IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 30, 2017, accepted December 8, 2017, date of publication December 14, 2017,
date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2783331

Evolutionary Feature Learning for 3-D Object Recognition

SYED AFAQ ALI SHAH, MOHAMMED BENNAMOUN, FARID BOUSSAID, AND LYNDON WHILE
The University of Western Australia, Crawley, WA 6009, Australia

Corresponding author: Syed Afaq Ali Shah (afaq.shah@uwa.edu.au)

This work was supported in part by The University of Western Australia and in part by the Australian Research Council Linkage under Grant LP130100138.

ABSTRACT 3-D object recognition is a challenging task for many applications including autonomous robot navigation and scene understanding. Accurate recognition relies on the selection/learning of discriminative features that are in turn used to uniquely characterize the objects. This paper proposes a novel evolutionary feature learning (EFL) technique for 3-D object recognition. The proposed novel automatic feature learning approach can operate directly on 3-D raw data, alleviating the need for data pre-processing, human expertise and/or defining a large set of parameters. EFL offers smart search strategy to learn the best features in a huge feature space to achieve superior recognition performance. The proposed technique has been extensively evaluated for the task of 3-D object recognition on four popular data sets including Washington RGB-D (low resolution 3-D Video), CIN 2D3D, Willow 2D3D and ETH-80 object data set. Reported experimental results and evaluation against existing state-of-the-art methods (e.g., unsupervised dictionary learning and deep networks) show that the proposed EFL consistently achieves superior performance on all these data sets.

INDEX TERMS 3-D object recognition, feature learning, evolutionary algorithms.

I. INTRODUCTION

The availability of high-dimensional data and its meaningful representation demands the use of feature learning and selection in many pattern recognition tasks [1]. In the real-world image datasets, a large number of irrelevant and redundant features generally exist. These features may significantly affect the performance of learned models and reduce their learning speed. Feature learning is a strategy that involves selection of salient features or removal of redundant features, from a set of given features, to improve predictive accuracy for challenging problems such as 3D object recognition. These salient features help the learned model to achieve good predictive accuracy. Accurate feature selection/learning is therefore very crucial and an active research area in pattern recognition and machine learning [2].

The conventional approaches for feature selection can be broadly categorized as: hand-crafted, automatic learning, and hybrid approaches [3], [4]. The first approach requires careful extraction and analysis of the feature set by human expertise without utilizing any learning model or optimization [5], [6]. The second approach aims to find the most representative features in a large feature space using a pre-determined learning model. The hybrid approach attempts to take advantage of the hand-crafted and automatic learning approaches [7]. Hybrid techniques are shown to find a good solution, while a single technique often traps into an immature solution.

The success of different feature selection methods depends on the search strategy used in feature selection process. One method is to start the search process with an empty set and successively add features. This is called the sequential forward search (SFS) [8]. Another approach, called the sequential backward search (SBS), is to start with a full set of features and successively remove redundant features [9]. This sequential strategy is less complex and computationally efficient but it is affected by nesting effect, which means that once a feature has been added, it cannot be deleted and vice versa.

The main problems with these search strategies is that they search the feature space locally rather than globally and therefore these approaches attempt to find solutions that range between sub-optimal and near optimal regions. These approaches involve partial search and therefore solution of optimal or near optimal is quite difficult to achieve. In addition, these search strategies also suffer from computational complexity. To address these problems, the recent research has been shifted towards the global search algorithms. They find the solution in the full search space by their global search capability. The global search algorithms work on basis of the activity of multi-agents, which ultimately enhance to find very high-quality solutions within a reasonable time. Because of their global search and parallelism nature, Evolutionary Algorithms (EA) can successfully and efficiently solve the feature selection/learning in a large feature space [10], [11]. These capabilities of EAs have been tested for training neural networks [12], fluorescence fingerprinting of plant species recognition, image classification and action recognition. In the literature, 3D object recognition using EAs has also been reported. However, instead of using 3D information, these techniques either rely on 2D projections obtained from 3D views [13] or intensity information assuming orthographic projection [14]. These techniques are therefore geared towards 2D data only and do not exploit any 3D information. Feature learning using EAs from *raw 3D data* for the challenging task of *3D object recognition* is therefore an open research problem.

In this paper, we fill this research gap by proposing a novel Evolutionary Feature Learning (EFL) technique for 3D object recognition. To the best of our knowledge, this is the first ever investigation of EAs for feature learning using raw 3D data for 3D object recognition. The proposed EFL exploits the strengths of Evolutionary Algorithms[1] (EAs) for a challenging task of 3D object recognition and offers a number of advantages: **(1)** EFL learns the best feature (candidate solution) by attempting many unconventional permutations of chromosomes/individuals (through crossover and mutation) and these permutations yield an exceptionally good recognition performance, **(2)** the parallelism inherent in EFL facilitates its efficient implementation on GPU architectures, and enables the algorithm to avoid being trapped in local optimal solution, thus greatly increasing the search speed for the candidate solution, **(3)** It could also facilitate efficient exploration of a larger feature space, thus significantly increasing the chance of learning the ''best'' features, **(4)** EFL works on the chromosomes, which are encoded version of potential solutions' parameter, rather the parameters themselves, **(5)** EFL uses probability selection rules and fitness score, which are obtained from fitness function, without any other complex information. **(6)** In contrast to deep learning techniques, which require large training data for learning and to avoid over-fitting, the evolutionary feature learning can be performed on smaller input data.

The proposed technique was extensively evaluated for the task of 3D object recognition on Washington RGB-D (low resolution 3D video), CIN 2D3D, Willow 2D3D and ETH-80 object datasets, and experimental results indicate that it achieves state-of-the-art performance (Section V).

The rest of this paper is organized as follows. The next section surveys the related work. Section III introduces the proposed evolutionary feature learning technique. Section IV discusses the proposed object recognition algorithm. Experimental results and evaluation against existing state-of-the-art techniques are provided in Section V. Finally, a conclusion is given in Section VII.

II. RELATED WORK

In this section, we briefly discuss the existing feature based techniques, which can be grouped into two categories depending on whether they use hand-crafted features or automatic feature learning. In addition, we briefly review prior applications of EAs to computer vision.

A. HAND-CRAFTED FEATURES

The most common approach for object recognition is to use well-designed hand-crafted features. In these techniques, features are defined by human experts in terms of local neighborhood operations applied to an input image [15].

