IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 18, 2020, accepted May 23, 2020, date of publication May 27, 2020, date of current version June 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997962

A Novel Action Recognition Framework Based on Deep-Learning and Genetic Algorithms

ABDULLAH ASIM YILMAZ, MEHMET SERDAR GUZEL, ERKAN BOSTANCI, AND IMAN ASKERZADE

Computer Engineering Department, Ankara University, 06830 Ankara, Turkey

Corresponding author: Erkan Bostanci (ebostanci@ankara.edu.tr)

ABSTRACT Recognition of human actions in partially cluttered environments is an important research field of computer vision and human–computer interaction. This field has recently garnered attention from a large number of academic researchers in various fields of application. This study proposes a novel deep-learning-based architecture for the recognition and prediction of human actions based on a hybrid model. The main contribution of this study is to propose a new hybrid architecture, integrating four wide-ranging pre-trained network models in an optimized manner, using a metaheuristic algorithm. This architecture consists of four main stages: namely, the creation of the data set, the design of deep neural network (DNN) architecture, training and optimization of the proposed DNN architecture, and evaluation of the trained DNN. By adapting the aforementioned architecture, reliable features are obtained for the training procedure. In order to validate the superiority of the proposed architecture over other state-of-the-art studies, a performance evaluation between these architectures is presented using benchmark datasets. The results reveal that the proposed architecture outperforms previously developed architectures in terms of predicting human actions.

INDEX TERMS Human action recognition, DNNs, transfer learning, deep learning, optimization, genetic algorithm.

I. INTRODUCTION

Human action recognition (HAR) is an important area of the computer vision and human–computer interaction fields; it is considered to be one of the most popular research fields. Additionally, HAR plays an important role in allowing society to gain in-depth knowledge of human activities from video or sensor inputs (such as 3D laser range finders, etc.) in their daily lives [1]. Application areas include, among others, the healthcare industry [2], video surveillance [3], driving safety [4], gesture recognition [5], video magnification [6], and smart home technology [7]; these are examples of application areas where human action recognition systems are extensively employed by researchers and engineers.

An action is basically defined as a sequence of movements of the human body. From a computer vision perspective, an action recognition procedure involves matching observations obtained from input values (e.g., sensor data, etc.) with a predefined pattern and then assigning a label based on the action type [8]. The main intention of a human action recognition procedure is to deliver automatic analysis of ongoing actions obtained from a video stream or image sequence. If it is essential to detail the purpose of recognizing human actions, two analytical approaches can be studied: specifically, simple and general cases. This can be elaborated as follows: In simple cases, the system basically monitors human activities and recognizes actions so as to classify them based on their activity category. However, for general cases, it is essential to ensure the incessant recognition of human activities by individually detecting the duration of all activities which occur in a video [9].

The recognition of human actions is classified into two groups in the literature, depending on whether the features are produced by traditional handcrafted-based (conventional-based) techniques or through the output of deep neural networks (DNN). In recent years, deep-learning-based human action recognition methods [10], [11] have outpaced traditional handcrafted-based human action recognition methods and achieved substantial progress in recognizing human action from video sequences [12]. However, training a new model of deep-learning-based human action recognition from scratch entails large amounts of data, plentiful operational

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Aljawarneh.

resources, and sometimes days for training processes. To overcome these challenges, it would be more appropriate to use a pre-trained model on a new training process [13], [14]. The structure involving this approach is called transfer learning [15].

This study proposes a novel deep-learning-based human activity recognition system in which the input data, consisting of video or sensor inputs, are first supplied to the system. After the video acquisition part is completed, the low- and high-level features are extracted by the convolution layers of the proposed hybrid architecture. Afterwards, a metaheuristic approach is applied to enhance the extracted features for better recognition performance. Once those steps are completed, the system is trained in a supervised manner. One of the critical contributions of this study is its proposition of a new hybrid layer that involves four pre-trained models instead of one model: namely, AlexNet [16], VGG-16 [17], GoogleNet [18], and ResNet-152 [19]. Overall, several comprehensive deep-learning models, relying on a transfer-learning method, are combined so as to produce a hybrid model. Besides a metaheuristic approach, a genetic algorithm (GA) is applied to approximate optimum feature selection for learning-based problems.

HAR is a crucial and challenging task used in a variety of application fields, including facial gesture analysis, human–computer interaction, video surveillance, etc. With the development of deep learning (DL), considerable progress has been achieved in computer vision and image understanding. Accordingly, several different pre-trained architectures using DL are proposed to overcome different problems of computer vision and machine learning. The challenge of the HAR problem and *the* limitation of the current techniques have mainly motivated the authors to develop novel hybrid architectures, involving a combination of different pre-trained DL architectures to obtain a better overall accuracy. A problem-specific metaheuristic algorithm for feature optimization process has also been designed with respect to the proposed architecture. This is mainly responsible for selecting qualified features for a better and shorter training process. It should be stated that the ultimate aim of this study is to design a general HAR architecture which is able to categorize lots of different human actions, from walking to riding a horse. The details will be defined in the methodology section.

The architecture is realized in four main stages. These are data set, design of the DNN architecture, training and optimization of the DNN architecture, and evaluation of the trained DNN. Overall, the rest of the paper is organized as follows. Section 2 presents the corresponding literature. The proposed framework and experimental results are discussed in Sections 3 and 4, respectively. The final section concludes the study.

II. LITERATURE REVIEW

In recent years, different solutions have been proposed which are associated with the recognition of human actions; these have been put forward due to the high demand in numerous application areas and their importance in the design of human–computer interaction systems. Examples of different solutions include frameworks which are skeleton based, semantic based, knowledge based, and so on. For example, in a related study, Chaudhry *et al.* [20] aimed to encrypt skeleton sequences to spatial–temporal hierarchical models and, next, apply linear dynamical systems to learn the dynamic structures; furthermore, they mainly aimed to model 3D representations of human motion. Vemulapalli *et al.* [21] also studied skeletal representation by considering relationships between different body parts based on translation and rotation operations in three-dimensional space. Additionally, this study characterized each action as a curve in a Lie group and utilized support vector machines (SVMs) to classify human actions. Alternatively, Culmone *et al.* [22] proposed a semantic-based system that implements a methodology based on the integration of an ontology to represent contextual knowledge and a rule-based system. This framework provides an infrastructure in which knowledge, organized in conceptual spaces, can be semantically discovered, shared, and queried across some applications. Salguero *et al.* [23] also proposed an ontology-based data mining framework for the recognition of the daily activities of humans. This framework is based on combining the entities in the ontology and tries to find the expressions which describe daily living activities. Zeng and Ji [24] proposed solutions based on dynamic Bayesian network models with domain knowledge for human action recognition. This study presents generic dynamic Bayesian network models which combine multiple features for human activity recognition, and also a framework to learn the deep belief network (DBN) model which relates training data with domain knowledge. With this approach, the authors claim to prevent the problem of a lack of adequate training data in the human action recognition process.

Human action recognition systems can be classified into two groups according to the type of data they exploit as input to represent features: namely, video-based and sensor-based action recognition systems. While video-based human action recognition systems involve human movements obtained from single or multiple cameras, sensor-based human action recognition systems rely on different sensors such as a magnetometer, thermometer, accelerometer, gyroscope, Bluetooth, smartphone, and so on [1], [9].

