Journal of Engineering and Applied Sciences 14 (8): 2672-2686, 2019
ISSN: 1816-949X

Burglar Detection using Deep Learning Techniques

[1]Rabia Riaz, [2]Sanam Shahla Rizvi, [1]Ayesha Mushtaq, [1]Sana Shokat and [3]Se Jin Kwon
[1]Department of CS and IT, University of Azad Jammu and Kashmir,
42714 Muzaffarabad, Pakistan
[2]Raptor Interactive (Pty) Ltd., Eco Boulevard, Witch Hazel Ave,
0157 Centurion, South Africa
[3]Department of Computer Engineering, Kangwon National University,
25806 Samcheok, Republic of Korea

Abstract: Burglar detection security systems have become a necessity in this age because of the increasing break-in cases in urban cities thus making these systems essential for residential as well as office usage. This study investigated how to model an intrusion detection system based on deep learning. Two deep learning approaches named generic Deep Neural Networks (DNN) and Convolution Neural Networks (CNN) are used for the training of the dataset. The experimental results showed that CNN approach is more suitable for burglar detection as it gives high accuracy with a superior performance as compared to the generic DNN approach. CNN provides a new research method with the improved accuracy of human intrusion detection. Experimental results found that CNN is compatible to solve classification problems and significantly faster and precise as compared to traditional object detection methods.

Key words: Burglar detection, security, intrusion detection, convolution neural networks, deep neural networks, human intrusion detection, object detection

INTRODUCTION

The increasing deep incorporation of technology in the society has changed the ways people live research and study. Technology is used to solve problems in the space of not only cyber security but also personal security as these threats are becoming serious day by day. Identifying the invader, especially, in unforeseen attacks is increasingly difficult and has become an inevitable technical issue.

Technically designed system for the detection of unauthorized entries is called burglar or security alarm. Two types of security alarms are common named single purpose security alarms and combination purpose security alarms. Single purpose security alarm serves only burglary protection and combination system serves both intrusion protection and fire protection. To record the performance of invasions, burglar alarms are also used in combination with CCTV systems.

To identify different types of attacks, machine learning methodologies are widely used. Traditional machine learning methodologies based on shallow learning emphasize feature selection and engineering. Shallow learning is not suitable for the high dimensional massive data as compared to the deep learning techniques. Deep learning theory was proposed by Professor Hinton *et al.* (2006). Deep learning is a part of machine learning methods which are based on learning data representations in contrast to the task specific algorithms. Deep learning is also called as a hierarchical learning or deep structured learning. Learning can be unsupervised, semi supervised and supervised (Farabet *et al*., 2010; Bengio *et al*., 2013; Schmidhuber, 2015; LeCun *et al*., 2015). Deep learning models are slackly correlated to communication pattern and information processing in a biological nervous system i.e., neural coding. Neural coding creates a relationship between associated neuronal responses and various stimuli (Olshausen and Field, 1996). Neural networks are developed to mimic the neural function of human brain, therefore, they are called artificial neural networks. These are mathematical functions which are executed on serial computers. Most of the research in the area of neural networks is done by development in mathematics and engineering rather than biology (Goodfellow *et al*., 2016).

Corresponding Author: Se Jin Kwon, Department of Computer Engineering, Kangwon National University, 25806 Samcheok, Republic of Korea

Deep learning architectures such as Recurrent Neural Networks (RNN), deep belief networks and Deep Neural Networks (DNNs) are applied to areas such as drug design, bioinformatics, machine translation, social network filtering, audio recognition, natural language processing, speech recognition and computer vision (Ghasemi *et al*., 2017). In some cases, neural networks have provided superior results (Ciregan *et al*., 2012). Deep learning uses multiple layers of nonlinear processing units to extract feature. In deep learning each succeeding layer use the output of the proceeding layer as its own input (Deng and Yu, 2014). Modern deep learning models are usually based on artificial neural network consisting of proportional formulas (Bengio, 2009) or latent variables which are ordered layer wise like nodes in deep Boltzmann machines and deep belief networks.