Zhang [16] proposed Harmonic Shape Image (HSI). The latter is based on harmonic map theory [17]. HSIs represent 3D surface regions as 2D images, thus reducing the 3D patch matching to less complex 2D image matching. HSI has, however, the following constraints. First, the construction of the boundary mapping for HSI requires ordering of the boundary vertices of the local surface patch in either clockwise or counter clockwise direction. Second, to facilitate matching of the surface patches, the consistency of two orders is also mandatory. Sun *et al.* [18] proposed Heat Kernel Signature (HKS). In their proposed technique, the 3D mesh is considered as a Riemannian manifold and the heat kernel $H_t(m, n)$ is restricted to the temporal domain $H_t(m, m)$. The HKS feature can be interpreted as a multi-scale notion of the Gaussian curvature, where the time parameter t provides a natural notion of scale. Tombari *et al.* [19] proposed a feature named Signature of Histograms of OrienTations (SHOT), whereby a local reference axis is first constructed for a feature point, and the neighborhood space is then divided into 3D spherical volumes. A local histogram is then generated for each volume by accumulating the number of points according to the angles between the normal at the feature point and those at the neighboring points. Recently, Shah *et al.* [20] proposed 3D-Div, which exploits the divergence of the vector field at each point of the local surface to construct the local feature. 3D-Div has been shown to achieve superior object recognition performance on low resolution data. Hulin and Troyanov [21] derived the relationship between the volume descriptor and the mean curvature, that was later on used by Gelfand *et al.* [22] for surface matching and global registration. Clarenz *et al.* [23], [24] used Principal Component Analysis (PCA) of 3D patches for feature detection. Their technique has been shown to achieve good performance and robustness.

These existing hand-crafted feature based methods suffer from low descriptiveness, require significant human

[1] In this paper, the Evolutionary Algorithms (EAs) used for feature learning are Genetic Algorithms, which are referred to as EAs throughout the paper.

intervention and the construction of a local reference axis to achieve superior performance [25]. To overcome these limitations, automatic feature learning methods have been proposed.

B. AUTOMATIC FEATURE LEARNING

In automatic feature learning, a transformation of raw inputs to a representation is learnt automatically [26]. The automatically learnt representation is then exploited in several tasks such as recognition. Feature learning using sparse coding and deep networks has recently received significant attention. Bo *et al.* [27] proposed Hierarchical Matching Pursuit (HMP) for automatic feature learning. In their proposed technique, K-SVD algorithm [28] is used to learn dictionaries over image patches. These learned dictionaries are then used to create feature hierarchies by using spatial pyramid pooling and orthogonal matching pursuit [27]. In deep network based approaches, the multiple layers of the deep belief nets [29] are trained in hierarchical fashion using the unsupervised Restricted Boltzmann Machine (RBM) for automatic feature learning. This unsupervised training helps to elude the problem of local minima. Next, the supervised training is done to adjust the learned weights. In convolutional deep belief nets proposed by Lee *et al.* [30], the weights are shared between the visible and hidden layers while a small filter is used to automatically learn features from full-size images. In the automatic feature learning approach proposed by Munawar *et al.* [31], Deep Reconstruction Model (DRM) is built by stacking autoencoders. The parameters of DRM are initialized using Gaussian RBM. DRM is then trained in a layer-wise fashion using hand-crafted Local Binary Pattern (LBP) features [32] to reconstruct the input images. Deep learning based DRM requires computationally expensive step of data pre-processing e.g. PCA whitening to achieve good object recognition performance. In addition, the performance of DRM also relies on hand-crafted LBP features. Similar to DRM, Convolutional Deep Network [33], denoising autoencoder [34] and deep Boltzmann machines [35] are other examples of automatic feature learning techniques.

C. EA BASED METHODS

In human recognition method proposed by Ijjina and Chalavadi [36], the weights of a Convolutional Neural Network (CNN) are initialized by using an evolutionary algorithm to minimize the classification error. They utilize the global search capabilities of EAs to find the most optimal solution. In a technique proposed by Fougerolle et al [37], EA is used to recover 2D rational Gielis curves, which can represent a wide range of shape and patterns. In their proposed technique, they exploit EA to define a cost function based on shortest Euclidean distance. Their technique has been shown to achieve good accuracy in the presence of noise and missing data. Stanhope and Daida [38] proposed EA based approach for clutter classification in synthetic aperture radar imagery. In their proposed technique, 10 different types of hand-crafted features are used. These features are first normalized by the mean and standard deviation of the feature vectors generated on the training data. These pre-processed/normalized features are then passed to the EA for processing and classification. In a craniofacial disorder estimation technique proposed by Atmosukarto *et al.*, [39], 2D histograms are formed for the given facial region by computing azimuth and elevation angles of surface normals. To optimize the classification, Adaboost learning is used to select the histogram bins as features. The latter are then combined using an EA which also quantifies the abnormality for a given face. In a technique proposed by Perez and Olague *et al.* [40], the EA synthesizes the mathematical expressions that are required to improve the well-known hand-crafted Scale Invariant Feature Transform (SIFT) feature [41]. Their improved SIFT feature has been shown to achieve better recognition compared to the original SIFT. However, their performance is quite comparable with other invariants of SIFT that include GLOH, DoG and SURF features. The main shortcoming of their technique lies in that they assume 50% overlap between the two images for accurate image match/recognition. In addition, their technique is only applicable to 2D gray scale images.

D. DISCUSSION

From the above reported literature review, one can note that: **(i)** Most existing hand-crafted and automatic feature learning techniques are only geared towards 2D gray scale images. **(ii)** EAs have been used either to improve or combine hand-crafted features extracted from 2D images [40], [39]. **(iii)** Most feature learning methods (e.g. EA based approach [38] and deep learning based [31]) require a large set of parameters and a computationally expensive data pre-processing step to improve the recognition performance. **(iv)** High performance (e.g. in the case of deep learning based [31] and EA based [39], [38]) also relies on hand-crafted features.

This paper overcomes the aforementioned shortcomings and proposes a novel evolutionary feature learning technique for multi-class 3D object recognition. The proposed approach does not require the pre-processing of the input images e.g. PCA/ZCA whitening. It operates directly on *raw* 3D and 2D RGB images. Furthermore, unlike deep networks, it does not require prior initialization of the parameters to achieve optimal results. Rather, the proposed technique automatically optimizes the candidate solution based on the fitness function and selects the best feature for superior object recognition. In the following, we describe our proposed evolutionary feature learning technique.

E. EA TERMINOLOGIES

In this section, we briefly define the EA terminologies, which will be used interchangeably in this paper.

1) POPULATION

Population holds all the possible solution. It is a multi-set of chromosomes.