Because the proposed system involves both video and sensor data, it falls into the category of both action recognition systems. Essentially, the methodology of the system consists of four main steps, illustrated in Fig. 1. The first step, data acquisition, involves gathering video or sensor data. The second step involves the extraction of features by employing pre-trained networks. The third step aims to optimize feature selection operations by using a metaheuristic approach, GA. Finally, the training phase of DNNs is carried out by considering a supervised learning methodology. There is a huge amount of literature on human recognition systems. Accordingly, this section only includes the corresponding

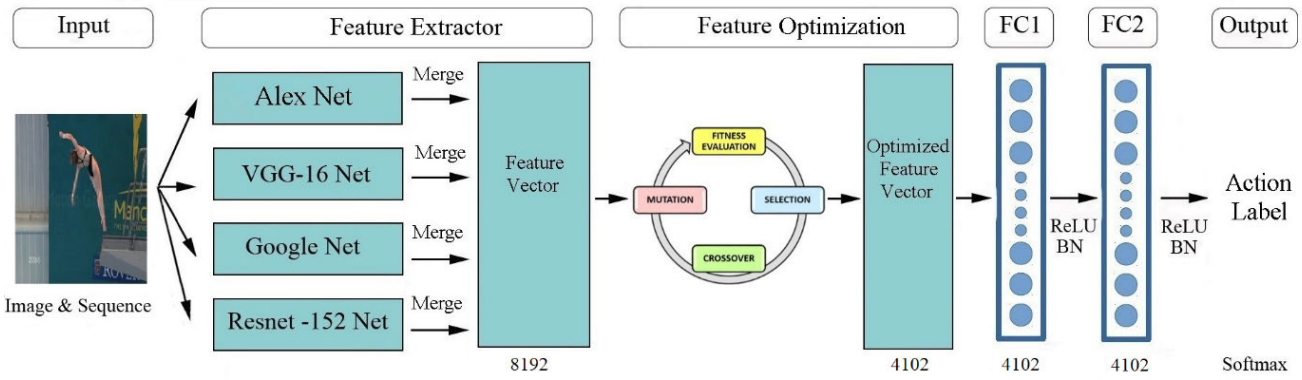

FIGURE 1. The general pipeline of proposed human action recognition system (The input picture frame is taken from [38]).

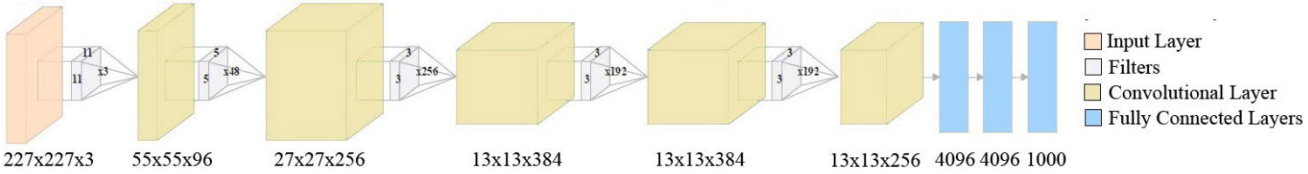

FIGURE 2. The basic architecture of AlexNet the network [16].

literature employed in the design of the proposed architecture. Consequently, only corresponding convolutional neural network (CNN) architectures and the optimization approach, specifically, GA-based studies, are considered within this section. Overall, four popular CNN architectures and the corresponding metaheuristic studies regarding GA are represented in a detailed manner.

AlexNet is one of the leading CNNs which was introduced in the "ImageNet Large Scale Visual Recognition Challenge." The AlexNet network, illustrated in Fig. 2, has a modest architecture and consists of the eight following layers. This network, in essence, consists of three "pooling," two "normalization," and seven "ReLU" layers, following the convolution layers. Regarding the learning and classification process, the architecture involves fully connected layers and a Softmax layer, as shown in Figure 2. The details of the corresponding architecture can be seen in [16].

The VGG-16 network, on the other hand, is also a CNN architecture, trained by more than 1.3 million images obtained from the ImageNet database [25]. ImageNet, one of the leading datasets in DL, is preferred for use in the pre-training part of the model because of its popularity and data diversity. The main purpose of this network model is to reveal the relation between the depth of CNNs and the overall accuracy rate for large-scale image recognition datasets. The model achieves significant improvement in the accuracy rate by pushing the depth to 16 layers. The architecture of the VGG-16 network, illustrated in Fig. 3, offers a homogeneous deep-learning architecture with 3×3 convolutions and 2×2 centering operations from start to finish. In addition to these, there exist fully connected layers followed by a Softmax layer for the classification problem [17]. Here, a Softmax layer is utilized to provide a multi-class learning process in which a set of features can be related to one of K classes. The mathematical formulation of the Softmax layer is shown in (1). Here, the equation, $\sigma(z)$ represents the Softmax result, whereas z represents a vector of the inputs to the output layer and j indicates the output units.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \quad j = 1, 2, \ldots, K \tag{1}$$

As opposed to the previous models, the GoogleNet Network was the first architecture to employ the concept of width in CNN models. The basic architecture of this model is given in Fig. 4. The most important feature of this architecture is that it has a module called "inception" which consists of a shortcut branch and few deeper branches. This module allows the width item in the model to be obtained. Essentially, the expansion effect of the inception modules is attained by employing 1×1, 3×3, and 5×5 filters in the convolution layers and 3×3 maximum pooling operation in a parallel manner [18]. The inception module involving dimensionality reduction is presented in Fig. 5. Here, the GoogleNet network, in essence, has a depth of 27 layers, including pooling layers, and has 7 million process parameters. The architecture is elaborated in Fig. 4 and 5 and consists of four main sections. These are stem, stacked, inception auxiliary classifiers and output classifier modules. The corresponding architecture has nine inception modules and 144 layers such as the input layer, output layer, convolution, maxpooling, ReLU, Softmax, and fully-connected layers [18].

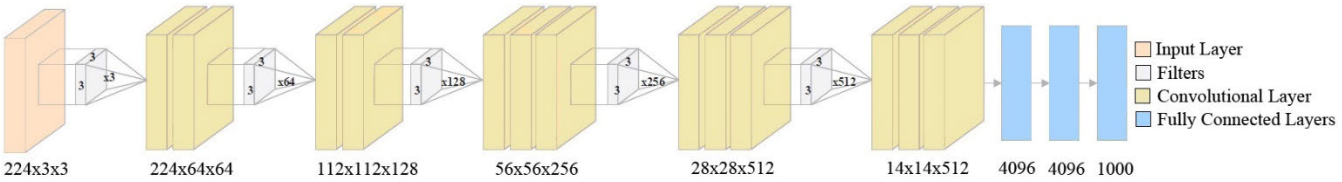

FIGURE 3. The basic architecture of the VGG-16 network [17].

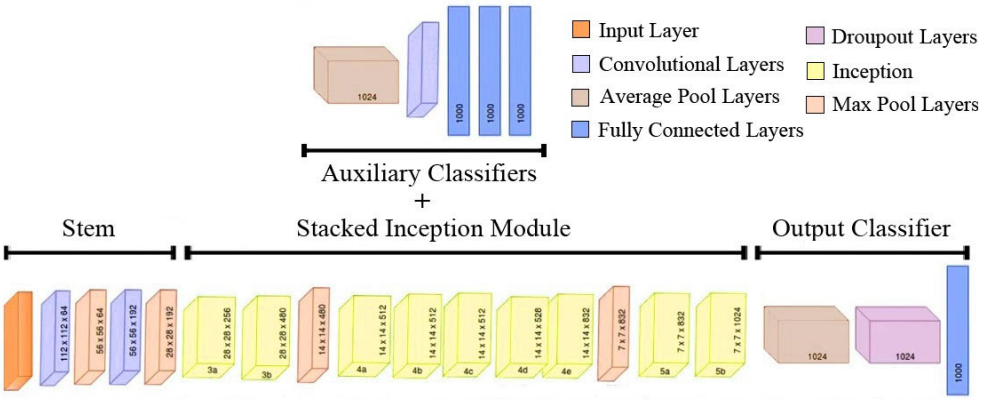

FIGURE 4. The basic architecture of the GoogleNet network [18].

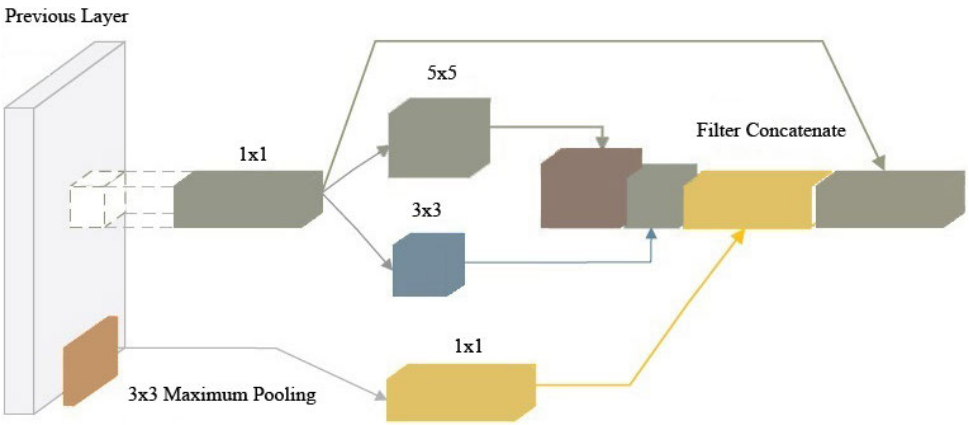

FIGURE 5. Inception module with dimension reductions [18].

