applied sciences

MDPI

Article

Face Recognition Based on Deep Learning and FPGA for Ethnicity Identification

Ahmed Jawad A. AlBdairi [1,2], Zhu Xiao [1], Ahmed Alkhayyat [3], Amjad J. Humaidi [4], Mohammed A. Fadhel [5], Bahaa Hussein Taher [1,5], Laith Alzubaidi [6,*], José Santamaría [7,*] and Omran Al-Shamma [8]

1 College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China; ahmed_albdairi@hnu.edu.cn (A.J.A.A.); zhxiao@hnu.edu.cn (Z.X.); ghrabiuk@gmail.com (B.H.T.)
2 Department of Computer Science, University of Babylon, Babylon 51001, Iraq
3 College of Technical Engineering, The Islamic University, Najaf 54001, Iraq; ahmedalkhayyat85@gmail.com
4 Control and Systems Engineering Department, University of Technology-Iraq, Baghdad 00964, Iraq; amjad.j.humaidi@uotechnology.edu.iq
5 College of Computer Science and Information Technology, University of Sumer, Rifai 64005, Iraq; mohammed.a.fadhel@uoitc.edu.iq
6 School of Mechanical, Medical and Process Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia
7 Department of Computer Science, University of Jaén, 23071 Jaén, Spain
8 AlNidhal Campus, University of Information Technology & Communications, Baghdad 10001, Iraq; o.al_shamma@uoitc.edu.iq
* Correspondence: laith.alzubaidi@hdr.qut.edu.au (L.A.); jslopez@ujaen.es (J.S.)

Abstract: In the last decade, there has been a surge of interest in addressing complex Computer Vision (CV) problems in the field of face recognition (FR). In particular, one of the most difficult ones is based on the accurate determination of the ethnicity of mankind. In this regard, a new classification method using Machine Learning (ML) tools is proposed in this paper. Specifically, a new Deep Learning (DL) approach based on a Deep Convolutional Neural Network (DCNN) model is developed, which outperforms a reliable determination of the ethnicity of people based on their facial features. However, it is necessary to make use of specialized high-performance computing (HPC) hardware to build a workable DCNN-based FR system due to the low computation power given by the current central processing units (CPUs). Recently, the latter approach has increased the efficiency of the network in terms of power usage and execution time. Then, the usage of field-programmable gate arrays (FPGAs) was considered in this work. The performance of the new DCNN-based FR method using FPGA was compared against that using graphics processing units (GPUs). The experimental results considered an image dataset composed of 3141 photographs of citizens from three distinct countries. To our knowledge, this is the first image collection gathered specifically to address the ethnicity identification problem. Additionally, the ethnicity dataset was made publicly available as a novel contribution to this work. Finally, the experimental results proved the high performance provided by the proposed DCNN model using FPGAs, achieving an accuracy level of 96.9 percent and an F1 score of 94.6 percent while using a reasonable amount of energy and hardware resources.

Keywords: face recognition; ethnicity identification; deep learning; real-time; HPC; FPGA; GPU

Citation: AlBdairi, A.J.A.; Xiao, Z.; Alkhayyat, A.; Humaidi, A.J.; Fadhel, M.A.; Taher, B.H.; Alzubaidi, L.; Santamaría, J.; Al-Shamma, O. Face Recognition Based on Deep Learning and FPGA for Ethnicity Identification. *Appl. Sci.* **2022**, *12*, 2605. https://doi.org/10.3390/app12052605

Academic Editors: Monica Perusquia Hernandez and Antonio Fernández-Caballero

Received: 7 December 2021
Accepted: 1 March 2022
Published: 2 March 2022

1. Introduction

Nowadays, surveillance systems play a significant role in freely available security [1]. The resulting videos of these surveillance systems [2,3] became more straightforward to analyze due to the progress in Artificial Intelligence (AI) [4] and its adoption in the field of CV (CV) [1]. Recently, numerous works have focused on the event detection problem using these videos being the necessary potential for identifying and localizing specific spatio-temporal patterns. In particular, the issue of person identification, which recently attracted the interest of researchers, is an additional problem within video surveillance [5].

This problem deals with the recognition of a person picked up from a whole image dataset in which single/multiple cameras were considered for the data acquisition. It is a significant issue for both human–computer interaction and surveillance systems to simplify the identification procedure using a sizable volume of both images and videos.

In race analysis, facial features have become a standard topic in CV and FR (FR) [6–8]. The field of HRR has shown considerable growth in conjunction with the increasing globalization in several real-world applications such as public security, border control, and customs checks. In the field of physical anthropology, HRR is also considered a relevant branch of research. Society, genes, the environment, and other factors have a significant impact on facial features. Due to various gene fragments, it is not easy to distinguish between the particular genes of one race group against another one, which has become a complex challenge. Thus, facial features seem similar across various races [9]. Besides identifying a suitable emergency supporter, recognizing the human race becomes very useful for surveillance systems in many situations. For instance, race-targeted pharmacogenomics and race-based medicine support race-information usage in diagnosing and treating different medical cases that have specific responses based on the organisms of diverse races. Improving the person's identification accuracy and narrowing down the possible matches can be achieved by embedding the race information (e.g., using soft biometrics information) into video surveillance systems. In addition, targeted advertising and human–computer interaction systems can also use racial information to offer employers ethnically convenient services, then preventing the offending risk present in many cultural taboos. Visagisme is an innovative concept that takes place in optometry and fashion fields. In the previous paragraph, the authors were talking about applications of surveillance. The role of HRR helps improve the surveillance and other potential applications of HRR how facial features can be useful in this application. This paper is focused on the Human Race Recognition (HRR) problem by means of video surveillance.

Moreover, video clustering targeted at classification tasks is another highly demanding application of HRR. However, several works on coding and HRR were conducted to address this challenge. In the study of race learning from the face, Fu et al. [10] presented an in-depth review including several up-to-the-date techniques. According to the review, the challenge may be answered in two directions: single model HRR and multiple model HRR. In the single model HRR, they made an effort to extract both the local discriminative areas and the appearance features. They combined features from both gait and face fusion in the second direction, or 3D and 2D information.

Recently, researchers have attracted significant attention to Deep Learning (DL) [11–14] owing to its numerous applications in speech processing [15], natural language processing [16], and CV [17,18]. In video recognition [19] and large-scale images, a model of DL so-called convolutional neural network (CNN) has lately attained several encouraging results. Simonyan and Zisserman [20] introduced in 2015 the VGG model, which accomplished an extremely well-behaved performance with ImageNet and became used-widely in different studies of CV.

On the other hand, one of the most significant drawbacks of DL is that it requires more processing time due to a large amount of data it must deal with. As a result, they employ various parallel hardware architecture platforms, such as FPGA and GPU, to overcome these difficulties. Field programmable gate array (FPGA) technology is an emerging hardware technology developed in recent years [21]. It has shown strong potential and capabilities for embedded intelligence (EI) applications. As compared to Graphics processing unit (GPU) and CPU technologies, the core benefit of the FPGA is its capability to customize the hardware implementation to achieve the applicability of the specified needs of the EI algorithm. Thus, the FPGA becomes an effective hardware solution for more significant energy saving and higher computation processing [22]. For example, the FPGA can use proper quantized memory representations along with fixed-point computations instead of floating-point. Compared to GPU and CPU systems, the FPGA dedicated hardware further benefits the FPGA-based systems. Therefore, the integration of embedded systems/devices with decision-making algorithms, Machine Learning (ML) techniques, as well as methods that are commonly achieved on central computation systems such as cloud computing was introduced by several researchers.