ANN (Artificial Neural Network) with multiple hidden layers is called DNN. These multiple hidden layers lie between output and input layers (Schmidhuber, 2015; Bengio, 2009). DNN represent non-linear complex relationships. DNN's structural design creates compositional models which express the objects as a layered composition of primordial (Szegedy *et al*., 2013). The extra layers in DNN enables the network to model the complex data potentially even by using fewer units as compared to a shallow network (Bengio, 2009). Generic DNN are mostly designed with a forward network through which data can only flow in forward direction. It means in generic DNN data can flow forward without looping back from input layer to output layer. But in case of Convolutional Neural Networks (CNNs) data can flow in any direction.

CNN is used for different applications like language modeling, pattern detection, image detection and others (Gers and Schmidhuber, 2001; Sutskever *et al*., 2014; Gillick *et al*., 2015; Mikolov *et al*., 2010). CNN is also applied for acoustic modeling to recognize automatic speech which is named as Automatic Speech Recognition (ASR) (Sainath *et al*., 2013). CNN is also used in computer vision (LeCun *et al*., 1998). CNN is a special case of Neural Network. Set of convolutional filters are pooled together to shape a convolutional level of neural network (Fukushima, 1988). Region with Convolutional Neural Network (R-CNN) is considered as a significant method for object detection by providing first practical solution with the use of CNN technique. Forward computation of CNN is executed individually for every object proposal even if the proposals overlap each other or originate from the same image (Girshick, 2015).

A neural network is trained to approximate target output by selecting all neuron's weight from the known input. It is difficult to solve the weights of neurons analytically. An effective and simple solution is provided by back propagation algorithm to solve the weights iteratively. The classical version of neural network uses gradient descent which is quite time consuming and even it is not guaranteed to find the global minimum error. But it research well with a proper configuration known as hyper-parameters. In an algorithm forward propagation of an input vector is done through the neural networks. Before doing forward propagation the weights of network neurons are initialized to small random values. By using a loss function, the comparison of the received output is done with the desired output of network. Computed gradient l oss function is called an error value. Error value is a difference between desired and current output. To calculate the error values they are propagated (Goodfellow *et al*., 2016; Bishop, 2006).

Even though many techniques based on Intrusion Detection Systems (IDSs) are proposed before they could not achieve the desired results. In this study a deep learning approach has been proposed for human intrusion detection by using CNN-IDS and DNN-IDS. The research evaluates the model's performance in binary classification and the effects of different learning rates on accuracy. The study contributes to advancing the state of research by using thermal images for burglar detection given that thermal images are rarely used for burglar detection. We provide a comparison of generic DNN and CNN, identifying technique that is faster and accurate for burglar detection. Our proposed system of Burglar alarms can be used in military, industrial, commercial and residential properties to protect against property damage and burglary (theft).

Literature review: Previous literature on object detection identified classification problems which motivated the researchers to investigate further on object detection. Joseph Redmon and colleagues worked on real time object detection and presented a new object detection approach. They have framed object detection as a regression problem to associated class probabilities and bounding boxes which are spatially separated. In one evaluation, class probabilities and bounding boxes are predicted directly from one image by a single neuron. They presented a unified architecture which was extremely fast. They processed images in real-time at 45 frames/sec. When generalized from natural images to artwork, YOLO (You Only Look Once) model outperforms other detection

methods, i.e., DPM and R-CNN (Redmon *et al*., 2016). Joseph Redmon and colleagues reframed object detection as a single regression problem which was studied straight from class probabilities to bounding box coordinates and image pixels. It was concluded that by using YOLO system at an image it can be predicted that what an object is and where the object is actually located.

