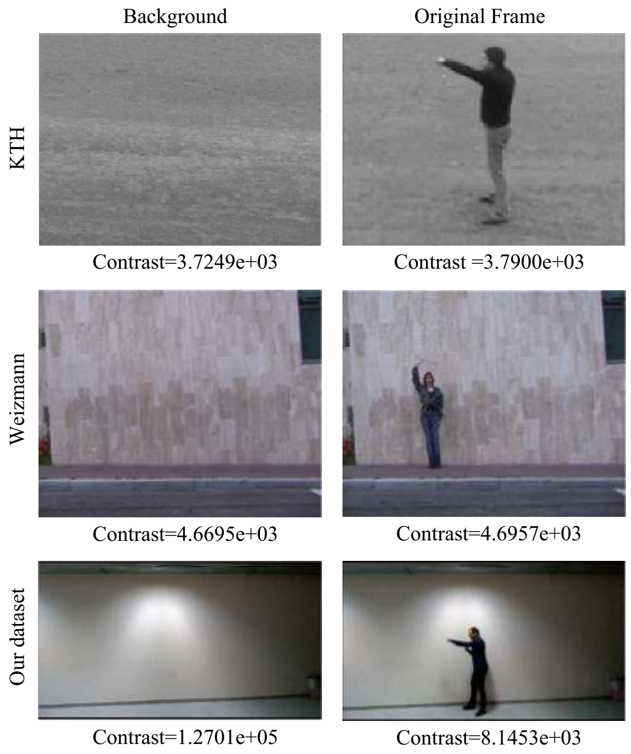


Fig. 10 Contrast change caused by foreground

Considering the training periods, the training time for k-NN, which is the most successful classification according to KTH results, is 5.57 s in binary images and 8.37 s in gray images. For Weizmann, these values were determined as 11.63 s and 46.39 s, respectively, by using SVM. Similarly, for our data, these values are 6.77 and 6.95 s, respectively, using SVM.

When all the results are evaluated, the final result is that the SURF for a binary image are more effective for HAR studies in frames which the foreground (human) reduces the contrast. In addition, training time is lower for binary image.