The following are the main contributions provided in this research:

- A network proposal was used to group the human face by its race, specifically from Chinese, Pakistani, and Russian citizens;
- The collected image dataset is a unique one in its kind and is publicly available upon request;
- Several HPC hardware accelerator architectures have been tested, such as FPGAs and GPUs;
- The proposed network was compared against four different pre-trained CNN models: ResNet50, DenseNet, AlexNet, and GoogleNet;
- The experimental results were enriched by reporting the power usage and execution time achieved by the proposed network in order to facilitate its future easy development as a mobile embedded system.

The structure of the paper is as follows. Section 2 is devoted to introducing the background and fundamentals regarding the technologies used in this research. Next, Sections 3 and 4 are aimed at describing in detail the deployment of DL in FPGAs and the DL-based ethnicity identification proposed in this work, respectively. Section 5 is focused on the experimental results reported by our proposal and its competitors from the state-of-the-art (SoTA). A summary of the conducted study and some future research suggestions is finally introduced in Section 6.

2. Background and Related Work

Classification of ethnics and races using facial images was introduced in a detailed survey [10]. The survey discussed the race classification challenge from the analytical and fundamental race sympathetic using interdisciplinary knowledge such as anthropometry, cognitive neuroscience, and psychology, highlighting the different representations of racial features. In addition, the survey compared the most related works in the automated race classification field based on facial images.

From the used racial features viewpoint, the up-to-the-date techniques employ either local features, global features, chromatic information, or a mixture of these techniques [23,24]. Techniques of local features-based classify the race using lower-level features such as Gabor filters of gradient directions or histograms. Techniques of chromatic-based are commonly based on the tone of skin and are extremely illumination sensible. The techniques of global-based are the most used-widely and take advantage of the interrelation between various facial areas to build racial belonging. In contrast, hybrid techniques mix all or some of the above techniques to achieve the optimum representation for race classification [25].

2.1. Deep CNNs

Recently, owing to their encouraging performance, deep CNNs have been utilized considerably in CV. For example, Zhang et al. [26] achieved a well-behaved performance using an innovative feature learning technique to classify halftone images. For feature extraction, this technique uses unsupervised learning (stacked spare auto-encoders (SAE)), while for image classification, it uses supervised learning (SoftMax regression) for fine-tuning the deep network. Wei et al. [27] presented Hypotheses-CNN-Pooling (as a flexible deep CNN model) to classify images of multiple labels. This framework uses an undefined number of object segment hypotheses as an input. Next, each hypothesis is connected to the shared CNN. Lastly, the max-pooling is summed with the hypotheses outputs to produce the final multi-label predictions. Numerous issues are present in face-based applications, such as facial expression analysis [28], face alignment [29], and face detection [30]. Li et al. [31] introduced in 2015 a cascade model constructed on a CNN that has a robust discriminative capability but keeping well-behaved performance for dealing with the variations in visual characteristics issue such as those owing to lighting, expression, and pose, in the practical FR. Park et al. [32], 2017, developed deep networks to align faces based on recurrent regression and facial landmark features. In contrast, Chen et al. [33] presented an intelligent technique for detecting smiles based on CNNs, since in our daily life, a smile is the most popular facial expression. However, facial expression analysis is also requested by other issues such as robotics, human–computer interface, lie detection, and medical assessment.

One of the crucial issues in video analysis is to identify/re-identify persons through images using either a single camera over time or multiple cameras. Ahmed et al. [34], in 2015, presented a corresponding similarity metric to re-identify persons and a technique for learning features in parallel. The authors achieved well-behaved results using a deep convolutional model with specifically designed layers to highlight the re-identification issue. An additional issue in video analysis is to track the target visually. This issue has several applications, such as video surveillance, augmented reality, and vehicle navigation. A CNN with outstanding performance was introduced to deal with such an issue. It is beneficial in real-time visual tracking [35]. Recently, one more issue in video analysis that magnetized great attention is the recognition of human activities. Ronao and Cho [36] introduced a CNN for performing the recognition of human activities effectively and efficiently. In their work, 1D time-series signals and the inherent activity characteristics are exploited using smartphone sensors. Based on various experimental datasets, they have achieved an excellent performance.

2.2. Transfer Learning

It is difficult to train a whole CNN from scratch owing to the limited undersized datasets. Alternatively, Simonyan and Zisserman [20] presented a CNN, namely VGG, which achieved an accuracy of 92.7% with ImageNet that has up to 1000 classes and over 14 million images. VGG-16 and VGG-19 are the two forms of the trained VGG model, where their parameters and structures are available online freely. Pre-trained models, fine-tuning CNNs, and a fixed feature extractor are the main three scenarios of the latter learning approach, named transfer learning (TL). In comparing these models, fine-tuning showed the most popular with different models such as ResNet50, DenseNet GoogLeNet, and VGG.

The RR-VGG model, which used VGG-16 inside, is presented in the coming sections of this study. Several works leveraged the VGG structure for performing transfer learning in various challenges. For extracting useful features, Paul et al. [37] utilized the trained VGG model to detect lung adenocarcinoma, while for image classification, different classifiers were used. Long et al. [38] adopted several pre-trained CNN models (GoogleNet, AlexNet, and VGG) inside a fully convolutional model. For segmentation purposes, they fine-tuned the model. Applying the pre-trained VGG models in problems of computer-aided detection was studied by Hoo-Chang et al. [39] and attained inspiring results. Transferring a pre-trained model such as VGG becomes more appropriate than using other techniques.

On the other hand, multiple small filters (e.g., 3×3) are used to build larger filters (e.g., 5×5) in the VGG-16 model, which has three fully connected and 13 convolutional layers. Thus, the whole convolutional layers have only a 3×3 filter size. Altogether, processing one 224×224 input image needs a VGG-16 model of 15.5 G multiply-and-accumulates and 138 M weights [40].

One of the most major disadvantages of DL is that it takes greater processing time due to a large amount of data it must deal with. As a consequence, to solve these challenges, they deploy different parallel hardware architectural platforms, such as FPGA and GPU.

3. Deployment of Deep Learning on FPGAs

Networks based on DL techniques can construct accurate decisions while they train/learn by themselves using the provided training data. Due to its capability to maximize parallelism and energy saving, the FPGA showed substantial growth in its usage in DL applications. As compared to GPU, the FPGA has comparatively lower computing, bandwidth, and memory resources. In contrast, it can offer a reasonable accuracy with high throughput. This subsection demonstrates the FPGA-based DL architectures, especially with CNNs.