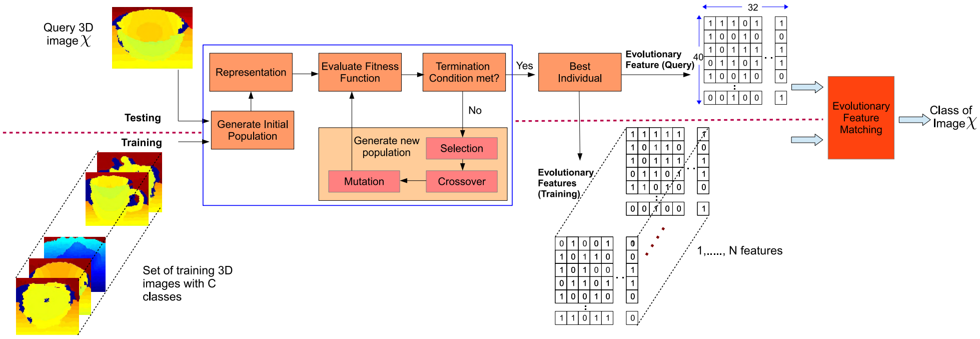


FIGURE 1. Illustration of the proposed EA based 3D object recognition system.

2) CANDIDATE SOLUTION/INDIVIDUAL

In the real world domain, candidate solution, individuals or genotype are used interchangeably to represent points of the space of possible solutions.

3) CHROMOSOME/GENOTYPE

In the EA domain, chromosome, genotype and even individual is used for the points in the space where EA process actually takes place.

4) REPRESENTATION

Representation is used to link the real world with EA space.

III. PROPOSED EVOLUTIONARY FEATURE LEARNING

In this section, we describe our proposed evolutionary feature learning technique (illustrated in Fig. 1) for 3D object recognition. Each individual in the population is represented as an encoding of a "hypothesized image",[2] and the fitness function measures the difference between this hypothesized image and the input image χ. Section III-B describes the representation used. Section III-C describes the fitness function. Sections III-D and III-E describe other relevant aspects of the EFL. Algorithm 1 summarizes our evolutionary feature learning based algorithm. The proposed evolutionary feature learning is initialized by first linking the real world i.e. image domain to the EA domain. The aim of a representation is to set up a bridge between the original problem context and the problem solving space where evolution will take place. During the representation, the genotype to phenotype mapping is defined. We propose two types of representations to cater for the type of input image i.e. 3D or *RGB*, we therefore propose 3D and 2D representations.

Algorithm 1 Proposed Evolutionary Feature Learning Algorithm

```
Input: Input Image $I$, Initial Population $P(t) = \{\Psi^{(1)}, \cdots, \Psi^{(N)}\}$
$G \leftarrow$ Maximum number of generations
for $t = 1, \cdots, G$ do
    while $NOT$ (Terminate condition for $N$) do
        $f(h_I(\Psi_i^{(z)})) \leftarrow E\left[\left\|h_I(\Psi_i^{(z)}) - I\right\|^2\right]$ (See Eq. 1)
        Sort chromosomes $\Psi_i$ based on fitness function $f(h_I(\Psi_i^{(z)}))$
        for $i = 1, \cdots, N$ do
            Select best parent chromosomes with minimum $f(h_I(\Psi_i^{(z)}))$
            $(p_1, p_2) \leftarrow ParentSelection(P(t))$
            $(c_1, c_2) \leftarrow$ crossover$(p_1, p_2)$
            $(c_1) \leftarrow$ mutation$(c_1)$
            $P(t) \leftarrow P(t) \cup \{c_1\}$
        end
        Select best chromosomes with minimum $f(h_I(\Psi_i^{(z)}))$
        $\Psi_i \leftarrow \underset{h_I(\Psi_i^z)}{arg\ min}\ f(h_I(\Psi_i^z))$
    end
end
Output: Evolutionary Feature $\leftarrow \Psi_i$
```

A. POPULATION AND GENERATIONS

In the proposed evolutionary feature learning technique, the population of chromosomes $P(t) = \{\Psi^{(1)}, \cdots, \Psi^{(N)}\}$ and number of generations, $t = 1, \cdots, G$ are defined for a given input image. The population size and number of generations are set to $N = 100$ and $G = 1000$ respectively, based on empirical tests.

B. PROPOSED REPRESENTATION

1) PROPOSED 3D REPRESENTATION

For each input depth image I of size $a \times b$ (Fig. 2(a)), a hypothesized image $h_I(\Psi_i^{(z)})$, $i = 1, \cdots, K$ and $z = 1, \cdots, N$, of the same size is initialized in the EA space for a given zth chromosome, and i represents the ith class of the input image. As opposed to intensity values in 2D images, the depth images contain $depth(d)$ information about an object which is the distance from the point on the surface

[2] The term hypothesized image refers to a reconstructed image. These two terminologies have been used interchangeably in this paper.

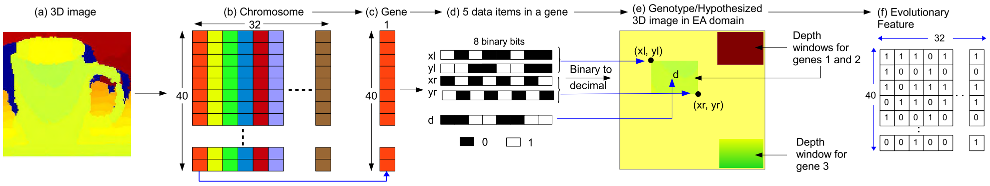


FIGURE 2. 3D Representation in EA domain. Each chromosome/individual of the population (dimension 40 × 32) represents a possible solution. Each chromosome consists of 32 genes, each of 1 × 40.

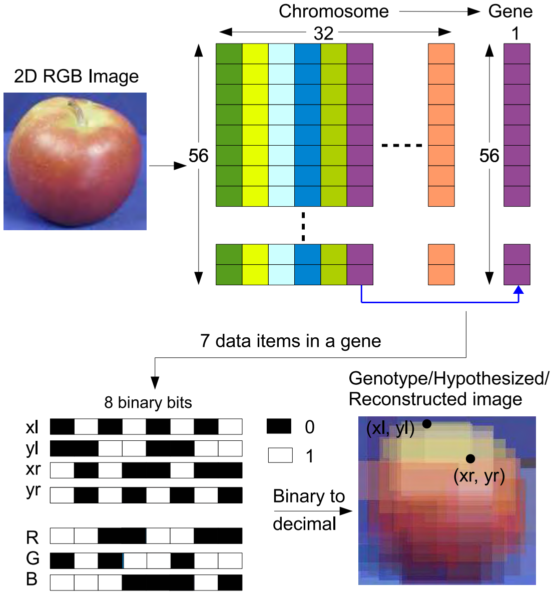


FIGURE 3. 2D Representation in EA domain. Each chromosome/individual of the population (dimension 56 × 32) represents a possible solution. Each chromosome consists of 32 genes, each with a 1 × 56 dimension.

