Human Activity Recognition based on Deep Belief Network Classifier and Combination of Local and Global Features

Azar Mahmoodzadeh
Department of Electrical Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.
mahmoodzadeh@iaushiraz.ac.ir

Received: 11/Feb/2020 Revised: 18/Aug/2020 Accepted: 18/Jan/2021

Abstract

During the past decades, recognition of human activities has attracted the attention of numerous researches due to its outstanding applications including smart houses, health-care and monitoring the private and public places. Applying to the video frames, this paper proposes a hybrid method which combines the features extracted from the images using the ‘scale-invariant features transform’ (SIFT), ‘histogram of oriented gradient’ (HOG) and ‘global invariant features transform’ (GIST) descriptors and classifies the activities by means of the deep belief network (DBN). First, in order to avoid ineffective features, a pre-processing course is performed on any image in the dataset. Then, the mentioned descriptors extract several features from the image. Due to the problems of working with a large number of features, a small and distinguishing feature set is produced using the bag of words (BoW) technique. Finally, these reduced features are given to a deep belief network in order to recognize the human activities. Comparing the simulation results of the proposed approach with some other existing methods applied to the standard PASCAL VOC Challenge 2010 database with nine different activities demonstrates an improvement in the accuracy, precision and recall measures (reaching 96.39%, 85.77% and 86.72% respectively) for the approach of this work with respect to the other compared ones in the human activity recognition.

Keywords*:* BoW; DBN; GIST; HOG; Human Activity Recognition; SIFT.

1- Introduction

As the diversity of the applications of supervisory and security systems grows, the need for smart algorithms which are able to detect activities and behaviors of the people is intensified. Progresses in data collecting and analysis technologies have led to wide usage of the human activity recognition (HAR) systems in the daily life. Applications such as security and surveillance, crowd management, content-based image retrieval, action retrieval in images, user interface design, human-computer interaction, robot learning, sport images analysis and eHealth have raised the attention of the researchers to propose various methods for recognizing the human activities [1, 2].

Based on the design methodology for data acquisitioning process, the HAR systems are mainly categorized into visual, non-visual, and multimodal sensor technologies. One of the most popular approaches for identifying the human activities is utilizing the visual sensors (cameras) and applying the pattern recognition techniques to the images. The most important difference between the cameras and other sensors is the method of perceiving information from the environment. While most sensors generate information as one-dimensional signals, in cameras the information is received as a set of two-dimensional signals (i.e., visible images). The applications of HAR systems based on visual approach are categorized to the following groups: (1) daily life and smart houses, (2) healthcare monitoring systems, (3) surveillance and security in public environments, and (4) sports and public outdoor applications [1]. Despite the great progresses in pattern classification, the human activity recognition in static images is still considered a big challenge. In this regard, several issues such as images with different and complicated backgrounds, high volume of data, images from different views, low intra-class similarity (doing a single action in different ways by different people), low inter-class variability (e.g., the similarity between drinking and eating) and lack of temporal information have led to difficulties in the human activity recognition using static images [1]. Several publications in this regard are reported in Section 2.

In general, the human activity recognition in static images includes four steps, as follow: (1) pre-processing: applying a set of operations to the images with the aim of image enhancement and reduction of noise and redundancy; (2) feature extraction: computing and finding effective and distinctive features; (3) feature reduction: decreasing the number of features by keeping or producing the most

*Corresponding Author

discriminative ones. This step moderates the computational load; (4) classification: this is the most important step of every machine learning system which identifies the human activities [1]. For this purpose, we extract some stable and useful features of the images using a combination of GIST, HOG and SIFT descriptors. A descriptor is a representation of an image that simplifies it by extracting useful information and throwing away unimportant information. Typically, a feature descriptor converts an image to a feature vector. Although the feature vector is not useful for the purpose of viewing an image, it is appropriate for tasks like image recognition and object detection. GIST is a global feature extraction algorithm which is used to detect the scenes and provide precise prediction of the activities in scenes. HOG algorithm works based on the image gradient. This algorithm is able to accurately detect the image edges and extract the features. The SIFT algorithm extracts the features from the image which are robust against image scale changes, rotation and lightening. The two latter algorithms are classified as the local feature extractors. Using the combination of global and local approaches enhances the recognition rate for classifying the human activities, as the results of this paper demonstrates.

Since the lengths of the feature vectors of all the methods are high, the bag of word (BoW) technique is used to map the high-dimension feature space to a lower-dimension one. Recently, the application of deep learning techniques in pattern recognition problems has been widely increased. Besides several advantages of deep learning methods, they suffer from two major drawbacks: (i) overfitting and (ii), much time-consumption. One of the robust and fast deep learning techniques is the deep belief network (DBN). This network consists of the Restricted Boltzmann Machines (RBMs) for training and a back-propagation neural network for tuning the network weights. Therefore, DBN is a good classifier for the problem of the human activity recognition with several classes. In this paper, all the extracted features are concatenated and a deep belief network (DBN) is applied to 1classify and detect the activity type.

The paper is organized as follows. Section 2 reviews the related literature and research methods. In section 3, the framework of the proposed algorithm is described in details. Particularly, the pre-processing, the feature extraction and reduction and classification by means of the deep belief network is presented. In section 4, the simulation results of applying the proposed approach and those of two other well-known methods are reported and compared. Finally, the conclusions are provided in section 5.