Consequently, another paper introduced a compressed CNN technique and employed a scalable and fast structure for a low-density FPGA. The image-batching and fixed-point arithmetic techniques are the core methods utilized in this structure. These techniques can save the memory bandwidth and allow the computational memory to concentrate on actual processing works such as object detection and surveillance monitoring. The weights and activations are represented in fixed-point format despite the window size in optimization techniques, where fixed-point quantization is a part of them. In addition, the memory bandwidth can be reduced by executing multiple images in a batch [41]. In contrast, each layer has its scale factor.

Even if the layers of convolution and fully connected are entirely the same, they should be separated into two unlike parts and performed in parallel. The on-chip memory is used to store the intermediate features and the input image. If it is undersize to store the image, then partitioning the image is required and convolved independently. The data dispatch and interconnected are used to store the OFM in an external memory if it is too big to fit into the on-chip memory. The PE (processing element) units run in parallel to compute the whole arithmetic. Figure 1 shows the PE cluster of the convolutional layers composed of the interconnection and matrix network to advance the activations of input and output among the PE.

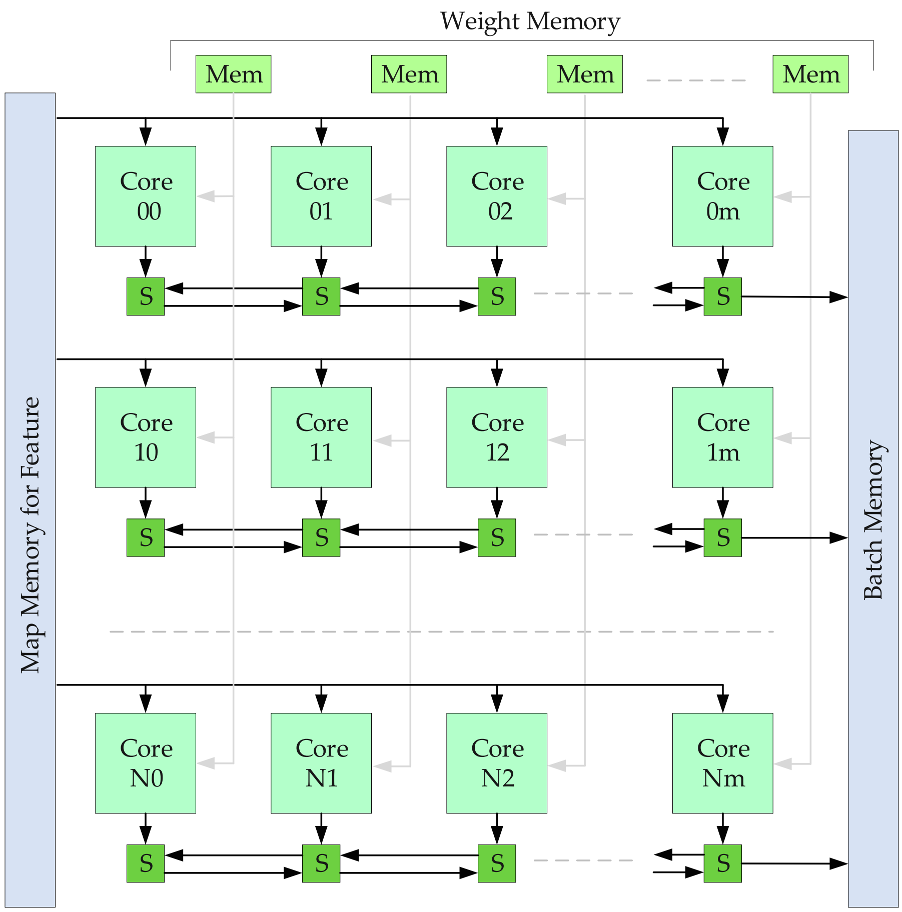


Figure 1. Interconnection between PE (processing element) and S (switch).

The data transfer is based on the type of the next layer; if it is convolution, the data moves back to the feature map memory; if not, then the data move to a fully connected layer, which has a similar structure to the convolution cluster. The switch is two small modules and performs the whole data transfer. For reducing the accuracy loss and improving the performance in the hardware, the weights and activation are represented by an 8-bit mixed fixed-point format. With a similar structure, the throughput increases by focusing on fixed-point arithmetic at the convolutional layer and the image batching at the fully connected layer. Finn is another proposed framework that focuses on the efficient mapping of the binarized neural network to the FPGA [42]. Using standard datasets (SVHM, CIFAR-10, and MNIST), the hardware maintains high classification performance and obtains low power consumption (lower than 25 W) on the ZC706 embedded FPGA platform.

Another work on FPGA that proposed a depth-wise separable CNN (essentially a factorization from conventional convolution into pointwise and depth-wise convolution to build new features) is being studied [43]. The framework can enhance computing speeds and reduce the model parameters, and it has become commonly used in applications of mobile-edge computing such as MobileNet and Xception. Double buffering memory channels are also proposed to handle the data transfer based on a custom computing engine structure. Moreover, the data tiling technique divides the matrix multiplication into small-form factors for applying in the fully connected layer.

In contrast, Figure 2 illustrates the depth-wise convolutional unit, composed of max pooling modules, nonlinearity activation modules, accumulators, and the multiply-accumulate computing unit. The max-pooling modules compute the data down-sampling, while the nonlinearity module carries out the ReLU function. The work of the pipeline structure is an essential technique for such depth-wise CNN. Lastly, lowering the numeric computing precision will lower the consumed power with a slight accuracy loss. Compared to 32-bit floating-point, a framework of 16-bit fixed-point multiplication reduces the power-consuming by 6.2 times, but with a reduction in accuracy of 0.5%. For the image size of $32 \times 32 \times 3$, the work also concluded a reduction in the consumed power of 29.4 times that of GPU. Further, an improvement in the performance of 17.6 times faster than that of the CPU [43].

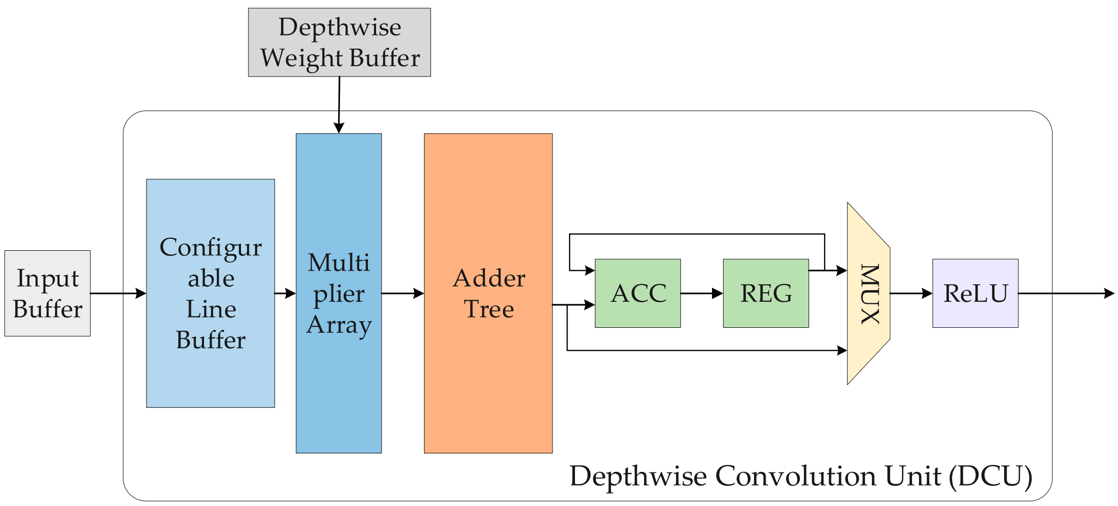


Figure 2. The block diagram of depth-wise convolution unit (DCU).