The ResNet-152 network allows the residual learning system to facilitate the training of deeper networks that are problematic to train. In this network model, a DNN architecture was designed to learn the residual functions by referring to the input layers rather than learning the unreferenced functions. The architecture of the ResNet-152 network, shown in Fig. 6, is based on the VGG-16 network. Details of the VGG-16 network are illustrated in Fig. 3. Within this network, there are two pooling operations, convolution layers with 3×3 and 1×1 filters, a fully connected layer, and a Softmax layer. In addition, the novelty of the model arises with its residual blocks and the depth of its architecture. In traditional deep-learning models, the stacked layers conform to a desired basic mapping, whereas the ResNet model allows these layers to fit a residual mapping [19].

GA—the first and best known evolutionary computational algorithm developed by John Holland—is a search method that tries to find the best solution in a solution space based on the principle of survival of the best [26]–[28]. GAs are said to have been based on Charles Darwin's words, "It is not the strongest of the species that survives, nor the most intelligent, but the one who is most adaptable to change," and they aim to solve optimization problems.

GAs first start working with an initial population of randomly generated individuals, each representing a solution. Subsequently, a new population is obtained by intercrossing

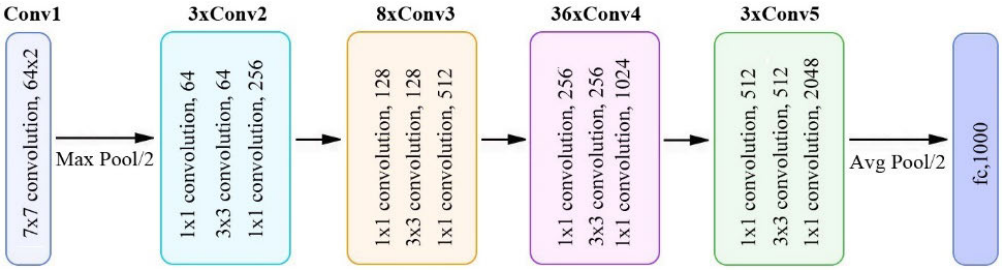

FIGURE 6. The basic architecture of the ResNet-152 architecture [19].

or mutating individuals selected from the current generation in each iteration process. Afterwards, this new generation replaces the previous generation. It is hoped that a population of individuals with better genetic characteristics will be formed within each new generation. Within the processes herein, the crossover and mutation operations are accomplished according to the probability of a crossover and mutation [29].

As an example of a solution to optimization problems using GAs, Minaei-Bidgoli and Punch [30] proposed a method for classifying students so as to forecast their final grade using features extracted from previously stored data in a database system. In this method, a GA was employed to optimize a group of classifiers. Here, a GA was employed to weight the feature vectors in an optimized manner. It should be noted that this approach provides a strong enhancement over raw classification. Handels *et al.* [31] offered a new approach to computer supported analysis of skin tumors in dermatology. In this work, high resolution skin surface profiles were investigated to automatically recognize ''nevocytic,'' ''nevi,'' and ''malignant'' melanomas. Here, feature selection was defined as an optimization problem, and genetic and greedy algorithms were employed for feature selection in order to enhance the classification performance of the recognition system. By using the optimized recognition system, classification is achieved to almost 98% accuracy. However, there are methods to solve optimization problems other than GAs. For example, Karim *et al.* [32] proposed a novel, deep auto-encoder (DA) based architecture which essentially aims to optimize DAs to process data. This deep-learning architecture integrates the Taguchi method into a DA-based architecture to optimize parameters. Results reveal that this architecture succeeds in optimizing multiple parameters. This essentially serves to extract more measurable and valuable data from a few experimental trials. Kumaran *et al.* [33] proposed a hybrid CNN–GWO (Grey Wolf Optimization) approach in order to recognize human actions in videos acquired from unconstrained environments. The goal of this approach is to increase the classification accuracy by training a CNN classifier with local and global exploration capabilities of ''gradient descent'' and ''GWO'' algorithms. Here, the GWO algorithm was employed to find the optimal weight initializations. In this work, the proposed approach yielded better results with 99.9% recognition accuracy compared with the conventional approaches. Karim *et al.* [34] proposed a novel framework for medical data processing which employs a DA and energy spectral density (ESD). The novelty of this work is the ability to feed selected features using the ESD function into a unique deep sparse auto-encoder model. This situation provides the proposed architecture to be optimized and allows production of better features in a time-efficient manner when compared with traditional methods. With this proposed deep sparse auto-encoder architecture, the overall accuracy was also found to be higher.

III. PROPOSED METHOD

The main aim of this study is to propose an optimized framework for human action recognition by designing a hybrid DNN architecture. The general architecture is illustrated in Fig. 1. The structure of the framework consists of four main stages. These steps are the creation of the data set, design of the DNN architecture, the training and optimization phase, and, finally, the evaluation stage. Fig. 11 in the system flowchart in Appendix illustrates a more detailed definition of those four steps in three sections. The first section, called ''pre-training'', is when the creation of the dataset was first performed, using comprehensive datasets containing video-based or sensor-based data, such as the Skoda, UCF sports action, UCI Smartphone, and KTH datasets. Sample images obtained from those human datasets are illustrated in Fig. 7. The details of these human action recognition datasets are described in the experimental results section.

After obtaining the dataset, the design of the DNN architecture was performed. Here, two critical preprocessing steps were made for the design of DNN architecture: The first step was to estimate an appropriate deep-learning architecture to be used in the human action recognition processes. Preliminary experiments revealed that a hash module can produce a better overall sensitivity. Therefore, a hybrid module was generated using pre-trained architectures. The hybrid module includes AlexNet, VGG-16, GoogleNet, and ResNet-152 architectures. The details of these DNN architectures are described in Section 1. Later, it is intended to explore the transfer learning approaches based on selected deep-learning architectures in order to design the proposed DNN architecture in the second preprocessing step. It should be