Leilei Jin and Hong Liang conducted research on deep learning for underwater image recognition in small sample size situations in 2017. The aim of their study was to find a resolution to underwater image detection in the situations where sample size is small. They tried to solve the underwater image recognition problem by developing an effective framework with the help of underwater cameras in the open sea to capture the set of fish images. They projected framework in a small sample size for underwater image recognition. To suppress fish image's noise an improved median filter was utilized. Images from Image-Net were used to pre-train and employ a CNN. Image-Net is a world's largest data base for image recognition. To test the classification performance and fine tune the pre-trained neural network, fish images were used which were pre-processed. Experimental results of this study demonstrated that the approach of CNN is proficient to recognize fish species. According to their study CNN provides an effective way to solve small sample size recognition tasks (Jin and Liang, 2017).

A research was conducted by Kikuo Asai1 and Norio Takase on finger motion estimation based on frequency conversion of Electromyogram (EMG) signals and image recognition using CNN. EMG based systems have been developed to smoothly control a robot hand. In their study, a simple CNN Model was used to estimate finger motion with the help of classified images produced by EMG signals via. a wavelet transformation. Originally this model has been used for document recognition. Simple CNN Model contains two pairs of pooling, convolutional and two fully connected layers. Inexpensive sensor devices were used to compose a prototype system which was fabricated to capture finger motion and acquire EMG signals. Their experimental study demonstrated that the test results showed 83% accuracy while classifying EMG signals. EMG signals were classified into 4 types, i.e. when a thumb is closed or open and when fingers are closed or open exempting of thumb. In this study, classification of EMG signals was considered as a problem of image recognition. 2D images were used as a data input to CNN. These 2D images were obtained from a wavelet transformation of EMG signals. In this study class labels of finger motion were considered as data output of CNN (Asai and Takase, 2017).

A study was conducted on image recognition by using deep convolutional network in 2015. Basically, this study was conducted on food image recognition by using DNN with fine tuning and pre-training of data. This study determined the Deep Convolutional Neural Network's (DCNN's) effectiveness for the task of food image detection. Fine grained visual recognition is generally called food recognition which is harder problem as compared to the conventional image detection. To overcome the problem of fine grained visual recognition Keiji Yanai and colleague used the best combination of DCNN related techniques, i.e., fine tuning and pre-training. For pre-training they used large scale image net data. The activation features were extorted from pre-trained DCNN. For this experimental study they made 2000 categories of image net data consisting of 1000 food related categories. 78.77% accuracy was achieved for UECFOOD100 and 67.57% accuracy was achieved by UEC-FOOD256. After conducting a challenging experimental study they concluded that these were the best results achieved, so far in the same field. Moreover, they also applied the food classifier to Twitter photo data to utilize the best combination of DCNN techniques. Keiji Yanai and colleague concluded that DCNN was a very compatible technique for large scale image data. Moreover, they revealed that by using GPU only 0.03 sec were taken by DCNN to classify one food photo (Yanai and Kawano, 2015).

A research conducted by Madhumita and MinXu (2017) was on image recognition based on facial micro-expression recognition at small dataset by using deep learning. Minute muscle changes in the face are called facial micro-expression which indicates that a person is either unconsciously or consciously controlling his mental health and true emotions. Hence, the field of micro-expression recognition has increased the scope of research efforts in both areas of computer vision and psychology. Madhumita and Min Xu have used hand crafted features for instance optical flow, gabor filter, Local Binary Pattern-Three Orthogonal Planes (LBP-TOP). For the difficult tasks of face recognition they have used generic deep convolutional neural system which proved very effective for face recognition task. Hence, this study has investigated the possible use of deep learning for face recognition (micro-expression recognition). Extensive training datasets were used to develop reliable DNN. Moreover, large number of labeled image samples was

used for micro expression recognition. Due to the short duration and repressed facial appearance, micro expression recognition has become a quite challenging task. This challenging task also resulted in lack of training data. In this study, researchers have used data augmentation on following two data bases named CASME and CASMEII to generate extensive synthetic images training datasets. Both datasets were combined later to tune a reliable CNN based micro expression recognizer.