The reduction in parameter requirements is an additional challenge attracting the researchers' interest. Regarding the hardware accelerator design, LetNet and SqueezeNet are models that utilize techniques of condensed parameter requirements [43]. Another work presented a reduction and size in the convolutional layers' kernels and removed nearly all the fully connected layers to leverage the previous research. A fire module layer replaced all removed layers [44], and it is a combined layer of the expanded convolutional layers and the squeeze layer. The model also utilizes a downsampling technique for reducing the size and preserving high-level accuracy. Calculating the number of parameters is performed by multiplying the number of output channels in the layer by the number of parameters in the individual kernel. In addition, calculating the AC operation depends upon the weight vectors presented in a square dimension of the layer output features [45].

On the other hand, 8-bit encoded image pixels and 8-bit encoded weight vectors are utilized in considering the bit-width issue. Based on the kernel size, a different number of parallel memories is presented in each layer. As compared to the standard network, a reduction in parameters by 11.2 times was achieved using this fire module technique with no impact on the accuracy of classification [43].

In concluding the above works, the FPGAs with flexible platforms to establish various DL applications in brief periods are desirable in embedded DL intelligence compared to application-specific integrated circuits. Performing arithmetic intensive operations become possible using embedded components and rich-set programmable logic cells such as DSP. However, additional techniques for DL are still in development and investigation by the researchers, such as the roofline model, graph partitioning, double buffering, and rearranging memory data. While the previous study [46] proposed a network based on the GPU platform, which consumes more power and is unsuitable for portable embedded systems, this research implements a network for human race classification by taking advantage of parallelism techniques on FPGA (type DE1-SoC) to speed up our classification while consuming less power, making it more suitable for portable embedded systems.

4. The Deep Learning-Based Ethnicity Identification Proposal

4.1. Ethnicity Identification

Our model proposal uses twelve DL layers. Initially, four convolutional layers are employed, each followed by a max-pooling layer. In addition, some of these convolutional layers have a dropout layer inserted after the max-pooling layer for extracting facial features. Next to these convolutional layers, there are two fully connected layers and a drop connect layer as a separator to rid the overfitting problem through the training process. After the drop-connect layer, a flattening layer is inserted to flatten the output before it passes to the fully connected layer. An additional dropout layer is inserted between the two fully connected layers. Finally, a SoftMax output layer is employed to recognize the classes. The total network architecture is shown in Figure 3. Note that it is possible to enlarge the recognized classes (n) to include more nationalities.

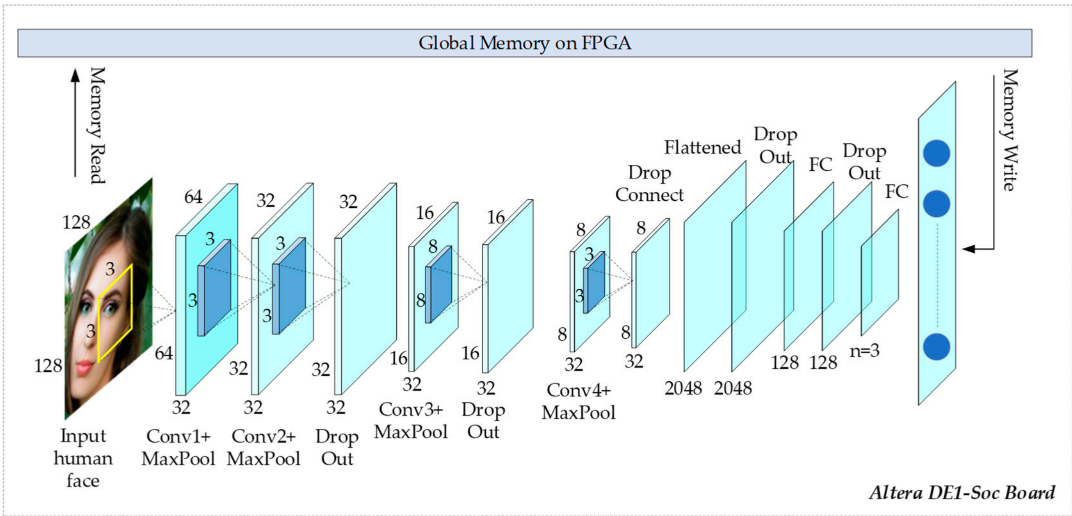


Figure 3. Our proposal system for ethnicity identification based on deep learning.

However, the input image has a size of $128 \times 128 \times 3$ since (3) can be feature maps. In addition, the patch size is 3×3 with unchanged padding in each convolutional layer. In contrast, the stride has a value (1) to make the convolutional layer output nearly similar to the input size. The input size is minimized, bypassing the convolutional layer output to a max-pooling layer and then to the ReLU activation function. The equation of the convolutional layer with the feature map is:

$$f(x)^{j(r)} = max\left(0, b^{j(r)} \sum_{n=1}^{\infty} k^{ij(r)} * x^{i(r)}\right) \tag{1}$$

Within a specific area r, $f(x)^{j(r)}$ is the j^{th} output patch of the convolution layer, $b^{j(r)}$ is the bias of the j^{th} output patch, $k^{ij(r)}$ is the kernel of the convolutional layer between the i^{th} input and j^{th} output patch, the multiplication sign ($*$) means the convolution, and $x^{i(r)}$ is the i^{th} input patch within a specific area r to the convolution layer. Figure 3 shows that the input image is partitioned into small areas; each area has a size of 3×3 based on the window patch size.

As mentioned earlier, the max-pooling layer receives the convolutional layer output. The equation of the max-pooling layer is:

$$f(x)^{i}_{jk} = \underset{0 \leq m,n < sz}{max}\left(x^{i}_{j \cdot sz+m, k \cdot sz+n}\right) \tag{2}$$

The neurons in the i^{th} output patch of $f(x)^i$ pool over a local area of $sz \times sz$ in the i^{th} input patch (x^i). The nonlinearity function ReLU keeps all positive values constant, whereas all negative values are set to zero. As compared to the sigmoid function, the RELU has a better fitting performance [45].

In this work, three dropout layers were used, one separating the two fully-connected layers and the other two are after the second and the third convolutional layers, respectively. The equation of the two fully connected layers is:

$$fc = max\left(0, \sum_i x^{i-1} \cdot w^{i-1,j-1}\right) + max\left(0, \sum_i DOut_{rate}\left(x^i \cdot w^{i,j}\right)\right) \tag{3}$$

Regarding the previous layer, x^{i-1} refers to the neurons, and w^{j-1} refers to the weights. The $DOut_{rate}$ separates the two fully connected layers where the rate is 0.5. Regarding the first connected layer, x^i refers to the neurons, and w^j refers to weights.