2- Related Work

The method proposed by Wang et al. in 2006 was one basic work in the static image-based HAR field [3]. In [4], a real-time algorithm was used for the human activity identification. In the feature extraction step, the algorithms of ‘scale-invariant features transform’ (SIFT) and ‘histogram of oriented gradient’ (HOG) were used. Indeed, the human skeleton was divided into five parts and the geometric configuration of these parts are determined. Finally, the Markov hidden model and the support vector machine (SVM) were used to classify the activities. In [5], HAR was investigated using the ‘global invariant features transform’ (GIST) algorithm. In that paper, the geometrical relations between the human body parts were used for the recognition. Finally, they addressed the images classification using the SVM algorithm. In [6], the human activities were identified using the convolutional neural network. The proposed method was appropriate for streaming video images, since the extracted features were taken from the images based on individual motions.

In [7] two algorithms of ‘speeded up robust features’ (SURF) and HOG were used to extract the images features. Then the authors made use of the SVM for the classification. That paper considered only five human activities. In [8], the recognition and classification of the activities using several body-worn sensory methods were proposed. In that work, the recognition system operates based on which sensor is activated. In 2008, Ikizler et al. [9] addressed the HAR in static images by presenting a rectangular area with oriented spatial histogram. They used linear discriminant analysis and a binary SVM to categorize the activities. In that paper, the human state was extracted using the SIFT algorithm and a SVM classifier. Li and Fei [10], classified the activities in the static images by combining the scenes and objects. They realized that combining the high-level signs may improve the recognition accuracy. The results of Thurau and Hlavac [11] showed that by combining the characteristics of objects and people states, the recognition rate increases. In 2011, Li et al. [12] studied the activities and behaviors in static images in the web. In [13], the human state was estimated using the HOG algorithm and the image scene model and features were obtained using the SIFT algorithm based on the bag of features method. Delaitre et al., [14] studied the activities recognition of the new dataset using bag of features method and combined them with SVM in static images.

Zheng et al., [15] addressed the human action recognition by extracting the features using the gradient-oriented histogram descriptor and two classifiers (one based on the Poselet and the other one based on the content-based learning). Sner et al., [16] proposed a multi signs-based method to recognize human activities in static images. Sharma et al., [17] proposed an expanded part model which is a strong distinctive descriptor of human detection. In [18] a poselet-based method was presented where poselet activation vectors obtain the pose of a person. In [19] proposed a method which learns a set of sparse features and part bases for HAR in still images. A human-centric method that identifies the location of humans and objects associated with an action is proposed in [20]. Khan et al. [21] evaluated the color descriptors and color-shape fusion methods for HAR. Moreover, in [22] they proposed some pose-normalized semantic pyramids employing body part detectors. In [23], the authors encoded multi-scale information during the image encoding stage.

3- Proposed Approach

The block diagram of the proposed approach for the human activity recognition based on the static images is shown in Fig. 1. First, a set of pre-processing operations are carried out to improve the quality of the image and prepare them for the next steps. Then some features are extracted from the image using three local and global descriptors. Following this, the bag of words technique is applied to decrease the dimension of each feature vector. In the next step, the reduced feature vectors are concatenated to form a single vector. Finally, these features are given to the classifier to predict the human activity type. In order to make comparison, the SVM and the artificial neural network (ANN) are also considered in the classification step besides the DBN.

3-1- Pre-processing

In the first step, sizes of all training and testing images are equalized to 64×64. Also, since none of the feature extraction algorithms use the color features, the color images are converted to grey ones. This conversion leads to a reduction in the number of computation operations for each image. Then, the images are passed through a low-pass filter to remove the noise. This task improves the feature extraction algorithms of the next step by avoiding the production of artifacts and wrong keypoints. Finally, edge sharpening filter is applied on the image to enhance the contrast of the edges. The enhanced image is sent to the next step for extracting features.

3-2- Feature Extraction

In this step, the features required for the classification step are computed using the HOG, GIST, and SIFT algorithms. In the following, a brief description of each algorithm is addressed.

HOG descriptor: Histogram of oriented gradient (HOG) is a local feature descriptor which is used in this paper to extract useful features from images containing the human activities. The HOG descriptor uses the distribution of directions of gradients as features. The reason for applying this descriptor is that local information of the image components can be represented by the intensity gradients or the path of the edges. This algorithm computes the gradients in local regions of an image. The general representation for computing the HOG descriptor is shown in Fig. 2. First, the image is divided into a grid of 8×8 cells. Then using the Sobel operator, the gradient magnitude and direction are calculated for every pixel in each cell. Following this, the histogram of the gradient orientations in each cell is computed with 8 bins. In fact, bin of histogram is defined based on the quantized directions and the votes (the values that go into the bins) are selected based on the magnitudes. The votes in each bin are added up to produce the 8-bin histogram for every cell. Since the gradients of an image are sensitive to overall lighting, they are normalized to become robust against lighting variations. For this purpose, a 16-cell window is used to form big-size 2×2 blocks. Notice that sliding this window by one cell constructs the neighbor block with two overlapping cells with the current block. Thus, each block has 4 histograms with 8 bins which can be concatenated to form a vector of length 4*8=32 and then it is normalized using the L2 norm. The final feature vector of HOG descriptor for the entire image is produced by concatenating all the 32×1 vectors into one giant vector. The size of this vector is (8-1)*(8-1) *32=1568 [24]. HOG has two important advantages in the application of human activity recognition. The first advantage is robustness against the lightning variations thanks to computing the gradient directions from the difference of the local intensities. The second advantage is robustness against deformation which is due to the shift and partial affine deformation. This property leads to ignorable changes in the histogram values.