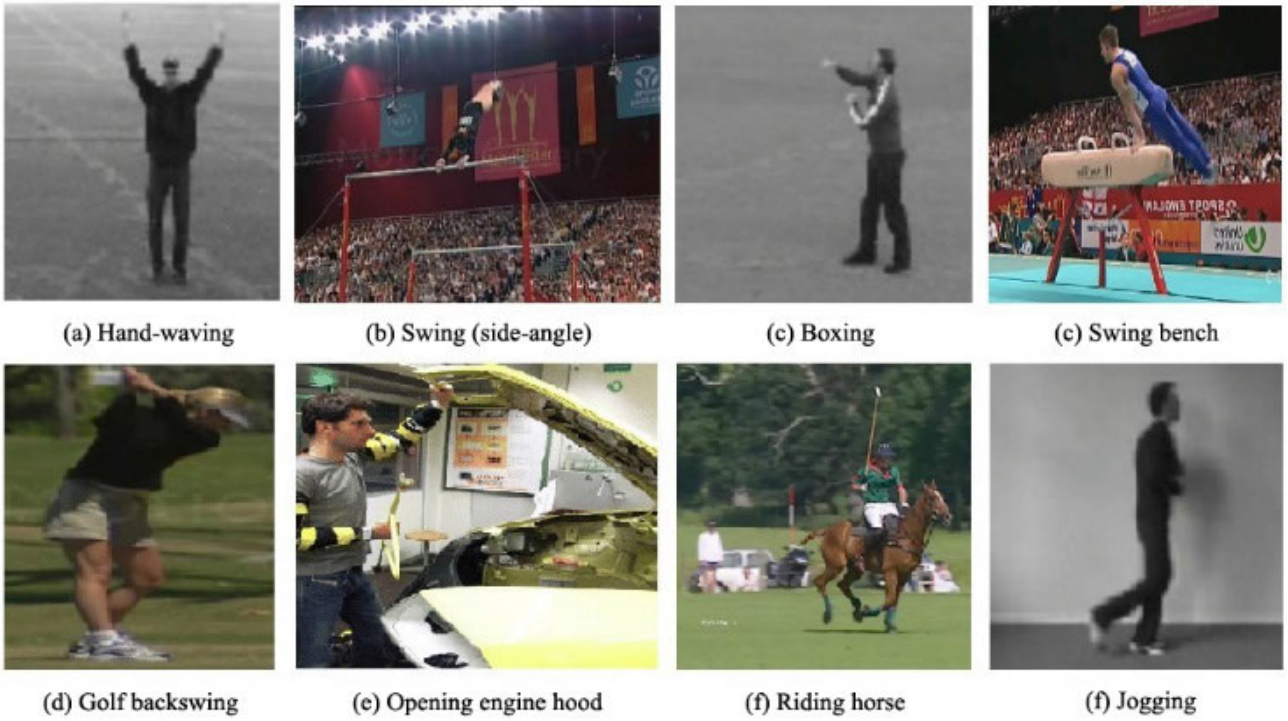

FIGURE 7. Sample images from human action recognition datasets [35]–[38].

noted that it may not be possible to obtain millions of datasets or process a large set of data in each application to achieve classification for use in a variety of processes or models due to various circumstances. Model complexity, excessive dataset size, hardware resource constraints, time constraints, and the like are examples of these circumstances. In such cases it may be more appropriate to adopt the transfer learning approach using pre-trained models. Two approaches are basically used in the transfer learning approaches. The first approach is the processing of weight updates using the new training dataset while maintaining the original pre-trained network. The second approach is the use of the pre-trained network as a feature extractor, followed by performing the classification process with a general classifier. Due to the circumstances mentioned, the latter transfer learning approach was adapted for the suggested architecture. The low- and high-level features of images obtained from different architectures were combined in the proposed architecture. The flow chart (Appendix) depicts the structure of the proposed DNN architecture. The fully connected layers of the proposed architecture consist of 4102 nodes and the Softmax layer has 66 outputs, referring to categories of 66 motions. According to this figure, the pre-trained networks shown in the pre-training section serve as feature extractors for general image features. Additionally, the layers shown in the feature optimization section show the feature optimization process performed by the GA. The tailing three layers represent fully connected layers for the operation and the Softmax classifier for the classification operation. The details of the features produced by the selected pre-trained deep-learning architectures are summarized as follows:

- Alexnet Network: On each picture frame, a 2048-dimensional feature is extracted from the final fully connected layer as shown in Figure 2.
- VGG-16 Network: On each picture frame, a 2048-dimensional feature is extracted from the final fully connected layer as shown in Figure 3.
- GoogleNet Network: On each picture frame, a 2048-dimensional feature is extracted from the final fully connected layer as shown in Figure 4.
- Resnet152 Network: On each picture frame, a 2048-dimensional feature is extracted from the final fully connected layer as shown in Figure 6.

This hybrid model is essentially a new pre-trained architecture which relies on transfer learning. The proposed model fundamentally merges four pre-trained networks by applying an equal weighting procedure so as to generate a dense feature vector, followed by an optimization process based on GA so as to both select the best features and obtain a sparse feature vector to decrease the training process. Each step is given as follows: Firstly, the pre-training operation was performed. Here, the selected deep-learning architectures were trained with an ImageNet data set [35]. Secondly, the features acquired from four deep-learning architectures were merged to form an 8192-dimensional feature vector. Once the merged feature vector with 8192 elements was obtained

without encountering any error or data loss, the pre-training phase of the system was completed (See 11).

The combined feature vector was optimized using a GA whose details are mentioned in the next paragraph and, as a result of this process, an optimized feature vector with 4102 dimensions was obtained. Afterwards, the combined feature vector was passed into the fully connected layers and the Softmax layer to achieve normalization. This layer aims to increase the learning capability of the proposed network (see Section 2 of Fig. 11). Finally, in the evaluation section, experimental analysis and validation were carried out by employing the comprehensive datasets as inputs to the trained model. The combined feature vector was optimized by using GA in order to allow the proposed architecture perform more efficiently during the action classification operations. In order to obtain qualified features with lower dimensions, preliminary tests were conducted: those tests reveal that a feature vector reduced to 4102 dimensions is enough to both obtain high accuracy and training performance. The flowchart of the feature optimization process performed is shown in the feature optimization section of Fig. 11 (see Appendix). In the feature optimization process in which the GA was performed first, an initial population was formed with a certain amount of individuals: each individual consisted of 8192 genes, corresponding to a feature vector of 8192 elements. Afterwards, a random feature selection process was performed from the individuals obtained. At this stage, each individual was used to obtain feature vectors. After the attribute selection process was performed, the classification process was carried out using the feature vector with the machine learning model (SVM and radial basis function (RBF)), details of which are mentioned in the following paragraphs. Then, a fitness evaluation was conducted and the accuracy calculation was performed according to the obtained classification value. Individuals were then selected in the next step. Here, the selection process was made randomly from individuals with high accuracy in the population and, in addition, from some other individuals to increase diversity. Afterwards, crossover and mutation processes were performed on selected individuals. After this process, the current iteration was completed and the next iteration started. Iterations were repeated as described above until the termination criterion was met.

The configuration of the feature optimization process with the GA was structured according to [36]. The population was initiated with 8192 individuals corresponding to the length of the feature vector. The mutation ratio was set to 1/8192. After each iteration, the GA selects the best individuals from 2% of the population through the tournament selection method, and random individuals are also selected from 2% of the remaining individuals with 0.02% probability. The tournament selection method [37] is a variant of ranking-based selection methods. The corresponding equation is given in (2). Here, a random set of ''k'' individuals was selected; k represents the number of individuals to compete in the tournament. These selected individuals are categorized according to their relative fitness and the most optimal individuals are selected for use in the next iteration. The entire process is repeated ''n'' times for the whole population. The probability of each individual being selected is illustrated by p(i). The terminating criterion of the GA is that it stops when the accuracy of the model remains unchanged or ten iterations pass.

$$p(i) = \begin{cases} \frac{c_{n-1}^{k-1}}{c_n^k}, & if\ i \in [1,\ n-k-1] \\ 0, & if\ i \in [n-k,\ n] \end{cases} \quad (2)$$

With regard to the machine-learning model applied within the GA, SVM and RBF kernels have been used in several different studies [38,39,40]. Here, SVM with an RBF kernel function is used as the fitness function for the GA designed for the proposed framework. SVM using an RBF kernel is employed as a multi-class classifier to match the combined feature with the corresponding categories of 66 motions. Here, 70% of the available data is used for training, 20% for testing, and 10% for validation data. In order to improve the model mentioned here, two basic parameters of the RBF kernel are taken into consideration (equation (3)). Here, the predefined interval of values are illustrated with a and b, ''γ'' defines the effect of a training sample, and ''C'' determines the compromise between the incorrect classification rate and the simplicity of the decision.

In addition, the loose grid search technique [38] was utilized to find the optimal γ and C parameter pair for SVM.

$$C = 2^a, \quad a \in \{-5, -3, -1,\ 1, \ldots,\ 13,\ 15\}$$
$$C = 2^b, \quad b \in \{-15, -13, -11,\ -9, \ldots,\ 1,\ 3\} \quad (3)$$

IV. EXPERIMENTAL RESULTS

This section presents the implementation details and experimental results of the evaluation for the proposed DNN model. Experiments were performed using an Intel Core i7 processor running at 3.5 GHz with 16 GB of RAM in a Linux environment. Python programming language and its libraries (Dlib, NumPy, OpenCV, TensorFlow, Scikit-Learn) were used to implement the proposed architecture. Here, training, test and validation data were chosen randomly for each data set and evaluations were carried out individually. Essentially, 70% of the available data were used for training, 20% for testing and 10% for validation data. As previously mentioned, this selection process was carried out for each dataset separately. Cross-validation that drastically increases the training time was not preferred. The training operation of the network architecture was realized, without GPU support, for about 60 hours and the training procedure was stopped at 300 epochs. Several evaluation metrics—including accuracy, sensitivity, specificity, and f-score—were used for the evaluation process. The accuracy value gives the accuracy rate of the models by measuring the ratio between the evaluated set and the correct results and is calculated by the ratio of the number of correctly classified samples to the total number of samples. The sensitivity and specificity values indicate the effectiveness of a model on positive and negative samples, respectively.