Among n different ethics, the n-way softmax predicts the face ethic based on the ConvNet output. The equation of the softmax is:

$$y_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)} \tag{4}$$

where x is the output index in n, which means, for example, the number of classes, and x_i is the input vector that refers to the very significant features utilized for FR to the output layer.

4.2. Dropout Layer

The reported results become inaccurate in the testing phase due to errors in the training phase. Dropout as a robust regularization is utilized to solve the overfitting problem [44]. The concept of the dropout includes dropping out from the network a few neurons that are selected randomly alongside the probability q = $1 - p$. These neurons become free after drop out (i.e., its input and output are discounted). They can learn somewhat helpful on their own, exclusive of depending in large amounts on the remaining neurons to modify their limitations.

Conversely, the inputs and outputs of all patches are calculated as follows before applying dropout:

$$x^{l+1} = w^{l+1} y^l + b^{l+1} \tag{5}$$

$$y^{l+1} = AF\left(x^{l+1}\right) \tag{6}$$

Note that the network layer index is l, the input patch is x^{l+1}, the weight is w^{l+1}, the bias is b^{l+1}, the output patch is y^{l+1}, and the activation layer is AF. After the dropout is executed, the following operations take place:

$$\sigma_i^l \approx Bernolulli(p) \tag{7}$$

$$y^{l^l} = \sigma^l \oplus y^l \tag{8}$$

$$x^{l+1} = w^{l+1} y^{l^l} + b^{l+1} \tag{9}$$

$$y^{l+1} = AF\left(x^{l+1}\right) \tag{10}$$

The layer l, i^{th} neuron, σ_i^l denotes the Bernoulli random variable alongside a probability value of 1. The multiplication of element by element is denoted as $\oplus$.

5. Results

5.1. Training Dataset

Despite the fact that there are several large-scale face image databases accessible online, none of these databases are acceptable for the purpose of the conducted study in our research. Furthermore, 3141 photographs were gathered from a variety of sources. Specifically, 1081, 1021, and 1039 Chinese, Pakistani, and Russian face photos were gathered, respectively. Following the collection of the images, they were processed in order to extract the faces from the whole set of images. Next, the entire dataset was partitioned into three, which include a training set, validation set, and test set at 60%, 10%, and 30%, respectively. A portion of the new dataset is shown in Figure 4.

Figure 4. A portion of the new dataset gathered from three separate countries.

5.2. FPGA Hardware Implementations

Using an FPGA DE1-SoC, a 64-computations array of 16-bit DSP was used to speed the proposed network and other four pre-training models. HPS, the control software utilized in this procedure, and the hardware responsible for the convolutional computations were the two components of this acceleration process. Because of the limits of FPGA fabrics on DE1-Soc, only 13-convolutional layers were employed in order to maximize efficiency while also addressing the restrictions of the architecture. The remaining calculations were carried out in hardware due to the fact that hardware can execute computations far quicker than software.

Unlike the convolution layers (CONV), which are mainly comprised of separate control logic and a parallel adder, the multiplication layers (MUL) serve as the principal computational engines and are connected across all levels, as seen in Figure 5. During the convolution process, the data input is recorded in on-chip buffers, and the outputs of the multiplier are sent to the CONV stage for summing and accumulating. After CONV was completed, the findings were routed to a variety of various on-chip memories, where they were used for the next step of the process.

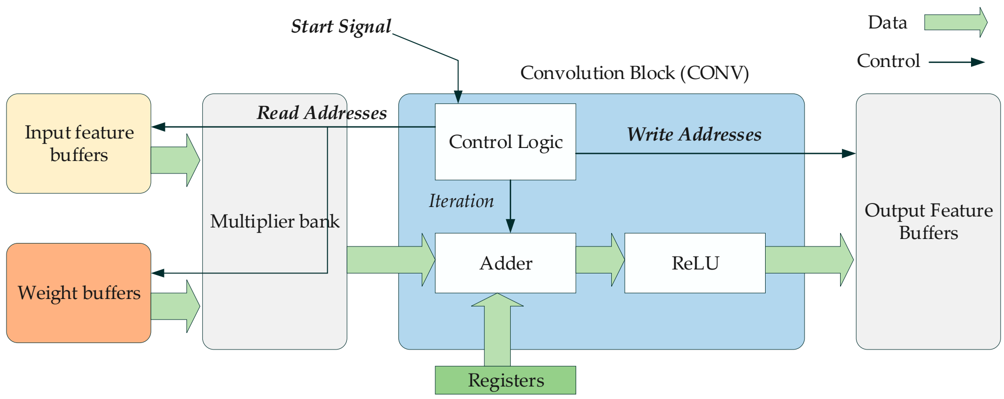


Figure 5. The schematic of a convolutional block within an FPGA.

5.3. A Preliminary Comparitive Analysis

This subsection is devoted to introducing a performance comparison considering several contributions from the SoTA. In recent years, numerous additional studies have attempted to address the classification problems of race [4] and ethnicity [46] by means of DL approaches, e.g., CNNs. Table 1 presents the results provided by several methods of the SoTA dealing with the tackled problem. Note that it was not feasible to carry out a direct numerical comparison against all methods due to several of them were trained and assessed on not accessible private datasets.

Table 1. Performance comparison results.

Method	Dataset	Number of Images	Groups	Accuracy	Hardware Platform
[4]	Internet images	over 175,000	Asian, Afro, Caucasian, and Indian	96.64%	GPU
[6]	VNFaces	6100	Vietnamese and others	88.87%	Not mention
[8]	Celeb-A	over 200,000	predominately Western celebrities	91%	Not mention
[23]	Private	22	WD, RBAL and BD	96%	GPU
Our	Private	3141	Chinese, Pakistani, and Russian	96.9	GPU and FPGA

In [4], the authors developed and made available a large-scale library of over 175,000 photos of faces of celebrities from multiple ethnicities and races, which were used for training and testing. The method was compared against four cutting-edge CNNs on the topic. Another dataset called VNFaces [6] was used to compare the accuracy of RR-CNN and RR-VGG methods, in which accuracy of 88.64 and 88.87 percent, respectively, was achieved.

Previous studies have used datasets consisting mostly of Western celebrities [8] as well (WD, RBAL, BD) [11] and obtained accuracy rates of 91 and 96 percent. It should be noted that GPU was the hardware platform used in all the previous studies. Our proposal provides a new extension of the SoTA by using FPGA, being one of the first successful attempts making use of this specific hardware platform for addressing the real-time classification of human races. The reported results in Table 1 demonstrated the higher performance achieved by our method with an accuracy rate of 96.9 percent.