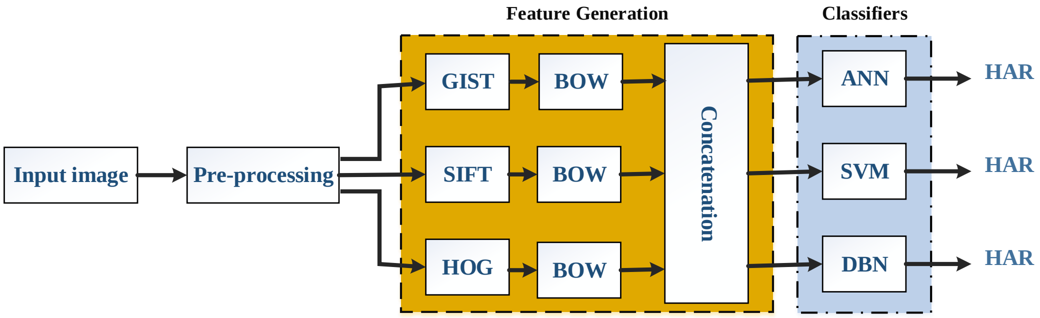

Fig. 1 The block diagram of the proposed algorithm for human activity recognition.

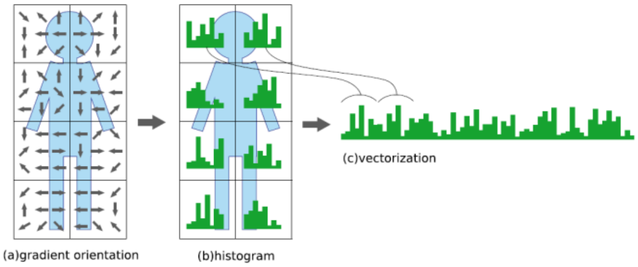

Fig .2 The general diagram of computing the HOG feature vector [24].

GIST descriptor: Global invariant features transform (GIST) descriptor, proposed by Oliva and Torralba in 2001 [25, 26], provides a small-size representation that contains sufficient information for recognizing the human activity in an image. This global descriptor represents the dominant spatial structure of an activity by analyzing the spatial frequency and orientation. The GIST algorithm was developed based on a phenomenon called spatial envelope which enables the algorithm to precisely predict the image scenes. In fact, the GIST is produced by combining the outputs of a number of Gabor filters at different scales and directions. Gabor filter is a linear filter which is a strong tool for analyzing the texture due to its multi-resolution property in the frequency and special domains. Given an image, the GIST descriptor is computed by applying 32 Gabor filters at four scales and eight directions to the image. Accordingly, 4*8=32 feature maps with the resolution same as the original image are constructed. Then, using an 8×8 grid, each feature map is divided into 64 cells and the average of values within each block is computed. Following this, for each feature map, a vector of length 64 containing this gradient information (averages) is generated. By concatenating the feature vectors of all feature maps, the GIST vector with length of 32*64=2048 is achieved [5].

SIFT descriptor: Scale-invariant feature transform (SIFT), introduced by Lowe in 2004 [27], is one the feature-based adaptive algorithms for pattern recognition in images. Extracting the features in the SIFT algorithm is done by an identifier. The feature extraction phase includes three steps: (*i*) extracting the scale space extremums, (*ii*) improving the location accuracy and removing unstable extremums, and (*iii*) allocating a direction to each feature. In the first step, to extract the scale space extremums, the stable features in different scales are extracted using a *scale space*. The scale space presents the image structures in different scales. This space is composed of a set of Gaussian images and difference of Gaussian (DoG) images in different scales which are sorted in different layers called *Octaves*. The scale space Gaussian images for the image $I(x,y)$ using the Gaussian kernel function $G(x,y,\sigma)= \exp(-(x^2+y^2)/2\sigma^2)/2\pi\sigma^2$ is calculated as follows [28]:

$$L(x,y,\sigma)=G(x,y,\sigma)\otimes I(x,y) \qquad (1)$$

Where σ indicates the scale of each image and its initial value is $\sigma_0=1/6$. This scale value is increased using a constant multiplier parameter k in different Octave levels. DoG images are computed as follows using the difference between two adjacent Gaussian images [28]:

$$D(x,y,\sigma)=L(x,y,k\sigma)-L(x,y,\sigma) \qquad (2)$$

The smaller scale is considered as the scale of the DoG image. After generating each Octave, the Gaussian image is scaled down to half using the resampling approach. Then, the resulted image is considered as the initial image of the next Octave; and this process is repeated. The aim of constructing the scale space is extracting features which are independent of the scale. Therefore, to extract the stable situations in DoG images, the intensity of each pixel in every Octave is compared with its eight neighboring pixels in a 3×3 window and nine pixels in neighboring upper and lower DOG images. If it is an extremum (maximum and minimum) comparing to those 26 pixels, it is stored as a candidate feature. Then, for each extracted feature based on the scale of the DOG image, the scale parameter is chosen. In the second step, to eliminate the unstable extremums, features with low contrast and those on the edges are removed. Also by interpolating the adjacent points, the exact location of each keypoint is determined. In the last step of feature extraction, a direction is dedicated to each stable keypoint [28].