TABLE 1. Hyperparameter configurations for KTH, UCF sports action, SKODA, and UCI smartphone datasets.

Parameters	Skoda	UCI Smartphone	KTH	UCF Sports Action
Epoch	70	60	90	80
Regularization	0.001	0.001	0.001	0.001
Batch Size	64	32	64	64
Learning rate				
0-500 iteration	0.01	0.01	0.01	0.01
500-1000 iteration	0.005	0.005	0.005	0.005
1000+ iteration	0.002	0.002	0.002	0.002
Momentum	0.9	0.9	0.9	0.9
Dropout	0.5	0.5	0.5	0.5
Weight Initialization	Xavier Initialization	Xavier Initialization	Xavier Initialization	Xavier Initialization
Optimization Algorithm	SGD	SGD	SGD	SGD
Loss Function	Categorical Cross Entropy	Categorical Cross Entropy	Categorical Cross Entropy	Categorical Cross Entropy

The f-score parameter shows the measurement of the accuracy of the tested data and, in addition, the f-score parameter is defined as the weighted harmonic mean of the precision and recall values of the tested data. In the study, these evaluation metrics were calculated using the equations (4), (5), (6), and (7) shown below. Here, TP refers to true positive, FP is false positive, whereas TN is true negative and TP is true positive. The evaluation metric values mentioned are the first step in evaluating the performance of the proposed model, where comparison operations were performed by comparison of the proposed hybrid model with four selected DNNs.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$F - score = \frac{(2 * TP)}{(2 * TP + FP + FN)} \tag{7}$$

While Table 1 indicates the initial values of standard configuration parameters for the proposed hybrid CNN architecture defined for each experimental dataset (See section A), Fig. 8 shows the accuracy, sensitivity, specificity and f-score metric values of the proposed network and the AlexNet, VGG-16, GoogleNet, and ResNet152 DNN models for the individual datasets. Configuration parameters of each model, shown in Table 1, were estimated by relying on previous experiments. Besides, a trial-and-error approach was enough to approximate the optimal parameters for the DL model due to the size of the parameter space.

A. EMPLOYED BENCHMARK DATASETS

Details of the KTH, UCF Sports Action, Skoda, and UCI Smartphone datasets are presented below.

- The Skoda dataset [41] includes captured video of assembly line workers at an automobile maintenance facility. Each worker was recorded using 20 accelerometers while performing 46 actions at one of the check points at the factory. The activities include ''close both left doors,'' ''check trunk gaps,'' ''open and close trunk,'' etc. Here, the sampling frequency used for the capture rate of the data was 96 Hz.

- The UCI Smartphone (human activity recognition with smartphone) dataset [42] consists of data obtained from accelerometer and gyroscope sensors. The data were collected from 30 subjects of ages ranging from 19 to 48. Here, each object was equipped with a smartphone and six standard activities were followed. These activities were sitting, lying, standing, walking, walking up and down stairs. With inertial sensors embedded within smartphones fitted to each person, data including 3-axial linear and angular velocities were captured with a sampling rate of 50 Hz. Captured data were then manually tagged.

- The **KTH dataset** [43] is a public dataset consisting of six types of human actions, including walking, clapping, hand-waving, boxing, jogging, and running. This database was created with 25 subjects who performed several times in four different scenarios such as outdoors with varying scales and illumination conditions. The data were obtained

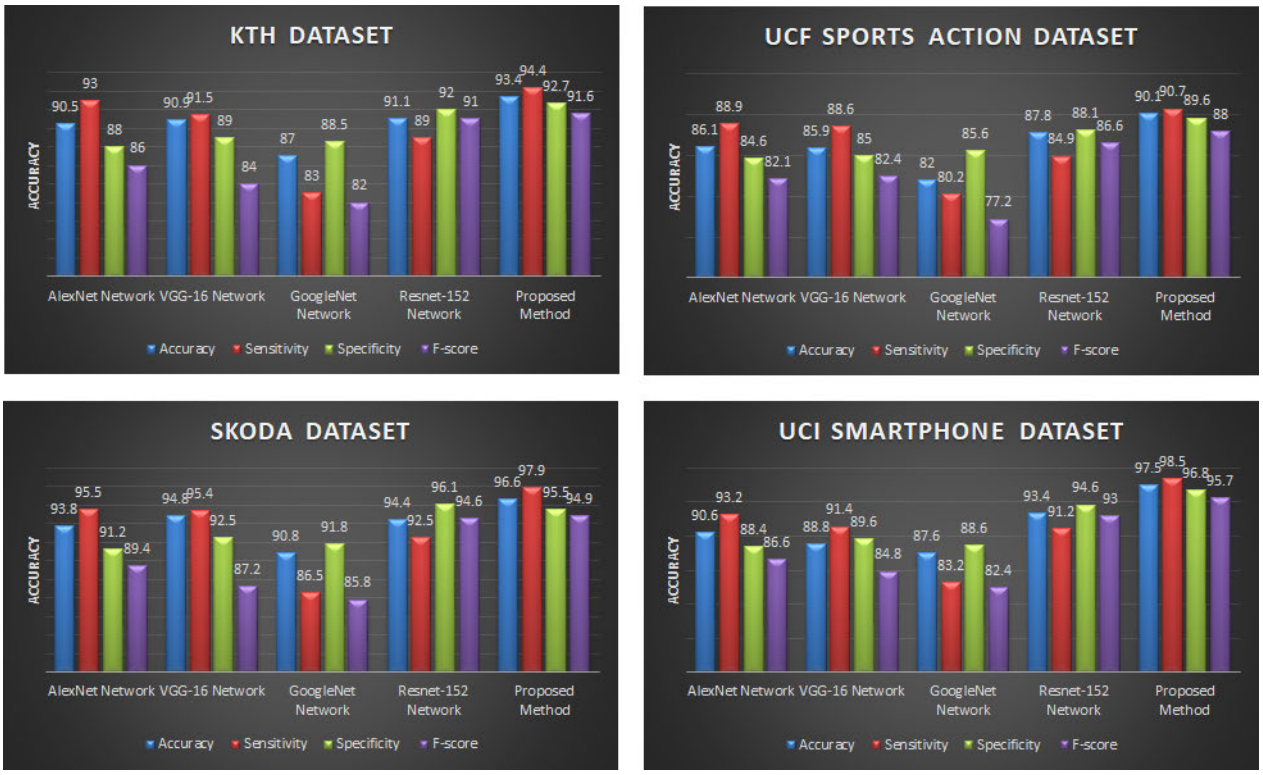

FIGURE 8. The obtained accuracy, sensitivity, specificity, and f-score metric values of the proposed architecture and other contemporary DNN models on the individual KTH (a), UCF Sports Action (b), Skoda (c), and UCI Smartphone (d) datasets.