5.4. Comparison for the Pre-Trained CNNs

A total of four pre-trained CNN models were chosen, in which the last four layers of each approach have been frozen, and with our fully connected layers being utilized to calculate the number of outputs based on the number of classes present in the dataset. Specifically, ResNet50, DenseNet, AlexNet, and GoogleNet were the chosen CNNs. The training was carried out using a system based on a CPU (Intel Core i7-10750H), a GPU (NVIDIA GeForce RTX 2070), and RAM (16 GB). According to the findings, our proposed network offers the best performance in terms of the parameters listed in Table 2, which illustrates the training outcomes of our network against the ones from the other four CNN models. As shown in Figure 6, our network outperformed the competition on Accuracy, F1 score, and other parameters. The accuracy and F1 score both reached a ratio rate of 96.9 percent and 94.6 percent, respectively. It is important to note that the power used by the GPU remains constant throughout workloads since it is intended to function at full capacity, as opposed to the FPGA, which uses just a subset of the reprogrammable logic components and then calculates the power usage by FPGA synthesis tool.

Table 2. Comparison between our proposed network and other CNN pre-train models in term of performance metrics.

Measure	ResNet50	DenseNet	AlexNet	GoogleNet	Our Proposed Network	Derivations
Sensitivity	0.945	0.916	0.723	0.852	0.966	TPR = TP/(TP + FN)
Specificity	0.950	0.977	0.957	0.984	0.956	TNR = TN/(FP + TN)
Precision	0.904	0.953	0.892	0.963	0.966	PPV = TP/(TP + FP)
False Positive Rate	0.049	0.022	0.042	0.015	0.043	FPR = FP/(FP + TN)
Accuracy	0.949	0.957	0.879	0.940	0.969	ACC = (TP + TN)/(P + N)
F1 Score	0.924	0.934	0.799	0.904	0.946	F1 = 2TP/(2TP + FP + FN)

Where: TP: True Positive, TN: True Negative, FP: False Positive and FN: False Negative.

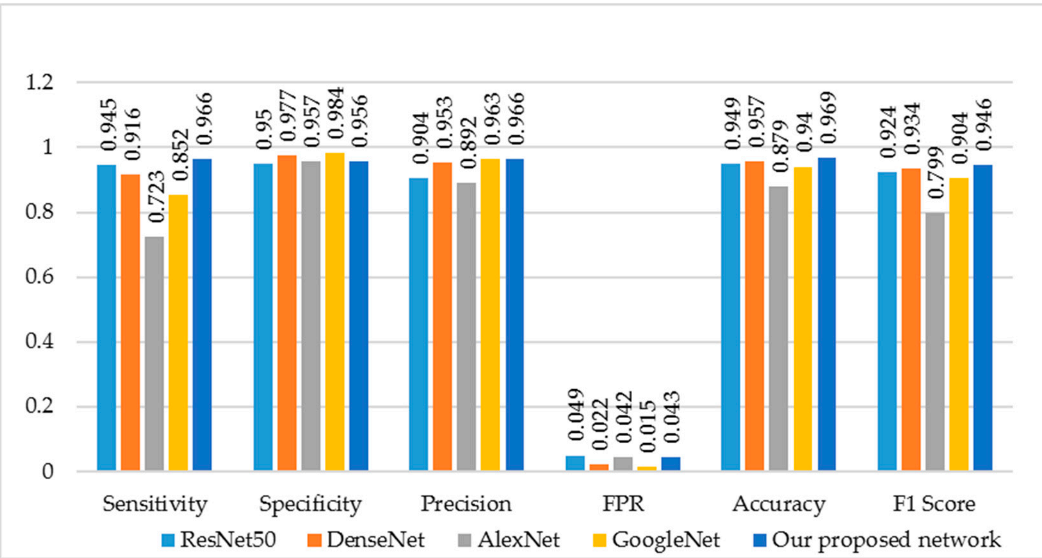


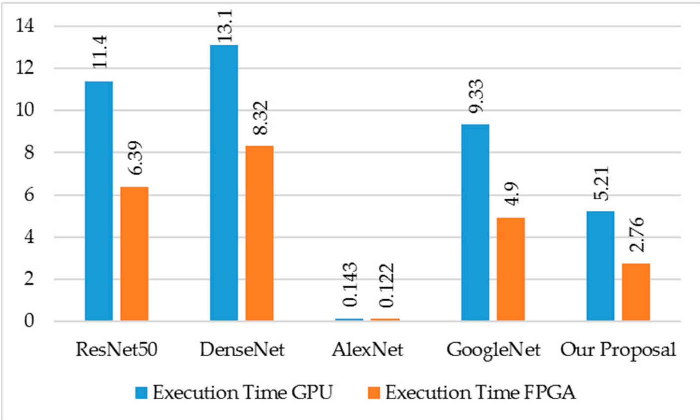
Figure 6. Summary of all models considering all evaluation parameters.

This research is an improvement from the previous study [46], which proposed a network based on GPU platform, which requires more power and is unsuitable for portable embedded systems, but in this work, the network for human race classification is implemented by taking advantage of parallelism techniques on FPGA (type DE1-SoC) to speed up our classification while consuming less power.

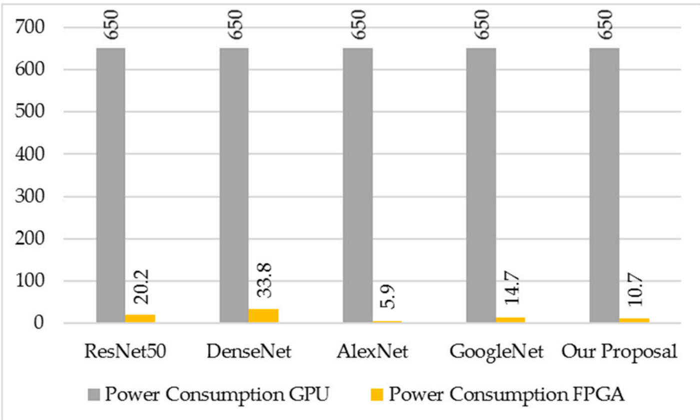
Another comparison was conducted in Table 3 between our proposed network and existing CNN models. AlexNet offers the lowest power usage and execution time due to its basic design. In contrast, our network outperforms in terms of accuracy, F1 score, appropriate power usage in FPGA, and processing time in GPU and FPGA, see Figure 7. The FPGA assessments were based on an acceptable amount of DE1-Soc resources, including total logic elements, total block memory, and total logic registers, as shown in Table 4 and Figure 8.

Table 3. Comparison between our proposed network and other pre-trained CNN models in terms of time execution and power consumption.

	ResNet50		DenseNet		AlexNet		GoogleNet		Our Proposal	
	GPU	FPGA	GPU	FPGA	GPU	FPGA	GPU	FPGA	GPU	FPGA
Execution Time	11.4 s	6.39 s	13.1 s	8.32 s	143 ms	122 ms	9.33 s	4.90 s	5.21 s	2.76 s
Power Consumption	650 w	20.2 w	650 w	33.8 w	650 w	5.9 w	650 w	14.7 w	650 w	10.7 w



(**A**)



(**B**)

Figure 7. (**A**) Time of execution (**B**) Power consumption.

Table 4. Overview of resources for DE1-Soc.