Once the keypoints are extracted, the next phase is to generate the features descriptor. To make them robust against the scale and rotation variations, descriptors are made according to the scale and direction of each feature. In addition, the descriptor is designed so that it is robust against lightning variations and those caused by imaging viewpoint. To do this in the standard SIFT algorithm, first a square block around every feature in the related Gaussian image is considered. The dimensions of this block are selected according to the scale of the feature so that every bin is a square with a side equal to three times of the scale. Then the coordination of the block is rotated to get aligned the main direction of the feature. Following this, the values and directions of the gradients of the pixels within the rotated region are calculated and the gradients directions are also rotated with the main direction of that feature. Then, a Gaussian function with the scale equal to the half width of the block is used to weighting the values of the gradients. Next for each bin in the block, a histogram of weighed gradient directions of the pixels within the bin is constructed. To prevent the effects of boundaries between the bins, a tri-liner interpolation is performed to distribute the gradient values in the histogram. Finally, a SIFT descriptor is produced as a vector with 128 components. In this descriptor, the amplitudes of the components are normalized in order to reduce the lightening variation effects. After this step, a threshold value of 0.2 is considered for the values of the descriptor to reduce the effects of angle variations of the imaging. Then, the normalizing process of the descriptor is repeated.

3-3- Feature Reduction and Final Concatenation

After applying each feature extraction algorithm, a long-length feature vector is produced for the image. Finding the useful features is considered an important topic in the human activity recognition; since with smaller number of features, the computational load is decreased. To reduce the dimension of the feature space, in this paper the BoW technique is applied to the features extracted by the descriptors. The recent studies showed that the BoW method presents a set of discriminative and robust features comparing other methods which use the texture or intensity [30, 31].

To generate a BoW model, all features of different images in different classes are collected in a set. These features are clustered using the *k*-Mean algorithm. The centers of the clusters represent the code-words and their union produces a code-book. For every input image, each feature vector is dedicated to one of the centers of the clusters using the nearest neighbor method. Then, a histogram is made for the image wherein the horizontal axis is the centers of the clusters and the vertical axis is the number of the features which are dedicated to each of the centers. Finally, this histogram is considered as the *new feature vector* generated for that feature extraction algorithm [30, 31]. Once the new feature vectors of the three descriptors are found, they are concatenated to generate the final feature vector, to be given to the classifier.

3-4- Deep Belief Network

In this paper the deep belief network (DBN), as one of the most important deep learning models, is used to model the human states. This network is a fast learning algorithm which can find the optimal responses with high speed. The learning model is composed of two steps: (*i*) pre-training and (*ii*) fine-tuning; see Fig. 3. The pre-training step is done using the restricted Boltzmann machine (RBM) which is a generative model. The RBM is a type of the Boltzmann machine in which the connections between the visible and hidden units are disconnected. The unsupervised pre-training system works effectively in solving classification problems with numerous data and high diversity [32]. In the second step, the network weights are precisely tuned using a supervised algorithm. For this step, a back-propagation neural network is used. A DBN can be trained by repeatedly maximizing the conditional probability of input vectors or observable vectors. By doing this, the hidden vectors and a specific set of layer weights are obtained. As the RBM is considered the heart of the deep belief network, this machine is briefly introduced.

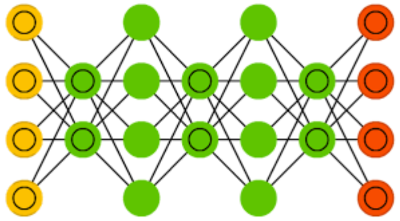

Fig. 3 Representation of the graph of a DBN.

Boltzmann machine is a special type of the Markov random field (MRF), which is represented by a symmetrical network with binary random units. This model has a set of D visible units $\boldsymbol{v} = \{0,1\}^D$ and a set of F hidden units $\boldsymbol{h} = \{0,1\}^F$. The common energy of these units in the Boltzmann machine is defined as follows [31]:

$$E(\boldsymbol{v},\boldsymbol{h}) = -\boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{v}^T W\boldsymbol{h} \qquad (3)$$

Where the components a_i and b_i - which from the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ - are the bias terms for the hidden and visible units, respectively. These vectors are updated in each iteration through some recursive formulas. The parameters of these formulas are the main parameters of the RBM in a DBN. Moreover, each weighting element W_{ij} in the matrix W indicates the symmetrical transactional term between the visible unit i and the hidden unit j. These weights are the parameters of the model. This network devotes a probability value to every possible pair of hidden and visible vectors in the energy function. The resulted probability distribution is defined as follows [31]:

$$P(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{Z}\exp\left(-E(\boldsymbol{v},\boldsymbol{h})\right),\ Z\ @\sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}) \qquad (4)$$