(a)

	Boxing	Ha. Clap.	Ha. Wav.	Jogging	Running	Walking
Boxing	87.8	5.8	2.6	1.9	3.8	1.9
Ha. Clap.	7.2	85.7	4.4	3.8	4.4	1.5
Ha. Wav.	3.9	2.1	93.3	2.7	6.8	1.5
Jogging	0.0	0.0	2.4	90.6	1.9	0.0
Running	3.9	1.9	2.7	2.1	92.8	1.7
Walking	0.0	0.3	1.2	0.0	2.1	92.5

(b)

	Boxing	Ha. Clap.	Ha. Wav.	Jogging	Running	Walking
Boxing	84.9	4.4	1.8	1.7	2.6	0.8
Ha. Clap.	5.8	85.2	4.3	2.6	1.8	0.9
Ha. Wav.	5.7	2.2	89.8	1.1	3.3	1.2
Jogging	0.0	0.0	1.6	88.3	1.7	0.0
Running	3.3	0.9	2.9	2.2	85.1	1.5
Walking	0.0	0.0	0.7	0.0	1.8	88.5

(c)

	Boxing	Ha. Clap.	Ha. Wav.	Jogging	Running	Walking
Boxing	88.2	5.8	2.6	1.9	3.8	1.9
Ha. Clap.	7.2	85.9	4.4	3.8	4.4	1.5
Ha. Wav.	3.9	2.1	93.6	2.7	6.8	1.5
Jogging	0.0	0.0	2.4	91.6	1.9	0.0
Running	3.9	1.9	2.7	2.1	93.0	1.7
Walking	0.0	0.3	1.2	0.0	2.1	92.8

(d)

	Boxing	Ha. Clap.	Ha. Wav.	Jogging	Running	Walking
Boxing	88.4	4.1	1.1	1.2	3.1	0.9
Ha. Clap.	5.5	86.3	4.9	3.7	2.2	1.1
Ha. Wav.	4.9	0.7	93.7	1.8	5.7	1.1
Jogging	0.0	0.0	2.1	91.9	2.7	0.0
Running	2.5	0.4	2.6	1.4	93.4	0.5
Walking	0.0	0.0	0.4	0.0	2.3	93.1

(e)

	Boxing	Ha. Clap.	Ha. Wav.	Jogging	Running	Walking
Boxing	92.7	3.2	0.7	0.4	1.4	0.1
Ha. Clap.	3.2	91.8	0.1	1.2	3.7	1.0
Ha. Wav.	1.7	1.9	94.8	3.8	2.9	1.3
Jogging	0.0	0.0	0.0	92.4	1.4	0.0
Running	2.6	0.1	1.9	0.2	93.2	1.0
Walking	0.0	0.0	0.8	0.0	0.0	95.3

FIGURE 9. The confusion matrix representations obtained for the KTH dataset for the six action types of AlexNet Network (a) VGG-16 Network (b) GoogleNet Network (c) ResNet-152 Network (d) and Proposed Network (e).

with a fixed camera and uniform background at a frame rate of 25 Hz.

- The **UCF sports action dataset** [44] includes 10 types of sports actions acquired from videos broadcast on TV channels such as BBC and ESPN. These actions include diving, golf swing, kicking, lifting, horse-riding, running, skateboarding, and walking. The dataset contains a total of 150 arrays with images of 720 x 480 pixels. Here, the data

FIGURE 10. Correctly classified examples obtained from the proposed framework.

were obtained by recording at 10 Hz rate in a real sports environment to show variations in illumination conditions, background, and occlusions.

B. DISCUSSION OF THE EXPERIMENTAL RESULTS

According to the graphs shown in Fig. 8, it can be determined that the proposed method outperforms those contemporary architectures, as expected. Besides, the performances of the other four DNNs vary considerably with respect to the four sets of data, while the performance of our network shows similar performance results to those four datasets. The aforementioned situations show that our network is more robust and has superior performance than the other four DNNs. In order to understand and evaluate the performance of the network model, secondly, in addition to the mentioned evaluation metrics, action types were examined together with the confusion matrices. Fig. 9 presents the confusion matrices obtained by the KTH dataset for the proposed network and other DNN models. This figure shows the confusion matrices obtained for the KTH dataset for six action types (boxing, hand-clapping, hand-waving, walking, jogging, and running) of the AlexNet, VGG-16, GoogleNet, and ResNet152 network models, from left to right and top to bottom, respectively. Accuracy rates for each action are presented along with the confusion matrices, and, in addition to this, the relation between the target classes along the x axis and the output classes along the y axis is presented. It can be observed that our proposed method, which shows the confusion matrix in Fig. 9(e), gives better results for all motion classification operations, except for running. In addition, five network models can easily distinguish walking and hand-waving. The ResNet-152 network model shown in Fig. 9(d) provides a better sense of the running action result than other network models. In order to evaluate the performance of the proposed model, a comparison was carried out against state-of-the-art results directly cited from these publications. Table 2 includes the accuracy values obtained for the KTH, UCF Sports Action, Skoda, and UCI Smartphone datasets for the proposed network model and other state-of-the-art studies, respectively. It should be noted that the performance of the proposed architecture is more successful than other state-of-the-art algorithms because it provides a higher accuracy value. Finally, in order to prove the efficiency of the GA algorithm, a T-test was applied to the merged DL model and the model using GA for optimization based on the benchmark datasets. The T-test is a commonly used statistical test which evaluates whether the difference between two sets of data is random or statistically significant. Table 3 illustrates the T-test results using an accuracy value between the combined

TABLE 2. The obtained accuracy metric values of the proposed architecture and other contemporary DNN models on the individual KTH (a), UCF Sports Action (b), Skoda (c), and UCI Smartphone (d) data sets.

Method	*Accuracy(%)*
Ahad et al. [45]	86.7
Chaaraoui at al. [46]	89.86
Ji et al [47]	90.2
Charalampous and Gasteratos [48]	91.99
Proposed Network	**93.4**

(a)

Method	*Accuracy(%)*
Atmosukarto et al. [49]	82.6
Wang et al. [50]	85.6
Le et al. [51]	86.5
Kovashka et al. [52]	87.27
Wang et al. [53]	88.0
Wang et al. [54]	88.2
Proposed Network	**90.1**

(b)

Method	*Accuracy(%)*
Ronao and Cho [58]	90
Ronao and Cho [59]	94.79
Ronao and Cho [60]	94.61
Jiang and Yin [61]	95.18
Proposed Network	**97.5**

(d)

Method	*Accuracy(%)*
Zeng et al. [55]	86.1
Alsheikh et al. [56]	89.3
Ordóñez and Roggen [57]	95.8
Proposed Network	**96.6**

(c)

TABLE 3. T-test results using accuracy value between the combined DL model and the model optimized using ga.

KTH		UCF SPORTS ACTION	
t Stat	20.9965	**t Stat**	29.3854
P(T<=t) one-tail	2.78E-12	P(T<=t) one-tail	2.78E-14
t Critical one-tail	1.7613	**t Critical one-tail**	1.7613
P(T<=t) two-tail	5.56E-12	P(T<=t) two-tail	5.55E-14
t Critical two-tail	2.1447	**t Critical two-tail**	2.1447
SKODA		UCI SMARTPHONE	
t Stat	20.9371	**t Stat**	20.5021
P(T<=t) one-tail	2.89E-12	P(T<=t) one-tail	3.84E-12
t Critical one-tail	1.7613	**t Critical one-tail**	1.7613
P(T<=t) two-tail	5.78E-12	P(T<=t) two-tail	7.68E-12
t Critical two-tail	2.1447	**t Critical two-tail**	2.1447

DL model and the model optimized using GA based on the four comprehensive datasets defined in this study. Here, t Stat values higher than t Critical for both one-tailed and two-tailed predictions suggest that the differences in performances between the two sets of accuracy were found to be statistically significant. $P(T \leq t)$ values very close to zero indicate that the confidence of the evaluation is higher than 99%, because $P(T \leq t) \ll 0.05$. Hence, the proposed approach offers a significant increase in the accuracy values obtained for human action recognition operations when considering four data sets. An example image illustrating the classified actions can be seen in Fig. 10. Despite the accuracy superiority of the proposed architecture over state-of-the-art architectures, it should be noted that it has disadvantage in terms of training time performance due to size of feature vectors as it is expected. It consumes almost two times more computational power than the other state-of-the-art architectures.