of the object to the sensor (3D scanner). Depth contains 3D information about the object's geometry. Next, a candidate solution, of dimension 40 × 32, is defined by generating a chromosome (Fig. 2(b)), comprising of 32 genes. Each gene has a dimension 1 × 40, as shown in Fig. 2(c). Each gene consists of 5 data items, namely depth d and 4 coordinate values, that is, the locations of the upper left $(x_l; y_l)$ and lower right corners $(x_r; y_r)$ of the image window, as shown in Fig. 2(e). Each of these 5 data items are further described by 8 bit binary integers (therefore the gene has a size of 1 × 40), as in Fig. 2(d). These 32 genes are then used to reconstruct the input image in EA space, by randomly placing 32 semi-transparent (overlapping) *depth* windows in the hypothesized image $h_I(\Psi_i^{(z)})$ (Fig. 2(e)). Each gene therefore encodes the parameters of the *depth* window and specifies a portion of the original image that will be used to reconstruct a similar image in the EA domain. As a result, our evolutionary information describing one individual would require 32 genes × 5 data items × 8 bits = 1280 bits. The genes and the resulting depth windows are then modified, using an EFL (Sec. III-E), to improve the 3D representation of the image in EA domain.

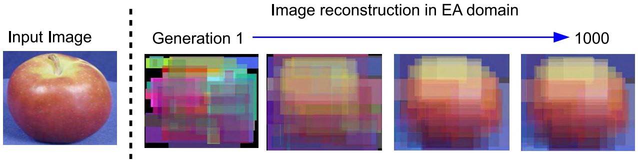


FIGURE 4. Reconstruction of Image in EA domain. Left. Input image. Right. The input image is reconstructed in each generation and compared with the input image. The chromosome, which reconstructs the image with minimum error qualifies as the best individual of that population. Figure best viewed in color.

2) PROPOSED 2D REPRESENTATION

To demonstrate that our technique is generic and can also be applied to 2D data, we also propose a 2D representation. For 2D color (*RGB*) images, the following changes are made to represent *RGB* information of the input image in the EA space. A candidate solution of dimension 56 × 32 is defined that also includes 32 genes, each having a slightly large dimension of 1 × 56 as shown in Fig. 3. In this case, each gene consists of 7 data items, namely three color channels R, G, B and 4 coordinate values, that is, the locations of the upper left $(x_l; y_l)$ and lower right corners $(x_r; y_r)$ of the image window. As in the case of 3D representation, each of these 7 data items are further described by 8 bit binary integers, as can be seen in Fig. 3. In this case, our *RGB* evolutionary information would require 32 genes × 7 data items × 8 bits = 1792 bits to describe one individual.

Fig. 4 illustrates the reconstruction of the hypothesized image in the EA domain. Only the images reconstructed using the best chromosomes are displayed here. The hypothesized images are generated using all the chromosomes of the population in a given generation. The hypothesized images are compared with the input image. The hypothesized image with minimum error is selected and the chromosome which reconstructs this image with a minimum fitness score qualifies as the best individual of that population.

C. FITNESS FUNCTION

For each candidate solution $\Psi_i^{(z)}$, the fitness is evaluated as:

$$f(h_I(\Psi_i^{(z)})) = E\left[\left\|h_I(\Psi_i^{(z)}) - I\right\|^2\right] \tag{1}$$

where $h_I(\Psi_i^{(z)})$ is the hypothesized image, reconstructed in the EA domain using the zth chromosome $\Psi_i^{(z)}$. Note that a perfect solution would have fitness 0, and that low fitness scores are better than high ones. The chromosomes are ranked on their fitness, so the best chromosome Ψ_i is given by,

$$\Psi_i = arg \underset{h_I(\Psi_i^{(z)})}{min} f(h_I(\Psi_i^{(z)})) \tag{2}$$

D. PARENT SELECTION

The role of parent selection is to distinguish among individuals based on their quality (computed using the fitness function), in particular to allow the better individuals to become parents to the next generation. An individual is a parent if it has been selected to undergo variation in order to create offspring. Based on the fitness values, high quality individuals have a higher chance to become parents than those with low quality. This, however, can cause the whole search (for the best solution) to become too greedy and get stuck in a local minimum.

In order to avoid this local minimum problem, instead of selecting the best parents for mating, we probabilistically select two parents p_1 and p_2 from a subset of the population i.e. $\{\Psi_i^{(1)}, \cdots, \Psi_i^{(N-2)}\}$. Thus, low quality individuals are also given a small but positive chance to become parents: each individual is assigned a probability that depends on its quality, i.e. on its fitness value. The parents (p_1, p_2) are then used for crossover and mutation to define new offspring, which replace individuals $\Psi_i^{(N-1)}$ and $\Psi_i^{(N)}$ to form a new population of chromosomes.

E. CROSSOVER AND MUTATION

During crossover, the parents p_1 and p_2 are recombined by randomly selecting the crossover points. The two chromosomes are then interchanged. This gives rise to two new child genes c_1 and c_2 for the next generation. c_1 and c_2 then replace $\Psi_i^{(N-1)}$ and $\Psi_i^{(N)}$ i.e. chromosomes with low quality (high fitness scores) in the previous generation.

Next, mutation is applied to one randomly selected chromosome and it results in a slightly modified mutant (i.e. the child or offspring of the selected chromosome). A random bit is selected in the chromosome and flipped to accomplish the mutation process.

IV. OBJECT RECOGNITION ALGORITHM

In this section, we describe our proposed object recognition algorithm. The algorithm consists of two parts: 1) *evolutionary learning* to learn class specific features in a supervised way and 2) *recognition* to decide on the identity of a query image.

Given m training images and their corresponding labels $\mathcal{L} \in [1, 2, \cdots K]$ where $K < m$, a genotype to phenotype mapping, the population size $P(t)$, and the number of generations t are defined for each training image. During the evolutionary learning process, feature representations are learnt by optimizing the fitness function for each training image to select the best individual/feature Ψ_i for each training image.

During recognition, given a test image $\mathcal{I}_{test}$, we first learn evolutionary features for the test image and select the best feature $\widetilde{\Psi}$, using the procedure stated above. We next calculate the error between the test $(\widetilde{\Psi})$ and the training $(\Psi_i^{(t)})$ EFs:

$$d_i(\Psi) = \left\| \widetilde{\Psi} - \Psi_i \right\|_2, \quad i = 1, \ldots, K \tag{3}$$

and rule in favour of the class with minimum distance i.e.,

$$\underset{i}{min}\, d_i(\Psi) \tag{4}$$

V. EXPERIMENTAL RESULTS

We evaluated and compared the performance of the proposed technique with existing state-of-the-art approaches for the task of 3D object recognition. The performance evaluation is presented for four popular publicly-available object datasets:

- Washington RGB-D (low resolution) [42]
- CIN 2D3D [43]
- Willow 2D3D [44]
- ETH-80 [45]

The detailed description of each of these datasets and our experimental results are reported in Sec. V-A.