The value of Z is recognized as a normalizing constant. The probability which a network dedicates to a training data can be increased by tuning the weights and bias in order to reach the lower energy. By defining $g(x)$ as a logistic sigmoid function, the conditional probability of the visible vector $\boldsymbol{v}$ and the hidden vector $\boldsymbol{h}$ can be obtained as follows [31].

$$P(v_i = 1|\boldsymbol{h}) = g\left(\sum_{j=1}^{F} W_{ij} h_j + b_i\right) \qquad (5)$$

$$P(h_j = 1|\boldsymbol{v}) = g\left(\sum_{i=1}^{D} W_{ij} v_i + a_j\right) \qquad (6)$$

$$g(x)\ @\frac{1}{1+\exp(-x)} \qquad (7)$$

When the modes of the hidden units are selected, the input data can be reproduced by putting every v_i equal to '1' according to (6). Then the modes of the hidden units are updated so that they show the reproduced features. To find the optimal weights in the matrix W , the contrastive divergence (CD) learning method [32] is applied. The variations of the weights are defined according to the following relation [32].

$$\Delta w_{ij} = \varepsilon\left(v_i h_j{}_{\text{data}} - v_i h_j{}_{\text{reconstruction}}\right) \qquad (8)$$

Where ε is the learning rate. The strong point of the restricted Boltzmann machines is learning with the aim of reconstruction. During the reconstruction process, this machine only uses the information of the hidden units as the feature of the learned input. If the model is able to retrieve the main input well, this means that the weights and bias are trained correctly. Because of the advantages of the RBM, in the recent years this model has been widely used in constructing DBNs. Also numerous papers are presented with the aim of improving this model and increasing its efficiency.

A RBM with a simple hidden layer is not adequate to find the features of a data. Greedy layer-wise training methodology is an efficient tool to improve the system accuracy. In this method, the first machine mapped the input data from the zero-layer to the first-layer. After training the machine, the trained features (output of the first-machine) are used as the inputs for the second RBM. The features of the last machine are considered as the learned features of the whole training process. This type of the layered learning system can be used to construct the DBNs [34, 35]. Then, the network is able to discover the deep features of the data. In fact, this network learns the deep features of the input by a pre-training process in a hierarchical process. Logistic regression (LR) layer is added to the end of the learning system as the second stage of the DBN. This classifier is used to tune the previously trained network so that the classification is performed using the learned features. The accurate tuning process is implemented using a back-propagation algorithm. The target of this algorithm is to find the minimum in the peripheral area of parameters which have already been determined by the DBN [35, 36].

4- Simulation Results

To evaluate the proposed algorithm, the PASCAL VOC Challenge 2010 database was used which includes 225 images and nine human activities [37]. The selected images have different sizes and they are color or gray-scale. For every activity, eighteen images are randomly selected for the training phase (about 70% of all available images) and the seven remaining ones are kept for the testing phase. Therefore, the set of the training and testing images include 162 and 63 images, respectively. Fig. 4 shows some of the sample images.

The proposed algorithm is run under the MATLAB R2014a programming environment on a PC equipped with 3.2 GHZ CPU and 8 GB RAM memory. To evaluate the performance of the proposed algorithm, three measures of precision (Pre.), recall (Rec.) and accuracy (Acc.) were used which are formulated in (9)-(11), respectively:

$$\text{Pre}. = \frac{TP}{TP+FP} \qquad (9)$$

Fig. 4 Some sample images of the dataset.

$$\text{Rec}. = \frac{TP}{TP+FN} \qquad (10)$$

$$Acc. = \frac{TP+TN}{TP+FN+TN+FP} \qquad (11)$$

Consider a two-class problem, called 'P' and 'N'. Then, TP is the number of truly detecting the class P ; also FP is the number of falsely detecting the class N as P . Similarly, TN is the number of truly detecting the class N ; also FN is the number of falsely detecting the class P as N . Table 1 shows the recall, precision and accuracy of the proposed algorithm for combining different features for nine human activities. In the first and second cases, the features extracted by SIFT only and HOG only were used. In the third case, the features extracted by combination of SIFT and HOG algorithms were used and in the fourth case the features extracted by all three algorithms of SIFT, HOG and GIST were applied. Notice that before concatenating the feature vectors, the BoW algorithm is applied to each descriptor using the *k*-Mean algorithm with *k*=20. Moreover, the parameters of the DBN are the learning rate ε=0.5, number of hidden layers=1, number of the hidden units *n*=10, momentum ϕ=0.006 and the weight decay λ=0.4.

According to the results obtained, the first case had the lowest precision compared to the others. Also, the performance of the HOG features was better than that of the SIFT. By applying these two feature extraction methods in the third case, the precision values in the activities such as 'Playing instrument', 'Riding horse' and 'Using computer' were higher than the other cases. In the fourth case, the precisions of the HAR system for activities such as 'Phoning', 'Reading', 'walking' and 'Taking photo' were increased compared with the third case. The average of the total precision in nine activities for the fourth case compared with the second and the third cases increased 6% and 2.5%, respectively. Because of the inherit complexity and the nonlinear behavior of the images and the features extraction methods, increasing the number of the features or combining the diverse features not only necessarily improve the system performance but also, in some cases, increase the redundancy. Therefore, in some cases of the Table, the non-homogeneous behaviors (improvement and reduction of the efficiency index) were reported.