Acknowledgements The authors are thankful to RAC-LAB (www.rac-lab.com) for providing the trial version of their commercial software for this study.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Dobhal T, Shitole V, Thomas G, Navada G (2015) Human activity recognition using binary motion image and deep learning. Procedia Comput Sci 58:178–185
2. Kim E, Helal S, Cook D (2010) Human activity recognition and pattern discovery. IEEE Pervasive Comput/IEEE Comput Soc IEEE Commun Soc 9(1):48
3. De Kleijn R, Kachergis G, Hommel B (2014) Everyday robotic action: lessons from human action control. Front Neurorobot 8:13
4. Dhamsania CJ, Ratanpara TV (2016) A survey on human action recognition from videos. In: 2016 Online international conference on green engineering and technologies (IC-GET). IEEE, pp 1–5
5. Koohzadi M, Charkari NM (2017) Survey on deep learning methods in human action recognition. IET Comput Vis 11(8):623–632
6. Ngoc LQ, Viet VH, Son TT, Hoang PM (2016) A robust approach for action recognition based on spatio-temporal features in RGB-D sequences. Int J Adv Comput Sci Appl 7(5):166–177
7. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
8. Mandal R, Roy PP, Pal U, Blumenstein M (2018) Bag-of-visual-words for signature-based multi-script document retrieval. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3444-y
9. Tang F, Lim SH, Chang NL, Tao H (2009) A novel feature descriptor invariant to complex brightness changes. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 2631–2638
10. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: European conference on computer vision. Springer, pp 404–417
11. Panchal P, Panchal S, Shah S (2013) A comparison of SIFT and SURF. Int J Innov Res Comput and Commun Eng 1(2):323–327
12. Karami E, Prasad S, Shehata M (2017) Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images. arXiv preprint arXiv:1710.02726
13. Yang J, Jiang Y-G, Hauptmann AG, Ngo C-W (2007) Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the international workshop on multimedia information retrieval. ACM, pp 197–206
14. Faraki M, Palhang M, Sanderson C (2014) Log-Euclidean bag of words for human action recognition. IET Comput Vis 9(3):331–339
15. Dawn DD, Shaikh SH (2016) A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis Comput 32(3):289–306
16. Xu S, Fang T, Li D, Wang S (2010) Object classification of aerial images with bag-of-visual words. IEEE Geosci Remote Sens Lett 7(2):366–370
17. Kim J, Kim B-S, Savarese S (2012) Comparing image classification methods: k-nearest-neighbor and support-vector-machines. Ann Arbor 1001:48109–48122
18. Farid DM, Zhang L, Rahman CM, Hossain MA, Strachan R (2014) Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Syst Appl 41(4):1937–1946
19. Ben-Hur A, Weston J (2010) A user's guide to support vector machines. In: Data mining techniques for the life sciences. Springer, pp 223–239
20. Abellán J, Castellano JG (2017) Improving the Naive Bayes classifier via a quick variable selection method using maximum of entropy. Entropy 19(6):247
21. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(Feb):281–305
22. Yao Y, Cao J, Ma Z (2018) A cost-effective deadline-constrained scheduling strategy for a hyperparameter optimization workflow for machine learning algorithms. In: International conference on service-oriented computing. Springer, pp 870–878
23. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th international conference on pattern recognition, 2004 ICPR 2004, vol. 3. IEEE, pp 32–36
24. Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space–time shapes. In: Proceedings of international conference computer Vision. IEEE, pp 1395–1402
25. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31(8):651–666
26. Plötz T, Guan Y (2018) Deep learning for human activity recognition in mobile computing. Computer 51(5):50–59
27. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2011) Sequential deep learning for human action recognition. In: International workshop on human behavior understanding. Springer, pp 29–39
28. Rahman S, Cho S-Y, Leung M (2012) Recognising human actions by analysing negative spaces. IET Comput Vis 6(3):197–213
29. Zhang Z, Hu Y, Chan S, Chia L-T (2008) Motion context: a new representation for human action recognition. In: European conference on computer vision. Springer, pp 817–829
30. Singh M, Basu A, Mandal MK (2008) Human activity recognition based on silhouette directionality. IEEE Trans Circuits Syst Video Technol 18(9):1280–1292
31. Bian W, Tao D, Rui Y (2012) Cross-domain human action recognition. IEEE Trans Syst Man Cybern Part B (Cybern) 42(2):298–307
32. Cao X-Q, Liu Z-Q (2015) Type-2 fuzzy topic models for human action recognition. IEEE Trans Fuzzy Syst 23(5):1581–1593
33. Uddin MZ, Kim T-S, Kim J-T (2013) A spatiotemporal robust approach for human activity recognition. Int J Adv Robot Syst 10(11):391
34. Ding W, Liu K, Cheng F, Shi H, Zhang B (2015) Skeleton-based human action recognition with profile hidden Markov models. In: CCF Chinese conference on computer vision. Springer, pp 12–21
35. Gao H, Chen W, Dou L (2015) Image classification based on support vector machine and the fusion of complementary features. arXiv preprint arXiv:1511.01706
36. Halima NB, Hosam O (2016) Bag of words based surveillance system using support vector machines. Int J Secur Appl 10(4):331–346
37. Liu A-A, Su Y, Gao Z, Hao T, Yang Z-X, Zhang Z (2013) Partwise bag-of-words-based multi-task learning for human action recognition. Electron Lett 49(13):803–805
38. Liu A-A, Xu N, Su Y-T, Lin H, Hao T, Yang Z-X (2015) Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing 151:544–553
39. Liu Y, Fung K-C, Ding W, Guo H, Qu T, Xiao C (2018) Novel smart waste sorting system based on image processing algorithms: SURF-BoW and multi-class SVM. Comput Inf Sci 11(3):35
40. Zhu Y, Nayak NM, Roy-Chowdhury AK (2013) Context-aware activity recognition and anomaly detection in video. J Sel Top Signal Process 7(1):91–101
41. Vo V, Ly N (2012) Robust human action recognition using improved BOW and hybrid features. In: 2012 IEEE International symposium on signal processing and information technology (ISSPIT). IEEE, pp 000224–000229
42. Gilbert A, Illingworth J, Bowden R (2009) Fast realistic multi-action recognition using mined dense spatio-temporal features. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 925–931
43. Grushin A, Monner DD, Reggia JA, Mishra A (2013) Robust human action recognition via long short-term memory. In: The