After the whole experimentation (Takalkar and Xu, 2017 concluded that CNN is an effective and reliable technique for micro expression recognition. Their study has contributed in the research in following ways: first of all they proposed a unique CNN architecture which gave satisfactory accuracy on micro-expression image recognition. Secondly, they gave a unique idea by combining two databases named CASEME and CASEMEII to boost up sample size for training CNN Model.

Ren *et al.* (2015) worked on real time object detection with region proposal networks and they introduced a Region Proposal Network (RPN). With detection network, RPN shares fill image features of CNN to enable cost free region proposals. RPN simultaneously predict object scores and object bounds at each position. High quality region proposals are generated by RPN after having end to end training. These high-quality proposals are used for detection by fast R-CNN. To share convolutional features, fast R-CNN and RPN could be trained with a simple alternating optimization. With their detection system a 5fps frame rate was achieved on a GPU with 73.2% mAP object detection accuracy on PASCAL VOC2007 and 2012 (Ren *et al.*, 2015).

Recently, deep learning technique has given reliable results in the area of face recognition. Especially, CNN has provided promising results. Still it is an open question that how to design a best CNN architecture and why CNN works well. Previous studies have focused on the best results of CNN and its architecture but the reason that why CNN works well is not investigated yet. Hu *et al.* (2015) have conducted an extensive research on the assessment of CNN-based Face Recognition System (CNNFRS) and tried to find the reason of how and why CNN works well. In their study, they have used public database LFW to train CNN in contrast to the private data bases which are commonly used to train CNN. LFW refers to labeled faces in the wild. They proposed three architectures of CNN and these are the pioneer reported architectures which were trained on LFW. In their study, a quantitative comparison of CNN's architecture has been made. Moreover, the implementation effect of different choices of CNN architecture has been studied. Further, they have identified a number of constructive features of CNN-FRS. Without adverse effect on the accuracy of face recognition, the dimensionality of the learned features can be significantly reduced (Hu *et al.*, 2015).

Li *et al.* (2015) worked on face detection with larger optical deviations i.e., lighting, expression and pose and made discriminative model which was used to accurately identify faces. Although, their created model was highly expensive to operate to deal with conflicting challenges they proposed cascade architecture. This cascade architecture was built on CNN having an influential discriminative capability. It was built to maintain high performance. Proposed CNN cascade operates at multiple resolutions which vary from low to high resolution. At low resolution it rejects background regions quickly and identifies the face similarly in high resolution it evaluates challenging candidates. Further, they introduced calibration stage to progress localization effectiveness. To adjust the detection window, position the output of each calibration stage is used as an input to the following stage. For VGA-resolution images, 14 FPS on single CPU core was used to run proposed model and to detect public face benchmarks 100 FPS was used on GPU. Hence, an up to date detection performance was achieved (Li *et al.*, 2015).

Another study was conducted by McCann and Reeshman in 2014 on face detection by using CNN. They chose neural network and applied it to detect object classification problem with the use of well-known datasets. With the help of CNN researcher have achieved best classification performance. For their dataset they used ZCA whitened grayscale images which gave good results. Results were measured by Kaggle leader board ranking. CNN technique is quite effective for face detection.

In the field of face detection another study was conducted by a number of researchers together to study implementation and evaluation of image recognition algorithm for an intelligent vehicle by using heterogeneous multi-core SoC. Now days image recognition has become very important technology for an intelligent vehicle application i.e., ADAS. ADAS refers to Advanced Driver Assistance System. ADAS is quite expensive to use for computation. For image recognition, special heterogeneous multi-core SoC was developed which could be operated on low power consumption (Tanabe *et al.*, 2012). After this research multiple

applications for image recognition were developed with the use of SoC. In Image recognition algorithms and ADAS applications were addressed and evaluated on SoC. Evaluations showed that in comparison to the general-purpose CPU, SoC runs the application with low power expenditure (Ozaki *et al*., 2015).