In addition to the precision, the accuracy of the proposed algorithm was investigated which indicated the correct detection of the algorithm. While the precision indicates the proximity of the repeated measurements to each other, accuracy is the proximity of a measurement to the actual value. The latter measure indicates that in the worst case, a measuring set to what extent is near to the real value. A correct system is not necessarily precise and vice versa. A system has appropriate performance if both the precision and accuracy are simultaneously high. From Table 1 it is inferred that the proposed algorithm generally has higher accuracy in case four than the other cases. Additionally, adding HOG to the SIFT improved the efficiency of the SIFT feature solely. Generally, it can be seen that all four cases have high accuracies. Nevertheless, for 'Phoning' the accuracy in the third case (HOG & SIFT) was lower than that of the second case (HOG alone).

By comparing the recall measure in Table 1 for the fourth case, it can be seen that for some activities such as 'running', the value of this measure is high and for some other activities such as 'Reading' and 'Taking photo' this value is low. Furthermore, using the HOG feature in the 'Riding horse' activity leads to a higher recall value compared to the SIFT and 'SIFT & HOG'. Therefore, the recall value depends on the activity type. Generally, given the results shown in Table 1 it can be seen that although the efficiency of the SIFT alone is lower than the other features in terms of the recall index, using this feature along with the HOG and GIST can improve the efficiency. The average recall value when all the features are used is enhanced 21%, 5.5% and 2% compared with (*i*) SIFT alone, (*ii*) HOG alone and (*iii*) HOG and SIFT, respectively.

To evaluate the proposed algorithm, the performance results of this method is compared with the results obtained from two well-known methods, i.e., artificial neural network (ANN) and multi-class support vector machine (SVM). The ANN is a multi-layer perceptron (MLP) with 20 neurons in one hidden layer. Also for the optimized SVM, the penalty parameter (*C*) is 3.5384 and the kernel parameter (σ) is 0.5147. The average accuracy results and the training and testing times for each of the four feature extraction techniques given to the SVM, ANN and DBN classifiers are reported in Table 2. Comparing according to the accuracy measure, Table 2 validates that the proposed method outperforms the SVM and ANN, in all the feature extraction methods. Nevertheless, using deep learning method in DBN leads to higher computational complexities and larger run times. Also divided results for each activity are shown in Fig. 5. According to this figure, the accuracy of the HAR based on the ANN and the SVM has reached the 74.42% and 91.80%, respectively. Meanwhile, the proposed method (based on the DBN) has achieved the accuracy of 96.39% which is higher than two other methods. It is worth nothing that all three above-mentioned classifiers were applied to the combined features set (SIFT& HOG & GIST).

Table 3 compares the proposed approach with some state-of-the-art methods for human action recognition. The approach of [21] attains a precision of 62.4%, while that of [22] achieves 63.5%. The method of [19] based on learning a sparse basis of features and parts obtains a precision of 65.1%. The approach of this work yields consistent improvement over the state-of-the-art methods with a precision of 85.8%.

5- Conclusion

The main goal of this paper is to develop a robust human activity recognizer based on the images' data. Using images for the application of HAR is feasible thanks to high-quality yet not-much expensive cameras. Easy installation and standard communication protocols make these cameras suitable for a variety of daily life uses. Thus, a novel approach was proposed in this work for the HAR application. First some pre-processing operations were performed on the images in order to enhance their quality and make them prepared for the next steps. Then, multiple robust features including SIFT, HOG and GIST were extracted followed by the BoW technique for feature reduction. Finally, the robust features were entered to a DBN for activity training and recognition on query images. The proposed method was compared with traditional multiclass SVM and ANN approaches where it showed the superiority of our technique. The HAR system is evaluated for nine different physical activities where it achieved a mean recognition rate of 96.39%. On the contrary, the SVM and ANN approaches obtained mean recognition rates of 91.80% and 74.42%, respectively. For the next work, we plan to use other robust features and learning approaches to achieve more efficient activity recognition systems for real-time applications in complex environments.

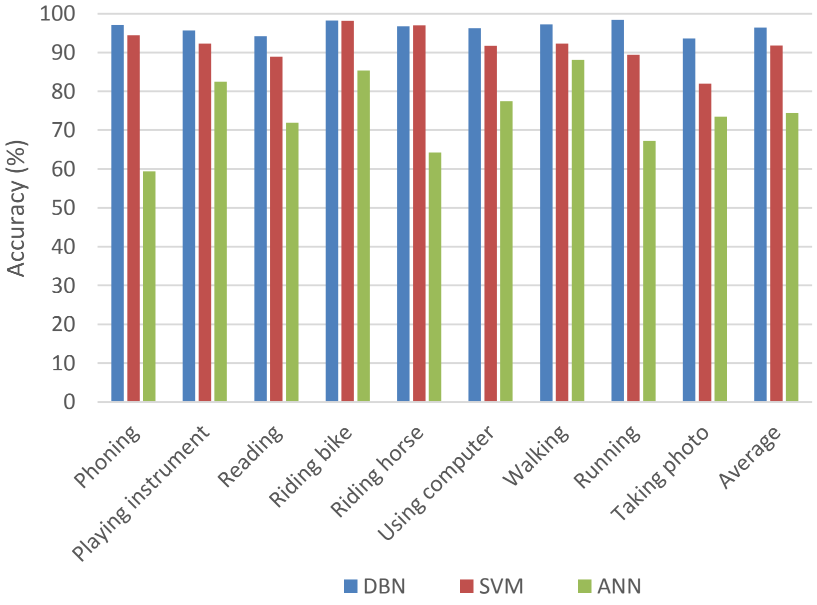

Fig. 5 The performance results of DBN, SVM and ANN methods in terms of the recognition accuracy.