A. DATASETS AND RESULTS

1) WASHINGTON RGB-D DATASET

This dataset is the largest available low resolution 3D video dataset. The dataset contains 300 objects in 51 different categories. There are roughly 600 images for each object captured from three different angles with respect to the horizon. Fig. 5 shows sample images of the objects from the Washington RGB-D dataset. The low resolution of the dataset makes the task of 3D object recognition challenging. We used the same experimental setup as [27], and the provided 10 random splits for the test sets. For experimental evaluation, every 5^{th} video frame was subsampled, leaving a total of 120 images per object. In addition, 51 test objects were used by sampling one object per category, each object having 120 images. Our training set therefore consisted of 34,000 images. To better generalize the results, our experiments were run 10-folds. The achieved performance in terms of recognition rates and standard deviations of our method and the compared methods is presented in Table 1. The results in [33] for other methods are reported here for comparison purposes. The results show that the proposed method achieves a higher performance of 83.2%, 82.7% and 88.7% for RGB, Depth only and RGBD object recognition, respectively, and outperforms other reported methods. The second best is achieved by unsupervised dictionary learning based technique, SP+HMP, with a recognition rate of 87.5% for RGBD object recognition. The results suggest that evolutionary algorithm based evolutionary feature learning provides a higher performance even when applied to low resolution 3D data. Note that the proposed technique uses raw 3D images for feature learning, while the method in [27] uses surface normals and gray scale images as additional features. These features are learnt with unsupervised methods based on sparse coding. The shortcoming of sparse coding is that for large input dimensions, it does not scale well in terms of speed.

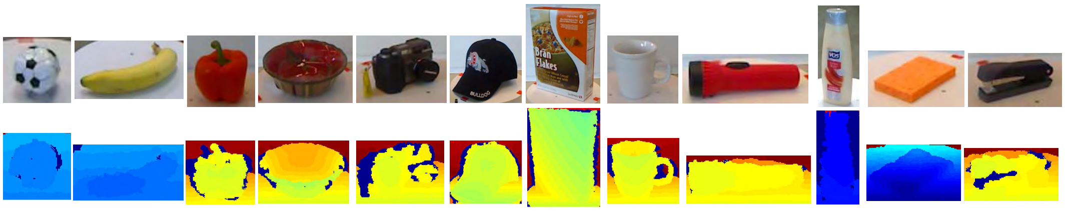


FIGURE 5. Sample images of the objects from the Washington RGB-D (low resolution) dataset. The first row shows RGB images, while the second row shows the corresponding 3D images.

TABLE 1. Comparison of our proposed EFL technique on Washington RGBD Object dataset with state-of-the-art methods.

Compared Techniques	RGB	Depth (D)	RGB-D
Linear SVM [42]	74.31 ± 3.3	53.1 ± 1.7	81.9 ± 2.8
Kernel SVM [42]	74.5 ± 3.1	64.7 ± 2.2	83.9 ± 3.5
Random Forest [42]	74.7 ± 3.6	66.8 ± 2.5	79.6 ± 4.0
DCRN [33]	80.8 ± 4.2	78.9 ± 3.8	86.8 ± 3.3
DKL+SVM [46]	77.7 ± 1.9	78.8 ± 2.7	86.2 ± 2.1
SP+HMP [27]	82.4 ± 3.1	81.2 ± 2.3	87.5 ± 2.9
Proposed Method	83.2 ± 4.1	82.7 ± 2.7	88.7 ± 3.8

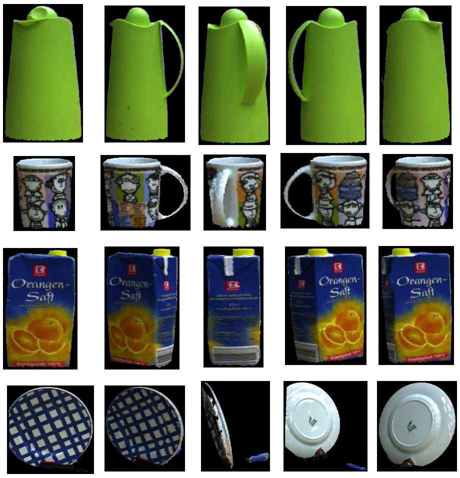


FIGURE 6. Sample images of the objects from CIN 2D3D dataset.

2) CIN 2D3D OBJECT RECOGNITION DATASET

The CIN dataset contains 3 to 14 household objects in 18 different categories. Fig. 6 shows sample images of the objects from the CIN 2D3D object dataset. For performance evaluation, we follow an experimental setup similar to [27] and [43]. The training and test images are randomly selected. Six objects from each category are used as training set, while remaining objects are used for testing. In addition, 18 views per object are selected for training and testing. The training set contains a total of 1476 views of 82 objects. The test set consists of 1332 views of 74 objects. For better generalization of the results, our experiments were run 10-folds. The recognition results and comparison with the recent state-of-the-art methods are reported in Table 2. The results in [27] for other methods are reported here for comparison purposes. It can be noted that the proposed technique outperforms the other methods by achieving a higher performance of 75.2% for 3D object (Depth) recognition and 83.6% for RGBD object recognition. These results clearly demonstrate the discriminative properties of the automatically learnt evolutionary features.

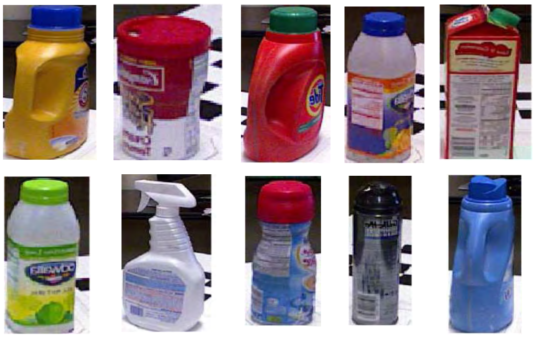


FIGURE 7. Sample images of the objects from Willow 2D3D dataset.

TABLE 2. Comparison of our proposed EFL technique on CIN dataset with state-of-the-art methods.

Compared Techniques/Descriptors	Performance (%)
SURF	42.4
PHOG	69.9
Self Similarity	41.7
Color	26.6
SVM+MLP [43] (RGB)	66.6
Proposed Method (RGB)	**68.7**
Shape Distribution	25.4
Shape Index	34.6
Shape Context 3D	55.2
Depth Buffer	72.9
SVM+MLP [43] (Depth)	74.6
Proposed Method (Depth)	**75.2**
SVM+MLP [43] (RGBD)	82.8
Proposed Method (RGBD)	**83.6**

3) WILLOW 2D3D OBJECT DATASET

This dataset contains rigid and textured household objects from the Willow and Challenge for training and testing, respectively. There are 35 objects in the training and test data set. Objects have been captured from different views [44].