Measure	ResNet50	DenseNet	AlexNet	GoogleNet	Our Proposal
Total logic elements out of 32,070	22,521	30,156	13,117	27,853	25,671
Total block memory out of 4,056,280 bits	217,937	263,319	43,561	64,890	96,734
Total logic registers	73,324	83,224	34,335	57,801	64,523

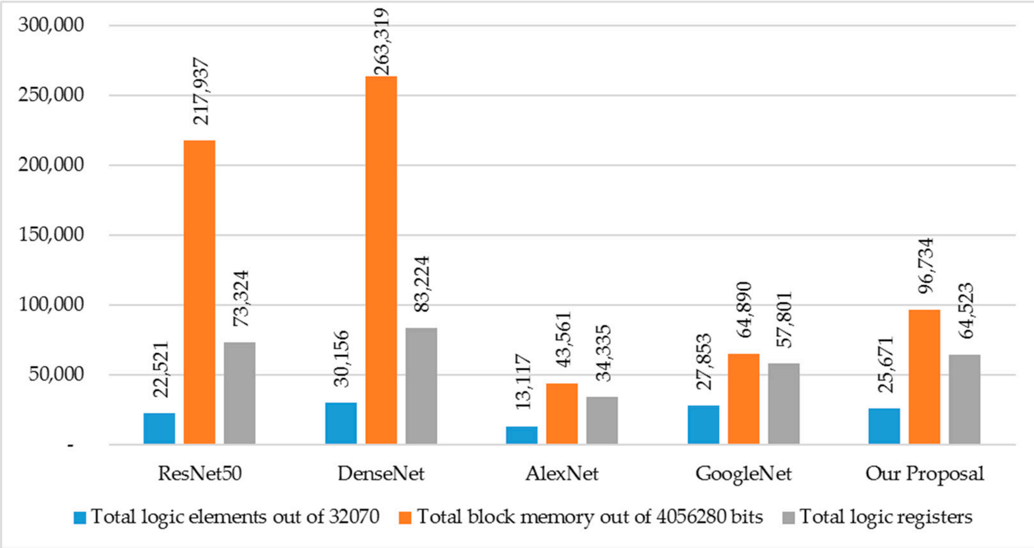


Figure 8. Numbers of resources for DE1-Soc.

6. Conclusions and Future Works

Recently, the interest in employing DL has increased, specifically CNNs, to highlight the race and ethnicity classification issue. In our research, a fresh dataset was acquired to be used in the training phase of a new DL proposal aimed at the ethnic identification of citizens. Therefore, photos taken in three separate countries were used: China, Pakistan, and Russia. This unique dataset is thought to be the first of its kind to be gathered for ethnic groups of individuals, and it was made accessible to the scientific community upon request.

In our study, several specialized high-performance computing (HPC) hardware such as field-programmable gate arrays (FPGAs) and graphics processing units (GPUs) were considered, which resulted in the development of classifier accelerators in order to increase the efficiency of the network in terms of power consumption and execution time, among other benefits. On the other hand, the conducted experiments reported that GPUs required more power consumption and are unsuitable for its future adoption as a portable embedded system, as the approach based on FPGA (type DE1-SoC), which is the one with the best performance to speed up the identification system while consuming less power. Moreover, our DL-based proposal was compared against four different pre-trained CNN models such as ResNet50, DenseNet, AlexNet, and GoogleNet. The experimental results reported that our model outperformed all the methods of state-of-the-art, achieving an accuracy and F1 score value of 96.9 percent and 94.6 percent, respectively.

Author Contributions: Conceptualization, A.J.A.A., L.A. and M.A.F.; methodology, A.J.A.A., L.A. and M.A.F.; software, A.J.A.A., L.A. and M.A.F.; validation, Z.X., A.A., B.H.T. and A.J.H.; formal analysis, A.J.A.A., L.A., M.A.F. and J.S.; investigation, L.A., M.A.F. and Z.X.; resources, J.S., L.A. and M.A.F.; data curation, L.A. and M.A.F.; writing—original draft preparation, A.J.A.A., L.A. and M.A.F.; writing—review and editing, L.A., M.A.F., A.J.H., O.A.-S., B.H.T., J.S. and A.A.; visualization, L.A. and M.A.F.; supervision, Z.X. and M.A.F.; project administration, L.A., M.A.F. and J.S.; funding acquisition, J.S., A.J.H. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The gathered dataset has been uploaded to the ULR below: https://drive.google.com/drive/folders/1SFDOlCbI4DmhSPdMiwiZ5hoBpSX8F6z9?usp=sharing (accessed on 2 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gallo, P.; Pongnumkul, S.; Nguyen, U.Q. BlockSee: Blockchain for IoT video surveillance in smart cities. In Proceedings of the 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), Palermo, Italy, 12–15 June 2018; pp. 1–6.
2. Al-Yassin, H.; Mousa, J.I.; Fadhel, M.A.; Al-Shamma, O.; Alzubaidi, L. Statistical accuracy analysis of different detecting algorithms for surveillance system in smart city. *Indones. J. Electr. Eng. Comput. Sci.* **2020**, *18*, 979–986. [CrossRef]
3. Kardas, K.; Cicekli, N.K. SVAS: Surveillance Video Analysis System. *Expert Syst. Appl.* **2017**, *89*, 343–361. [CrossRef]
4. Darabant, A.S.; Borza, D.; Danescu, R. Recognizing Human Races through Machine Learning—A Multi-Network, Multi-Features Study. *Mathematics* **2021**, *9*, 195. [CrossRef]
5. Cosar, S.; Donatiello, G.; Bogorny, V.; Garate, C.; Alvares, L.O.; Bremond, F. Toward Abnormal Trajectory and Event Detection in Video Surveillance. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 683–695. [CrossRef]
6. Vo, T.; Nguyen, T.; Le, T. Race Recognition Using Deep Convolutional Neural Networks. *Symmetry* **2018**, *10*, 564. [CrossRef]
7. Dagnes, N.; Marcolin, F.; Nonis, F.; Tornincasa, S.; Vezzetti, E. 3D geometry-based face recognition in presence of eye and mouth occlusions. *Int. J. Interact. Des. Manuf.* **2019**, *13*, 1617–1635. [CrossRef]
8. Khan, A.; Marwa, M. Considering race a problem of transfer learning. In Proceedings of the IEEE Winter Applications of Computer VisionWorkshops, Waikoloa Village, NI, USA, 7–11 January 2019; pp. 100–106.
9. Jian-wen, H.A.O.; Lihua, W.; Lilongguang, L.; Shourong, C. Analysis of morphous characteristics of facial reconstruction and the five organs in Chinese north five national minorities crowd. *J. Chongqing Med. Univ.* **2010**, *35*, 297–303.
10. Fu, S.; He, H.; Hou, Z.-G. Learning Race from Face: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2483–2509. [CrossRef]
11. Alzubaidi, L.; Fadhel, M.A.; Al-Shamma, O.; Zhang, J.; Santamaría, J.; Duan, Y. Robust application of new deep learning tools: An experimental study in medical imaging. *Multimedia Tools Appl.* **2021**, 1–29. [CrossRef]
12. Alzubaidi, L.; Duan, Y.; Al-Dujaili, A.; Ibraheem, I.K.; Alkenani, A.H.; Santamaría, J.; Fadhel, M.A.; Al-Shamma, O.; Zhang, J. Deepening into the suitability of using pre-trained models of ImageNet against a lightweight convolutional neural network in medical imaging: An experimental study. *PeerJ Comput. Sci.* **2021**, *7*, e715. [CrossRef]
13. Alzubaidi, L.; Fadhel, M.; Al-Shamma, O.; Zhang, J.; Santamaría, J.; Duan, Y.; Oleiwi, S. Towards a Better Understanding of Transfer Learning for Medical Imaging: A Case Study. *Appl. Sci.* **2020**, *10*, 4523. [CrossRef]
14. Nasser, A.R.; Hasan, A.M.; Humaidi, A.J.; Alkhayyat, A.; Alzubaidi, L.; Fadhel, M.A.; Santamaría, J.; Duan, Y. IoT and Cloud Computing in Health-Care: A New Wearable Device and Cloud-Based Deep Learning Algorithm for Monitoring of Diabetes. *Electronics* **2021**, *10*, 2719. [CrossRef]
15. Arias-Vergara, T.; Klumpp, P.; Vasquez-Correa, J.C.; Noeth, E.; Orozco-Arroyave, J.R.; Schuster, M. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Anal. Appl.* **2021**, *24*, 423–431. [CrossRef]
16. Lauriola, I.; Lavelli, A.; Aiolli, F. An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools. *Neurocomputing* **2021**, *470*, 443–456. [CrossRef]
17. Alzubaidi, L.; Al-Amidie, M.; Al-Asadi, A.; Humaidi, A.J.; Al-Shamma, O.; Fadhel, M.A.; Zhang, J.; Santamaría, J.; Duan, Y. Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data. *Cancers* **2021**, *13*, 1590. [CrossRef] [PubMed]
18. Alzubaidi, L.; Abbood, A.A.; Fadhel, M.A.; AL-Shamma, O.M.R.A.N.; Zhang, J. Comparison of hybrid convolutional neural networks models for diabetic foot ulcer classification. *J. Eng. Sci. Technol.* **2021**, *16*, 2001–2017.
19. Cust, E.E.; Sweeting, A.J.; Ball, K.; Robertson, S. Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *J. Sports Sci.* **2018**, *37*, 568–600. [CrossRef] [PubMed]
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
21. Al-Shamma, O.; Fadhel, M.A.; Hameed, R.A.; Alzubaidi, L.; Zhang, J. Boosting convolutional neural networks performance based on FPGA accelerator. In *International Conference on Intelligent Systems Design and Applications*; Springer: Cham, Switzerland, 2018; pp. 509–517.
22. Frasser, C.F.; de Benito, C.; Skibinsky-Gitlin, E.S.; Canals, V.; Font-Rosselló, J.; Roca, M.; Ballester, P.J.; Rosselló, J.L. Using Stochastic Computing for Virtual Screening Acceleration. *Electronics* **2021**, *10*, 2981. [CrossRef]
23. Coe, J.; Atay, M. Evaluating Impact of Race in Facial Recognition across Machine Learning and Deep Learning Algorithms. *Computers* **2021**, *10*, 113. [CrossRef]
24. Nassih, B.; Amine, A.; Ngadi, M.; Azdoud, Y.; Naji, D.; Hmina, N. An efficient three-dimensional face recognition system based random forest and geodesic curves. *Comput. Geom.* **2021**, *97*, 101758. [CrossRef]
25. Klare, B.; Jain, A.K. On a taxonomy of facial features. In Proceedings of the 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), Washington, DC, USA, 27–29 September 2010; pp. 1–8.