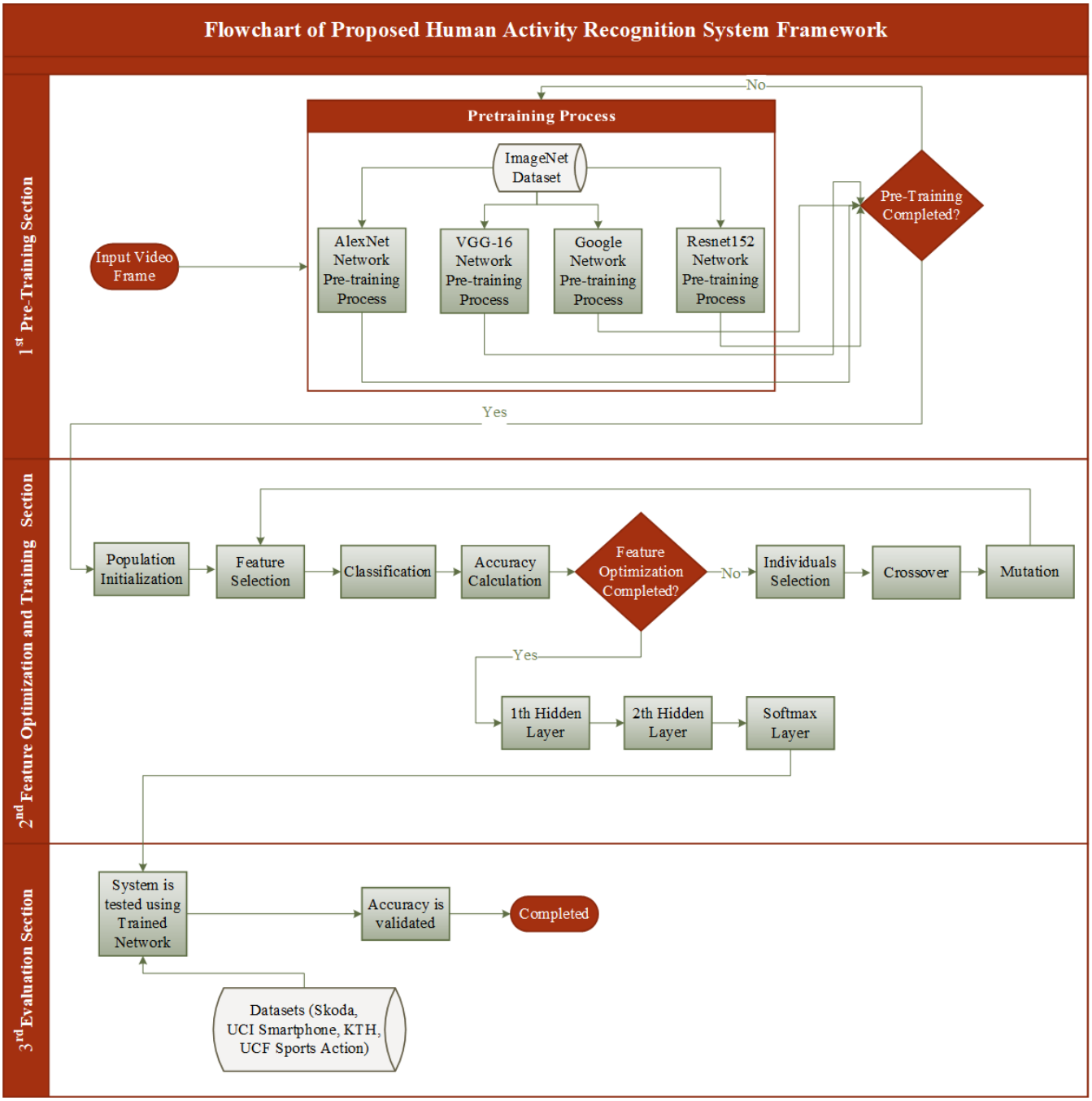

FIGURE 11. Flowchart Representation of the Proposed Deep-Learning Architecture.

V. CONCLUSION

The problem of HAR and limitations of current techniques have motivated the authors of this study to find new and innovative techniques to improve the process. Accordingly, this study proposes a novel deep-learning-based architecture for the recognition of human actions. The architecture proposes a hybrid approach which involves several comprehensive pre-trained networks, relying on the transfer learning method. The contribution of those networks to the overall architecture is implemented and tuned in an optimized manner. This is achieved by employing a popular metaheuristic algorithm, GA, which aims to merge the features obtained from each network. That essentially allows the architecture to approximate an optimized feature vector to be used in the training phase.

As mentioned above, the main contribution of the proposed method is to introduce a hybrid model by combining four well-known pre-trained network models in an optimized manner. The performance of the proposed architecture is evaluated and validated by utilizing benchmark datasets. Thus, the proposed hybrid model is first compared with each individual model separately. It is revealed that the proposed

model outperforms individual models for all scenarios. In addition, the proposed hybrid model is also compared with state-of-the-art studies published in various academic journals. These results also reveal and validate the superiority of the proposed method over state-of-the-art methods. In addition, a T-test was applied to the merged DL model and the model using GA for optimization based on benchmark datasets. Consequently, the differences in performances were found to be statistically significant. The authors believe that this study has the potential to encourage other authors to focus on generating hybrid architectures from deep-learning models by employing metaheuristic algorithms to overcome more complex tasks within areas such as computer vision, human–computer interaction, etc. It should be noted that different metaheuristic or optimization algorithms can be applied to improve the performance of hybrid classifiers by adjusting parameters or input vectors. As opposed to conventional machine learning algorithms, DL algorithms mostly work with huge volumes of unstructured data involving images, audio recordings, and videos. Consequently, the authors plan to apply the proposed hybrid architectures to challenging problems such as zero-shot learning, active learning, etc.

APPENDIX

see Figure 11.

REFERENCES

[1] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[2] K. Zhao, J. Du, C. Li, C. Zhang, H. Liu, and C. Xu, "Healthy: A diary system based on activity recognition using smartphone," in *Proc. IEEE 10th Int. Conf. Mobile Ad-Hoc Sensor Syst.*, Oct. 2013, pp. 290–294.

[3] J. Qin, L. Liu, Z. Zhang, Y. Wang, and L. Shao, "Compressive sequential learning for action similarity labeling," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 756–769, Feb. 2016.

[4] A. A. Yilmaz, M. S. Guzel, I. Askerbeyli, and E. Bostanci, "A vehicle detection approach using deep learning methodologies," in *Proc. Int. Conf. Theor. Appl. Comput. Sci. Eng.*, Nov. 2018, pp. 64–71.

[5] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[6] A. A. Yilmaz, M. S. Güzel, and I. Askerbeyli, "Algılanması güç olan değişimlerin ortaya çıkarılması için resim teleskobu çalışmalarının analiz çalışmaları," *Dokuz Eylül Univ. Fac. Eng. J. Sci. Eng.*, vol. 19, no. 57, pp. 723–732, 2017.

[7] I. Fatima, M. Fahim, Y. K. Lee, and S. Lee, "A genetic algorithm-based classifier ensemble optimization for activity recognition in smart homes," *KSII Trans. Internet Inf. Syst.*, vol. 7, pp. 2853–2873, Nov. 2013.

[8] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: A survey," 2015, *arXiv:1501.05964*. [Online]. Available: http://arxiv.org/abs/1501.05964

[9] K. Aggarwala and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[10] X. Xiao, D. Xu, and W. Wan, "Overview: Video recognition from handcrafted method to deep learning method," in *Proc. Int. Conf. Audio, Lang. Image Process. (ICALIP)*, Jul. 2016, pp. 646–651.

[11] D. Wu, N. Sharma, and M. Blumenstein, "Recent advances in video-based human action recognition using deep learning: A review," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2865–2872.

[12] Z. Zhang, X. Ma, R. Song, X. Rong, X. Tian, G. Tian, and Y. Li, "Deep learning based human action recognition: A survey," in *Proc. Chin. Autom. Congr. (CAC)*, Jinan, China, Oct. 2017, pp. 3780–3785.

[13] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 36–45.

[14] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*. [Online]. Available: http://arxiv.org/abs/1405.3531

[15] A. Kaya, A. S. Keçeli, and A. B. Can, "Akciğer nodül özelliklerinin tahmininde çeşitli sınıflama stratejilerinin incelenmesi," *J. Fac. Eng. Archit. Gazi Univ.*, vol. 34, no. 2, pp. 709–726, 2018.