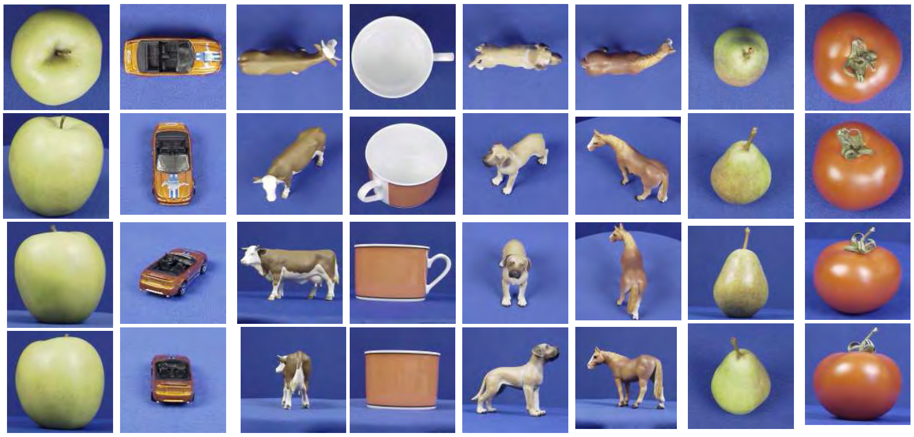


FIGURE 8. Sample images of the objects from ETH-80 dataset.

TABLE 3. Comparison of our proposed EFL technique on willow dataset with state-of-the-art methods.

Methods	Precision	Recall
ICRA12 [44]	96.7	97.4
SP+HMP [27]	97.4	100
Proposed Method	96.8	100

Fig. 7 shows example images of objects from the Willow 2D3D dataset. For performance evaluation, we follow an experimental setup similar to [27] and [44]. To better generalize the results, our experiments were run 10-folds.

Table 3 reports the recognition results in terms of precision/recall and comparison with recent state-of-the-art methods. The results in [27] for other methods are reported here for comparison purposes. It can be noted that the achieved precision is comparable with other methods, whilst recall is on par with [27].

a: ETH-80 OBJECT DATASET

The ETH-80 dataset contains images of eight object categories which include apples, cars, cows, cups, dogs, horses, pears and tomatoes. Fig. 8 shows eight object categories in the ETH-80 dataset. Each object category further includes ten subcategories such as different types of cups or different breeds of horses and cows. Each subcategory has images under 41 orientations. For performance evaluation, we followed the experimental setup outlined in [47] and [48]. For each object, five subcategories are selected for training and the remaining five are used for testing. The achieved performance in terms of recognition rates and standard deviations against other reported methods is presented in Table 4. The results in [48] for other methods are reported here for comparison purposes. The proposed EFL technique achieves a comparable performance on ETH-80 dataset which is in par with BCF [48] and Kernel-edit [49].

TABLE 4. Performance evaluation of state-of-the-art techniques on ETH-80 dataset.

Compared Techniques/Descriptors	Recognition Accuracy (%)
Color Histogram [45]	64.86
PCA gray [45]	82.99
PCA masks [45]	83.41
SC + DP [45]	86.40
IDSC + DP [50]	88.11
IDSC + Morphological strategy [51]	88.04
Height Function [52]	88.72
Robust Symbolic [53]	90.28
Kernel-edit [49]	91.33
BCF [48]	91.49
Proposed Method	91.29

VI. ABLATIVE ANALYSIS

The two parameters that determine the performance and the computational complexity of the proposed EFL algorithm are the number of generations t and size of the population $P(t)$. This is because their product yields the number of fitness function evaluations. In the following, we study the effect of these two parameters using an ablative analysis on the ETH-80 object dataset.

A. EFFECT OF GENERATIONS

To study the effect of the number of generations t on the performance of the proposed algorithm, we varied the generation size from 100 to 1000 for population $P(t)$ sizes ranging from 20 to 100 individuals. Fig. 9 shows the fitness values computed as a function of $P(t)$ and t. Note that for t equal to or above 700, the fitness function starts to converge for all $P(t)$. A population size of 100 achieves the lowest fitness score at the end of G computations. The results suggest that the generation size is an important element of the proposed algorithm, as smaller values of G result into higher fitness scores for the individuals.

B. EFFECT OF POPULATION SIZE

We also studied the effect of population size $P(t)$ on the performance of the proposed algorithm. With the number of

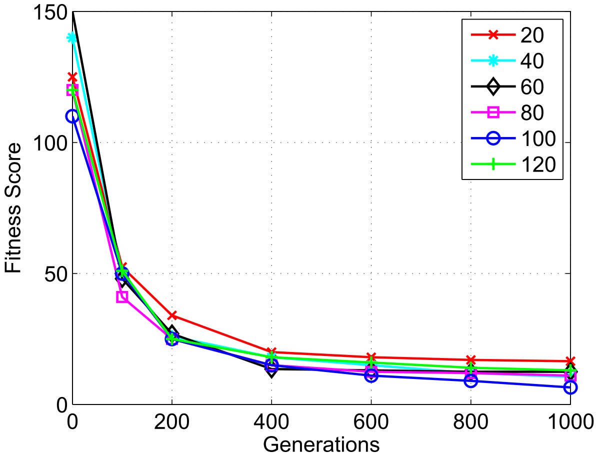


FIGURE 9. Fitness score computed as a function of the the number of generations t and the population size $P(t)$. Note that a population size of 100, achieves the lowest fitness score for G=1000. Figure best viewed in color.

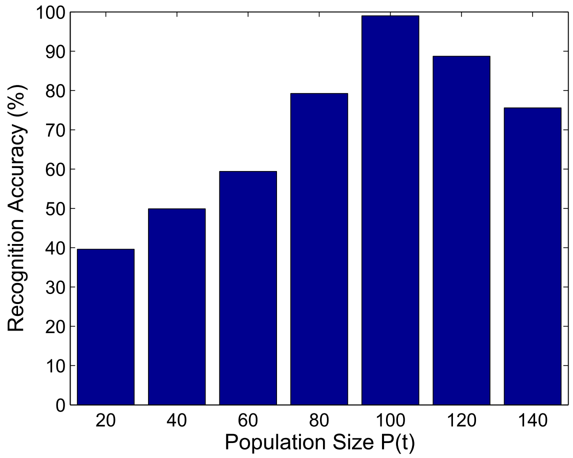


FIGURE 10. Recognition rate computed as a function of the population size $P(t)$. Note that a population size of 100, achieves the highest recognition performance.

generations set to 1000 (Sec. VI-A), the population size $P(t)$ was varied from 20 to 120 and recognition rate computed for each $P(t)$. Fig. 10 shows the recognition performance achieved by the proposed technique for different population sizes. The recognition rate is seen to increase as the population size is varied from 20 to 100. The proposed technique achieves the highest recognition rate of 99% for the population size of 100. Note that the recognition rate starts to decrease for the population sizes above 100. These results suggest that the selection of accurate population size plays an important role on the performance of the proposed EFL technique.