Table 1: Comparing the performance of different feature extraction methods for nine human activities. The following abbreviations are used: ‘Pre.’: Precision, ‘Rec.’: Recall, ‘Acc.’: Accuracy, ‘Ave.’: Average.

	SIFT			**HOG**			**SIFT & HOG**			**SIFT & HOG & GIST**		
Activity	**Pre.**	**Rec.**	**Acc.**	**Pre.**	**Rec.**	**Acc.**	**Pre.**	**Rec.**	**Acc.**	**Pre.**	**Rec.**	**Acc.**
Phoning	52.14	80.93	92.88	73.85	92.85	96.22	72.00	93.49	96.09	80.28	95.03	97.11
Playing instrument	79.85	62.43	91.69	76.71	72.23	93.79	86.71	78.21	95.50	82.42	81.21	95.67
Reading	65.42	53.79	89.36	69.71	65.90	92.23	74.57	68.24	93.00	88.00	71.37	94.23
Riding bike	75.71	94.24	96.50	89.57	85.80	96.50	93.00	93.36	98.15	92.28	94.56	98.22
Riding horse	68.85	55.20	90.07	89.71	86.33	96.85	91.57	89.04	97.42	89.71	85.63	96.77
Using computer	58.85	59.22	90.44	71.71	89.76	95.60	75.85	92.81	96.30	70.85	97.34	96.27
Walking	56.14	57.11	90.19	83.28	85.22	96.14	86.42	88.30	96.87	92.00	87.33	97.22
Running	57.71	73.12	92.38	91.14	97.88	98.50	90.00	97.73	98.39	87.85	99.98	98.43
Taking photo	56.14	52.91	89.30	79.14	59.07	91.20	84.28	61.18	91.88	88.57	68.02	93.66
Ave.	**63.42**	**65.47**	**91.42**	**80.53**	**81.59**	**95.23**	**83.28**	**84.70**	**95.96**	**85.77**	**86.72**	**96.39**

Table 2: Comparing the average accuracy results and training and testing times for four feature extraction techniques given to the SVM, ANN and DBN.

Classifier	Measure	**SIFT**	**HOG**	**SIFT & HOG**	**SIFT & HOG & GIST**
SVM	Ave. ACC.	88.73	90.54	90.88	92.23
	Training Time (s)	0.287	0.523	0.774	1.188
	Testing Time (s)	0.035	0.071	0.101	0.151
ANN	Ave. ACC.	72.94	75.26	75.59	76.38
	Training Time (s)	0.136	0.172	0.592	0.981
	Testing Time (s)	0.029	0.047	0.089	0.134
DBN	Ave. ACC.	91.42	95.23	95.96	96.39
	Training Time (s)	2.835	3.418	6.172	8.529
	Testing Time (s)	0.081	0.094	1.053	1.105

Table 3: Comparison with the state-of-the-art results according to the precision measure. ‘PrM.’ stands for the proposed method.

	Phoning	*Playing instrument*	*Reading*	*Riding bike*	*Riding horse*	*Using computer*	*Walking*	*Running*	*Taking photo*	***Ave.***
Poselet[18]	49.6	43.2	27.7	83.7	89.4	59.1	67.9	85.6	31.0	*59.7*
IaC [13]	45.5	54.5	31.7	75.2	88.1	64.1	62.0	76.9	32.9	*59.0*
POI [14]	48.6	53.1	28.6	80.1	90.7	56.1	69.6	85.8	33.5	*60.7*
LAP [19]	42.8	60.8	41.5	80.2	90.6	66.1	74.4	87.8	41.4	*65.1*
WPOI [20]	55.0	81.0	69.0	71.0	90.0	50.0	44.0	59.0	36.0	*62.0*
CF [21]	52.1	52.0	34.1	81.5	90.3	59.9	66.5	88.1	37.3	*62.4*
SM-SP[22]	52.2	55.3	35.4	81.4	91.2	59.6	68.7	89.3	38.6	*63.5*
BDF [23]	64.3	94.5	65.1	96.9	96.8	87.7	78.9	93.4	77.1	*83.7*
PrM.	79.2	82.4	88.0	92.2	89.7	70.8	92.0	87.8	88.5	*85.7*

References

[1] S. Ranasinghe, F. Al Machot, and H.C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," International Journal of Distributed Sensor Networks, vol. 12, no. 8, p. 1550147716665520, 2016.

[2] S.S. Agaian, J. Tang, and J. Tan, "Electronic imaging applications in mobile healthcare," 2019.