[16] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[20] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bioinspired dynamic 3D discriminative skeletal features for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 471–478.

[21] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595.

[22] R. Culmone, G. Paolo, and M. Quadrini, "Human activity recognition using a semantic ontology-based framework," *Int. J. Adv. Intell. Syst.*, vol. 8, pp. 159–168, 2015.

[23] A. G. Salguero, P. Delatorre, J. Medina, M. Espinilla, and A. J. Tomeu, "Ontology-based framework for the automatic recognition of activities of daily living using class expression learning techniques," *Sci. Program.*, vol. 2019, pp. 1–19, Apr. 2019.

[24] Z. Zeng and Q. Ji, "Knowledge based activity recognition with dynamic Bayesian network," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2010, pp. 532–546.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[26] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence* (Complex Adaptive Systems). Cambridge, MA, USA: MIT Press, 1992.

[27] J. H. Holland, "Genetic algorithms," *Sci. Amer.*, vol. 267, no. 1, pp. 66–73, 1992.

[28] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Mach. Learn.*, vol. 3, nos. 2–3, pp. 95–99, 1988.

[29] A. Gülcü and Z. Kuş, "Konvolüsyonel sinir aglarinda hiper-parametre optimizasyonu yöntemlerinin incelenmesi," *Gazi Uni. Sci. J. Part C, Des. Technol.*, vol. 7, no. 2, pp. 503–522, 2019.

[30] B. Minaei-Bidgoli and W. Punch, "Using genetic algorithms for data mining optimization in an educational Web-based system," in *Proc. Genetic Evol. Comput. Conf.*, Jul. 2003, pp. 2252–2263.

[31] H. Handels, T. Rob, J. Kreuschb, H. H. Wolffb, and S. J. Pöppl, "Feature selection for optimized skin tumor recognition using genetic algorithms," *Artif. Intell. Med.*, vol. 16, pp. 283–297, Jul. 1999.

[32] A. M. Karim, M. S. Güzel, M. R. Tolun, H. Kaya, and F. V. Çelebi, "A new generalized deep learning framework combining sparse autoencoder and Taguchi method for novel data classification and processing," *Math. Problems Eng.*, vol. 2018, pp. 1–13, Jun. 2018.

[33] N. Kumaran, A. Vadivel, and S. S. Kumar, "Recognition of human actions using CNN-GWO: A novel modeling of CNN for enhancement of classification performance," *Multimedia Tools Appl.*, vol. 77, no. 18, pp. 23115–23147, Sep. 2018.

[34] A. M. Karim, M. S. Güzel, M. R. Tolun, H. Kaya, and F. V. Çelebi, "A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing," *Biocybernetics Biomed. Eng.*, vol. 39, no. 1, pp. 148–159, Jan. 2019.

[35] (2020). *The ImageNet Dataset From the Facial Expression Recognition Challenge*. [Online]. Available: http://www.image-net.org/

[36] R. Cilla, M. Patricio, J. García, A. Berlanga, and J. Molina, "Recognizing human activities from sensors using hidden Markov models constructed by feature selection techniques," *Algorithms*, vol. 2, no. 1, pp. 282–300, Feb. 2009.
[37] K. Jebari and M. Madiafi, "Selection methods for genetic algorithms," *Int. J. Emerg. Sci.*, vol. 3, no. 4, pp. 333–344, Dec. 2013.
[38] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci. Inform. Eng., Univ. Nat. Taiwan, Taipei, Taiwan, Tech. Rep., 2003. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
[39] T. D. T. Nguyen, T.-T. Huynh, and H.-A. Pham, "An improved human activity recognition by using genetic algorithm to optimize feature vector," in *Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2018, pp. 333–344.
[40] H. Cao, M. N. Nguyen, C. Phua, S. Krishnaswamy, and X.-L. Li, "An integrated framework for human activity classification," in *Proc. ACM Conf. Ubiquitous Comput. (UbiComp)*, 2012, pp. 333–344.
[41] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster, "Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection," in *Proc. Eur. Conf. Wireless Sensor Netw.*, vol. 4913, Feb. 2008, pp. 17–33.
[42] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Proc. Int. Conf. Ambient Assested Living*, 2012, pp. 216–223.
[43] M. Rodriguez, "Spatio-temporal maximum average correlation height templates in action recognition and video summarization," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. Central Florida, Orlando, FL, USA, 2010.
[44] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, Sep. 2004, pp. 32–36.
[45] M. A. R. Ahad, M. N. Islam, and I. Jahan, "Action recognition based on binary patterns of action-history and histogram of oriented gradient," *J. Multimodal User Interfaces*, vol. 10, no. 4, pp. 335–344, Dec. 2016.
[46] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 1799–1807, Nov. 2013.
[47] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[48] K. Charalampous and A. Gasteratos, "On-line deep learning method for action recognition," *Pattern Anal. Appl.*, vol. 19, no. 2, pp. 337–354, May 2016.
[49] I. Atmosukarto, N. Ahuja, and B. Ghanem, "Action recognition using discriminative structured trajectory groups," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 899–906.
[50] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 124.1-124.11.
[51] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3361–3368.
[52] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2046–2053.
[53] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.
[54] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3169–3176.
[55] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th Int. Conf. Mobile Comput., Appl. Services*, 2014, pp. 197–205.
[56] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep activity recognition models with triaxial accelerometers," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 8–13.
[57] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, pp. 1–25, 2016.
[58] C. A. Ronaoa and S. B. Cho, "Evaluation of deep convolutional neural network architectures for human activity recognition with smartphone sensors," in *Proc. Int. KIISE Korea Comput. Congr.*, 2015, pp. 858–860.
[59] C. A. Ronao and S. B. Cho, "Deep convolutional neural networks for human activity recognition with smartphone sensors," in *Proc. Int. Conf. Neural Inf. Process.*, Nov. 2015, pp. 46–53.
[60] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016.
[61] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, 2015, pp. 1307–1310.

ABDULLAH ASIM YILMAZ received the B.S. degree from the Computer Engineering Department, Başkent University, Turkey, in 2010, and the master's degree from the Computer Engineering Department, Graduate School of Natural and Applied Sciences, TOBB University of Economics and Technology, Turkey, in 2012. He is currently pursuing the Ph.D. degree with the Computer Engineering Department, Graduate School of Natural and Applied Sciences, Ankara University, Turkey. He has been working as a Computer Engineer with the Republic of Turkey Ministry of Agriculture and Forestry, Ankara, Turkey, since 2011. His research interests include the artificial intelligence, image processing, computer vision, pattern recognition, classification, software development, image and video analysis, and object recognition and tracking.

MEHMET SERDAR GUZEL received the B.S. and M.S. degrees from the Computer Engineering Department, Ankara University, Ankara, Turkey, and the Ph.D. degree from the Mechanical and Systems Engineering Department, Newcastle University, U.K., in 2012. From 2006 to 2012, he was a Research Assistant with Ankara University, where he was an Assistant Professor, from 2014 to 2019. His research interests include the image processing, software development, control theory, and robotics.

ERKAN BOSTANCI received the B.Sc. degree and the M.Sc. degree in real-time battlefield simulation from the Computer Engineering Department, Ankara University, Turkey, in 2007 and 2009, respectively, and the Ph.D. degree from the School of Computer Science and Electronic Engineering, University of Essex, U.K., in 2014. He joined the Computer Engineering Department, Ankara University, as a Research Assistant. His research interests include different yet closely related aspects of computer science from image processing, computer vision and graphics to artificial intelligence and fuzzy logic, as well as mathematical modeling and statistical analysis. He has been involved in technical committees for several conferences and has been organizing an international conference for several years as well as acting as a reviewer for various journals.

IMAN ASKERZADE received the B.S. and M.S. degrees from the Physics Department, Moscow State University, Russia, and the Ph.D. degree from Moscow State University and Azerbaijan Academy of Sciences, Azerbaijan, in 1995. From 2003 to 2005, he was an Associate Professor with Ankara University, Ankara, Turkey. From 2006 to 2007, he was an Associate Professor with TOBB ETÜ University, Ankara. His research interests include the fuzzy Logic, quantum computing, modeling, and simulation.