C. COMPUTATIONAL COMPLEXITY OF EFL

From a computational complexity point of view, the most expensive part of the proposed algorithm is the evaluation of the fitness function for a given individual. N such evaluations are performed in each generation. Thus the product $N \cdot G$ determines the computational complexity of the proposed technique. It should be noted that within each generation, the computations for each individual are independent of the other individuals. This inherent computational parallelism can be exploited to achieve very efficient implementations with GPU architectures.

VII. CONCLUSION

In this paper, we propose a novel Evolutionary Feature Learning (EFL) algorithm for the challenging task of 3D object recognition. The proposed EFL adopts a smart search strategy to learn the best features in a large feature space from raw 3D data. Irrelevant and redundant features are omitted based on their fitness score. Only the best candidate solution, termed here as Evolutionary Feature (EF), is selected for each input image. In contrast to existing automatic feature learning methods, the proposed EFL requires neither data pre-processing, defining a large sets of parameters or additional features to achieve superior performance. This has been validated through an extensive evaluation on three publicly-available Washington RGB-D (low resolution 3D video), CIN 2D3D, Willow 2D3D and ETH-80 object datasets. Our experimental evaluations against existing state-of-the-art methods show that the proposed method consistently achieves good performance on all these datasets.

VIII. ACKNOWLEDGMENT

The author would like to thank NVIDIA for their Tesla K40 GPU donation.

REFERENCES

[1] H. Al-Sahaf, A. Al-Sahaf, B. Xue, M. Johnston, and M. Zhang, "Automatically evolving rotation-invariant texture image descriptors by genetic programming," *IEEE Trans. Evol. Comput.*, vol. 21, no. 1, pp. 83–101, Feb. 2017.

[2] Z. Si, H. Yu, and Z. Ma, "Learning deep features for dna methylation data analysis," *IEEE Access*, vol. 4, pp. 2732–2737, 2016.

[3] H. M. Bui, M. Lech, E. Cheng, K. Neville, and I. S. Burnett, "Object recognition using deep convolutional features transformed by a recursive network structure," *IEEE Access*, vol. 4, pp. 10059–10066, 2016.

[4] S. A. A. Shah, M. Bennamoun, and F. Boussaid, "A novel feature representation for automatic 3D object recognition in cluttered scenes," *Neurocomputing*, vol. 205, pp. 1–15, Sep. 2016.

[5] S. A. A. Shah, M. Bennamoun, and F. Boussaid, "A novel 3D vorticity based approach for automatic registration of low resolution range images," *Pattern Recognit.*, vol. 48, no. 9, pp. 2859–2871, 2015.

[6] L. Jiao *et al.*, "A novel image representation framework based on Gaussian model and evolutionary optimization," *IEEE Trans. Evol. Comput.*, vol. 21, no. 2, pp. 265–280, Apr. 2017.

[7] S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, G. Fortino, and M. Buckland, "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis," *IEEE Access*, vol. 4, pp. 9145–9154, 2016.

[8] C.-C. Chang and T.-Y. Lin, "Linear feature extraction by integrating pairwise and global discriminatory information via sequential forward floating selection and kernel QR factorization with column pivoting," *Pattern Recognit.*, vol. 41, no. 4, pp. 1373–1383, 2008.

[9] G. Ghinea, R. Kannan, and S. Kannaiyan, "Gradient-orientation-based PCA subspace for novel face recognition," *IEEE Access*, vol. 2, pp. 914–920, 2014.

[10] H.-C. Chang, Y.-P. Chen, T.-K. Liu, and J.-H. Chou, "Solving the flexible job shop scheduling problem with makespan optimization by using a hybrid Taguchi-genetic algorithm," *IEEE Access*, vol. 3, pp. 1740–1754, 2015.

[11] P.-H. Kuo, T.-H. S. Li, Y.-F. Ho, and C.-J. Lin, "Development of an automatic emotional music accompaniment system by fuzzy logic and adaptive partition evolutionary genetic algorithm," *IEEE Access*, vol. 3, pp. 815–824, 2015.

[12] S. Ding, H. Li, C. Su, J. Yu, and F. Jin, "Evolutionary artificial neural networks: A review," *Artif. Intell. Rev.*, vol. 39, no. 3, pp. 251–260, 2013.

[13] T. Kawaguchi and T. Baba, "3-D object recognition using a genetic algorithm," in *Proc. IEEE Int. Symp. Circuits Syst. Connecting World (ISCAS)*, vol. 3. May 1996, pp. 321–324.

[14] G. Bebis, S. Louis, and S. Fadali, "Using genetic algorithms for 3-D object recognition," in *Proc. 11th Int. Conf. Comput. Appl. Ind. Eng.*, 1998, pp. 13–16.

[15] S. A. A. Shah, M. Bennamoun, and F. Boussaid, "Keypoints-based surface representation for 3D modeling and 3D object recognition," *Pattern Recognit.*, vol. 64, pp. 29–38, Apr. 2017.

[16] D. Zhang, "Harmonic shape images: A 3 D free-form surface representation and its applications in surface matching," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, USA, 1999.

[17] D. Zhang and M. Hebert, "Harmonic maps and their applications in surface matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 1999, pp. 524–530.

[18] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1383–1392, 2009.

[19] F. Tombari, S. Salti, and L. Stefano, "Unique signatures of histograms for local surface description," in *Computer Vision* (Lecture Notes in Computer Science), vol. 6313, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010, pp. 356–369. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-15558-1_26

[20] S. A. A. Shah, M. Bennamoun, F. Boussaid, and A. A. El-Sallam, "A novel local surface description for automatic 3D object recognition in low resolution cluttered scenes," in *Proc. Int. Conf. Comput. Vis. (ICCV) Workshop*, Jun. 2013, pp. 638–643.

[21] D. Hulin and M. Troyanov, "Mean curvature and asymptotic volume of small balls," *Amer. Math. Monthly*, vol. 110, no. 10, pp. 947–950, 2003.