[3] Y. Wang, H. Jiang, M.S. Drew, Z.N. Li, and G. Mori, "Unsupervised discovery of action classes," in Proceedings of CVPR, pp. 17-22.

[4] S. Yan, J.S. Smith, W. Lu, and B. Zhang, "Multibranch Attention Networks for Action Recognition in Still Images," IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 4, pp. 1116-1125, 2017.

[5] Y. Wang, Y. Li, X. Ji, "Human action recognition based on global gist feature and local patch coding," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 2, pp. 235-246, 2015.

[6] E. Park, X. Han, T.L. Berg, and A.C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1-8, 2016.

[7] H.A. Qazi, U. Jahangir, B.M. Yousuf, and A. Noor, "Human action recognition using SIFT and HOG method," in 2017 International Conference on Information and Communication Technologies (ICICT), pp. 6-10, 2017.

[8] H.F. Nweke, Y.W. Teh, G. Mujtaba, and M. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," Information Fusion, vol. 46, pp. 147-170, 2019.

[9] N. Ikizler, R.G. Cinbis, S. Pehlivan, and P. Duygulu, "Recognizing actions from still images," in 2008 19th International Conference on Pattern Recognition, pp. 1-4, 2008.

[10] L.J. Li, and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," In 2007 IEEE 11th international conference on computer vision, pp. 1-8, 2007.

[11] C .Thurau and V. Hlavác, "Pose primitive based human action recognition in videos or still images," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[12] P. Li, J. Ma, and S. Gao, "Actions in still web images: visualization ,detection and retrieval," in International Conference on Web-Age Information Management, pp. 302-313, 2011.

[13] N. Shapovalova, W. Gong, M. Pedersoli, F.X. Roca, and J. Gonzalez, "On importance of interactions and context in human action recognition ",in Iberian conference on pattern recognition and image analysis, pp. 58-66, 2011.

[14] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in Advances in neural information processing system, pp. 1503-1511, 2011.

[15] Y. Zheng, Y.J. Zhang, X. Li, and B.D. Liu, "Action recognition in still images using a combination of human pose and context information," in 2012 19th IEEE International Conference on Image Processing, pp. 785-788, 2012.

[16] F. Sener, C. Bas, and N. Ikizler-Cinbis, "On recognizing actions in still images via multiple features," in European Conference on Computer Vision, 2012, pp. 263-272.

[17] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3506-3513, 2012.

[18] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in CVPR 2011, pp. 3177-3184, 2011.

[19] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in 2011 International Conference on Computer Vision, pp. 1331-1338, 2011.

[20] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 601-614, 2011.

[21] F.S. Khan, R.M. Anwer, J. Van De Weijer, A.D. Bagdanov, and M. Felsberg, "Coloring action recognition in still images," International journal of computer vision, vol. 105, no. 3, pp. 205-221, 2013.

[22] F.S. Khan, J. Van De Weijer, R.M. Anwer, M. Felsberg, and C. Gatta, "Semantic pyramids for gender and action recognition," IEEE Transactions on Image Processing, vol. 23, no. 8, pp. 3633-3645, 2014.

[23] F.S. Khan, J. Van De Weijer, R.M. Anwer, A.D. Bagdanov, M. Felsberg, and J. Laaksonen, "Scale coding bag of deep features for human attribute and action recognition," Machine Vision and Applications, vol. 29, no. 1, pp. 55-71, 2018.

[24] T. Watanabe, S. Ito, and K. Yokoi, "Co-occurrence histograms of oriented gradients for pedestrian detection," in Pacific-Rim Symposium on Image and Video Technology, pp. 37-47, 2009.

[25] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of computer vision, vol. 42, no. 3, pp. 145-175, 2001.

[26] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," Progress in brain research, vol. 155, pp. 23-36, 2006.

[27] G. Lowe, "SIFT-The Scale Invariant Feature Transform," Int. J, vol. 2, pp. 91-110, 2004.

[28] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, pp. 91-110, 2004.

[29] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in null, p. 1470, 2003.

[30] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 524-531, 2005.

[31] M.A. Carreira-Perpinan and G.E. Hinton, "On contrastive divergence learning," in Aistats, pp. 33-40, 2005.

[32] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural computation, vol. 14, no.8, pp. 1771-1800, 2002.

[33] N. Le Roux, and Y. Bengio, "Deep belief networks are compact universal approximators," Neural computation, vol. 22, no. 8, pp. 2192-2207, 2010.

[34] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in Artificial Intelligence and Statistics, pp. 448-455, 2009.

[35] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in Neural Networks for Perception, ed: Elsevier, pp. 65-93, 1992.

[36] I. Sutskever and G.E. Hinton, "Deep, narrow sigmoid belief networks are universal approximators," Neural computation, vol. 20, no. 11, pp. 2629-2636, 2008.

[37] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International journal of computer vision, vol. 88, no. 2, pp. 303-338, 2010.

Azar Mahmoodzadeh received B.Sc., M.Sc. and Ph.D. degrees in Electrical Engineering from University of Shiraz, University of Shahed and University of Yazd, Iran, in 2005, 2008 and 2013, respectively. From 2009, she was with the Islamic Azad University, Shiraz Branch, Shiraz, Iran. Her research interests include pattern recognition and image and signal processing.