26. Zhang, Y.; Zhang, E.; Chen, W. Deep neural network for halftone image classification based on sparse auto-encoder. *Eng. Appl. Artif. Intell.* **2016**, *50*, 245–255. [CrossRef]
27. Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. HCP: A Flexible CNN Framework for Multi-Label Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1901–1907. [CrossRef] [PubMed]
28. Dagnes, N.; Marcolin, F.; Vezzetti, E.; Sarhan, F.R.; Dakpé, S.; Marin, F.; Nonis, F.; Mansour, K.B. Optimal marker set assessment for motion capture of 3D mimic facial movements. *J. Biomech.* **2019**, *93*, 86–93. [CrossRef]
29. Gogić, I.; Ahlberg, J.; Pandžić, I.S. Regression-based methods for face alignment: A survey. *Signal Process.* **2020**, *178*, 107755. [CrossRef]
30. Li, X.; Lai, S.; Qian, X. DBCFace: Towards Pure Convolutional Neural Network Face Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**. [CrossRef]
31. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.
32. Parka, B.H.; Oha, S.Y.; Kim, I.J. Face alignment using a deep neural network with local feature learning and recurrent regression. *Expert Syst. Appl.* **2017**, *89*, 66–80. [CrossRef]
33. Chen, J.; Ou, Q.; Chi, Z.; Fu, H. Smile detection in the wild with deep convolutional neural networks. *Mach. Vis. Appl.* **2016**, *28*, 173–183. [CrossRef]
34. Ahmed, E.; Jones, M.J.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3908–3916.
35. Marvasti-Zadeh, S.M.; Cheng, L.; Ghanei-Yakhdan, H.; Kasaei, S. Deep learning for visual tracking: A comprehensive survey. *IEEE Trans. Intell. Transp. Syst.* **2021**. [CrossRef]
36. Ronao, C.A.; Cho, S.-B. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Syst. Appl.* **2016**, *59*, 235–244. [CrossRef]
37. Paul, R.; Hawkins, S.; Balagurunathan, Y.; Schabath, M.; Gillies, R.; Hall, L.; Goldgof, D. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival among Patients with Lung Adenocarcinoma. *Tomography* **2016**, *2*, 388–395. [CrossRef] [PubMed]
38. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
39. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [CrossRef] [PubMed]
40. Sze, V.; Chen, Y.-H.; Yang, T.-J.; Emer, J.S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proc. IEEE* **2017**, *105*, 2295–2329. [CrossRef]
41. Véstias, M.P. A Survey of Convolutional Neural Networks on Edge with Reconfigurable Computing. *Algorithms* **2019**, *12*, 154. [CrossRef]
42. Umuroglu, Y.; Fraser, N.J.; Gambardella, G.; Blott, M.; Leong, P.; Jahre, M.; Vissers, K. Finn: A framework for fast, scalable binarized neural network inference. In Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, New York, NY, USA, 22–24 February 2017; pp. 65–74.
43. Seng, K.P.; Lee, P.J.; Ang, L.M. Embedded Intelligence on FPGA: Survey, Applications and Challenges. *Electronics* **2021**, *10*, 895. [CrossRef]
44. Laith, A.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 1–74.
45. He, K.; Wang, Y.; Hopcroft, J. A powerful generative model using random weights for the deep image representation. *arXiv* **2016**, arXiv:1606.04801.
46. Albdairi, A.J.A.; Xiao, Z.; Alghaili, M. Identifying Ethnics of People through Face Recognition: A Deep CNN Approach. *Sci. Program.* **2020**, *2020*, 6385281. [CrossRef]