[22] N. Gelfand, N. J. Mitra, L. J. Guibas, and H. Pottmann, "Robust global registration," in *Proc. Symp. Geometry Process.*, vol. 2. 2005, pp. 1–10.

[23] U. Clarenz, M. Rumpf, and A. Telea, "Robust feature detection and local classification for surfaces based on moment analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 10, no. 5, pp. 516–524, Sep. 2004.

[24] U. Clarenz, M. Griebel, M. Rumpf, M. A. Schweitzer, and A. Telea, "Feature sensitive multiscale editing on surfaces," *Vis. Comput.*, vol. 20, no. 5, pp. 329–343, 2004.

[25] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3D object recognition in cluttered scenes with local surface features: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2270–2287, Nov. 2014.

[26] S. Li, Z.-Q. Liu, and A. B. Chan, "Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network," *Int. J. Comput. Vis.*, vol. 113, no. 1, pp. 19–36, 2015.

[27] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for RGB-D based object recognition," in *Experimental Robotics*. Springer, 2013.

[28] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[29] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[30] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.

[31] H. Munawar, B. Mohammed, and S. An, "Deep reconstruction models for image set classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 713–727, Apr. 2015.

[32] D.-C. He and L. Wang, "Texture features based on texture spectrum," *Pattern Recognit.*, vol. 24, no. 5, pp. 391–399, 1991.

[33] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 665–673.

[34] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[35] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2009, pp. 448–455.

[36] E. P. Ijjina and K. M. Chalavadi, "Human action recognition using genetic algorithms and convolutional neural networks," *Pattern Recognit.*, vol. 59, pp. 199–212, Nov. 2016.

[37] Y. D. Fougerolle, J. Gielis, and F. Truchetet, "A robust evolutionary algorithm for the recovery of rational gielis curves," *Pattern Recognit.*, vol. 46, no. 8, pp. 2078–2091, 2013.

[38] S. A. Stanhope and J. M. Daida, "Genetic programming for automatic target classification and recognition in synthetic aperture radar imagery," in *Evolutionary Programming VII*. Berlin, Germany: Springer, 1998, pp. 735–744.

[39] I. Atmosukarto, L. G. Shapiro, and C. Heike, "The use of genetic programming for learning 3d craniofacial shape quantifications," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2010, pp. 2444–2447.

[40] C. B. Perez and G. Olague, "Evolutionary learning of local descriptor operators for object recognition," in *Proc. 11th Annu. Conf. Genetic Evol. Comput.*, 2009, pp. 1051–1058.

[41] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 1150–1157.

[42] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2011, pp. 1817–1824.

[43] B. Browatzki, J. Fischer, B. Graf, H. H. Bülthoff, and C. Wallraven, "Going into depth: Evaluating 2D and 3D cues for object classification on a new, large-scale object dataset," in *Proc. ICCV Workshops*, Nov. 2011, pp. 1189–1195.

[44] J. Tang, S. Miller, A. Singh, and P. Abbeel, "A textured object recognition pipeline for color and depth image data," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2012, pp. 3467–3474.

[45] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2003, pp. 1–7.

[46] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. 24th IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 821–826.

[47] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. CVPR*, Jun. 2010, pp. 2567–2573.

[48] X. Wang, B. Feng, X. Bai, W. Liu, and L. J. Latecki, "Bag of contour fragments for robust shape classification," *Pattern Recognit.*, vol. 47, no. 6, pp. 2116–2125, 2014.

[49] M. R. Daliri and V. Torre, "Shape recognition based on kernel-edit distance," *Comput. Vis. Image Understand.*, vol. 114, no. 10, pp. 1097–1103, 2010.

[50] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, Feb. 2007.

[51] R.-X. Hu, W. Jia, Y. Zhao, and J. Gui, "Perceptually motivated morphological strategies for shape retrieval," *Pattern Recognit.*, vol. 45, no. 9, pp. 3222–3230, 2012.

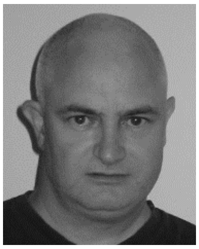
[52] J. Wang, X. Bai, X. You, W. Liu, and L. J. Latecki, "Shape matching and classification using height functions," *Pattern Recognit. Lett.*, vol. 33, no. 2, pp. 134–143, 2012.

[53] M. R. Daliri and V. Torre, "Robust symbolic representation for shape recognition and retrieval," *Pattern Recognit.*, vol. 41, no. 5, pp. 1782–1798, 2008.

SYED AFAQ ALI SHAH received the Ph.D. degree in computer vision and machine learning from The University of Western Australia, Crawley, Australia. He is currently a Research Associate with the School of Computer Science and Software Engineering, The University of Western Australia. His research interests include deep learning, 3-D object/face recognition, 3-D modeling, and image processing. He received the Start Something Prize for Research Impact through Enterprise for 3-D Facial Analysis Project funded by the Australian Research Council.

MOHAMMED BENNAMOUN received the M.Sc. degree in control theory from Queen's University, Kingston, ON, Canada, and the Ph.D. degree in computer vision from Queensland University of Technology (QUT), Brisbane, Australia. He was the Director of the University Center, QUT, i.e., the Space Centre for Satellite Navigation from 1998 to 2002. He is currently a Winthrop Professor and has been the Head of the School of Computer Science and Software Engineering, The University of Western Australia, Perth, Australia, from 2007 to 2012. He is a co-author of the book *Object Recognition: Fundamentals and Case Studies* (Springer-Verlag, 2001) and the edited book on Ontology Learning and Knowledge Discovery Using the Web published in 2011. He has authored or co-authored over 200 journal and conference publications and secured highly competitive national grants from the Australian Research Council (ARC). Some of these grants were in collaboration with industry partners (through the ARC Linkage Project scheme) to solve real research problems for industry, including, Swimming Australia, the West Australian Institute of Sport, Beaulieu Pacific (a textile company), and AAM-GeoScan.

FARID BOUSSAID received the M.S. and Ph.D. degrees in microelectronics from the National Institute of Applied Science, Toulouse, France, in 1996 and 1999, respectively. He joined as a Post-Doctoral Research Fellow with Edith Cowan University, Perth, Australia, and a member of the Visual Information Processing Research Group in 2000. He joined The University of Western Australia, Crawley, Australia, in 2005, where he is currently a Professor. His current research interests include smart CMOS vision sensors and image processing.



LYNDON WHILE received the B.Sc.Eng. and Ph.D. degrees from the Imperial College, London, U.K., in 1985 and 1988, respectively. He is currently a Senior Lecturer with the School of Computer Science and Software Engineering, The University of Western Australia, Perth, Australia. His current research interests include evolutionary algorithms, multiobjective optimization, and the semantics and implementation of functional programming languages.