2013 international joint conference on, neural networks (IJCNN). IEEE, pp 1–8

44. Jhuang H, Serre T, Wolf L, Poggio T (2007) A biologically inspired system for action recognition. In: IEEE 11th international conference on computer vision, 2007 ICCV 2007. IEEE, pp 1–8
45. Kläser A (2010) Learning human actions in video. Ph.D. Thesis, Université de Grenoble
46. Lin Z, Jiang Z, Davis LS (2009) Recognizing actions by shape-motion prototype trees. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 444–451
47. Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos “in the wild”. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE, pp 1996–2003
48. Liu J, Shah M (2008) Learning human actions via information maximization. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–8
49. Rodriguez M (2010) Spatio-temporal maximum average correlation height templates in action recognition and video summarization. Electronic Theses and Dissertations, 4323
50. Schindler K, Van Gool L (2008) Action snippets: How many frames does human action recognition require? In: IEEE conference on computer vision and pattern recognition CVPR 2008. IEEE, pp 1–8
51. Sun X, Chen M, Hauptmann A (2009) Action recognition via local descriptors and holistic features. In: IEEE computer society conference on computer vision and pattern recognition workshops, 2009 CVPR workshops 2009. IEEE, pp 58–65
52. Veeriah V, Zhuang N, Qi G-J (2015) Differential recurrent neural networks for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 4041–4049
53. Wu X, Liang W, Jia Y (2009) Incremental discriminative-analysis of canonical correlations for action recognition. In: 2009 IEEE 12th international conference on computer vision, 2009. IEEE, pp 2035–2041
54. Suto J, Oniga S, Lung C, Orha I (2018) Comparison of offline and real-time human activity recognition results using machine learning techniques. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3437-x
55. Niebles JC, Wang H, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis 79(3):299–318
56. Ramage D, Hall D, Nallapati R, Manning CD (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing: volume 1. Association for Computational Linguistics, pp 248–256
57. Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space–time shapes. In: Tenth IEEE international conference on computer vision (ICCV’05). IEEE, pp 1395–1402
58. Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on multimedia. ACM, pp 357–360
59. Bregonzio M, Xiang T, Gong S (2012) Fusing appearance and distribution information of interest points for action recognition. Pattern Recognit 45(3):1220–1234
60. Dollár P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance. IEEE, pp 65–72
61. Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-gradients. In: BMVC 2008 19th British machine vision conference. British Machine Vision Association, pp 275: 1–10
62. Liu H, Ju Z, Ji X, Chan CS, Khoury M (2017) Study of human action recognition based on improved spatio-temporal features. In: Human Motion sensing and recognition: a fuzzy qualitative approach. Springer, Berlin, pp 233–250
63. Moussa MM, Hamayed E, Fayek MB, El Nemr HA (2015) An enhanced method for human action recognition. J Adv Res 6(2):163–169
64. Singh YK, Singh ND (2017) Binary face image recognition using logistic regression and neural network. In: 2017 International conference on energy, communication, data analytics and soft computing (ICECDS). IEEE, pp 3883–3888
65. Pandey RK, Vignesh K, Ramakrishnan A (2018) Binary Document image super resolution for improved readability and OCR performance. arXiv preprint arXiv:1812.02475
66. Perner P, Perner H, Müller B (2002) Mining knowledge for HEp-2 cell image classification. Artif Intel Med 26(1–2):161–173
67. Santofimia MJ, Martinez-del-Rincon J, Nebel J-C (2014) Episodic reasoning for vision-based human action recognition. Sci World J 2014:270171
68. Laptev I, Lindeberg T (2006) Local descriptors for spatio-temporal recognition. In: Spatial coherence for visual motion analysis. Springer, pp 91–103
69. Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F (2011) An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recognit 44(8):1761–1776
70. Haralick RM (1979) Statistical and structural approaches to texture. Proc IEEE 67(5):786–804

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