A recent study on pedestrian detection with a large field of deep network was conducted by Angelova *et al*. (2015). According to their study, pedestrian detection is very important for autonomous driving applications. Better accuracy could be achieved by deep learning methods, especially for pedestrian detection. In case of pedestrian detection miss rate is very important and deep learning techniques resolve this problem and give better results. For pedestrian detection, they presented a Large Field of View (LFOV) deep network. LFOV deep network was designed to detect faster detection problems. LFOV was used to make classification decisions accurately and simultaneously at multiple locations. Large image areas can be processed at higher speed with the help of LFOV network as compared to the typical deep networks. Even LFOV can reuse computations intrinsically. Pedestrian detection solution achieved 35.85% average miss rates at 280 msec per image on GPU. Pedestrian detection solution was a combination of standard deep network and LFOV network (Angelova *et al*., 2015).

Face verification or face recognition is still a challenging task from accuracy point of view. Although, number of studies is being conducted on face detection by using different techniques but still there is gap in getting accuracy. Therefore, Taigman *et al*. (2014) conducted an experimental study on Deep face to close the gap on human level performance in face verification. In case of modern face detection technique there are four stages of conventional pipeline. These four stages include detect, align, represent and classify. They revisited both representation and alignment steps. To apply a piecewise affine transformation both steps were revisited by engaging a clear 3D Face Model. Moreover, they derived a face representation from DNN which was having nine-layers. With the use of a number of locally connected layers, DNN was consisting of more than 120 million parameters without sharing weight as compared to the layers of standard convolutional network. Therefore, for the training of dataset they used up to date largest facial dataset. It was an identity labeled dataset which was consisting of almost 4 million facial images. These facial images were belonging to more than 4000 identities. The learned representations gave an accurate model-based alignment even with a simple classifier and in the unstable environment. Hence, the accuracy of face detection was 97.35% on LFW dataset (Labeled Face in the Wild) and the reduced error was not more than 27% which approached nearly to the human level performance (Taigman *et al*., 2014).

An advance research in the field of pedestrian detection has been done by Tian *et al*. (2015). They worked together on deep learning for pedestrian detection. They made a theoretical contribution in this field by transferring the learned features of CNN to pedestrian. CNN was pre-trained with massive object categories, i.e., image net. Learned features of CNN were made able to control variations like lighting, viewpoints and poses. Although, features were strong enough but still there was present some probability of their failure with complex occlusions. In pedestrian detection, occlusion handling is an important problem. In contrast to the deep models they proposed deep parts.

Deep models consist of single detector but deep parts consist of extensive part detectors with a number of appealing properties. Appealing properties of deeppartsare as follows: It can b etrained on weekly labeled data. It can control low IOU positive proposals. These proposals shift away from ground truth. In deep parts each part detector can distinguish pedestrian with the help of part of proposals. Therefore, each part detectoris consider edasa highly strong detector. Results of experiments revealed that deeppartsare highly effective with a missrate of 11.89% and out performed second best method by 10% (Tian *et al*., 2015).

Another study was conducted on real time object detection through 3D CNN by Maturana and Scherer (2015). They found that in the real-world environment strong object detection is a critical skill to operate robots autonomously. In modern robotic systems, the range sensors i.e., RGBD and LIDAR provide a rich source of 3D information. Already developed systems have trouble to deal with large amount of point cloud data. They proposed VoxNet architecture to deal with point cloud data efficiently by 3D CNN. They evaluated their proposed system on publicly available benchmarks by using CAD data, RGBD and LiDAR. VoxNet achieved accuracy more than previous systems with labeling hundreds of instances per second (Maturana and Scherer, 2015).

MATERIALS AND METHODS

This study presents the research methodology adapted and the issues encountered during the data collection. Figure 1 shows the steps defined for research design.

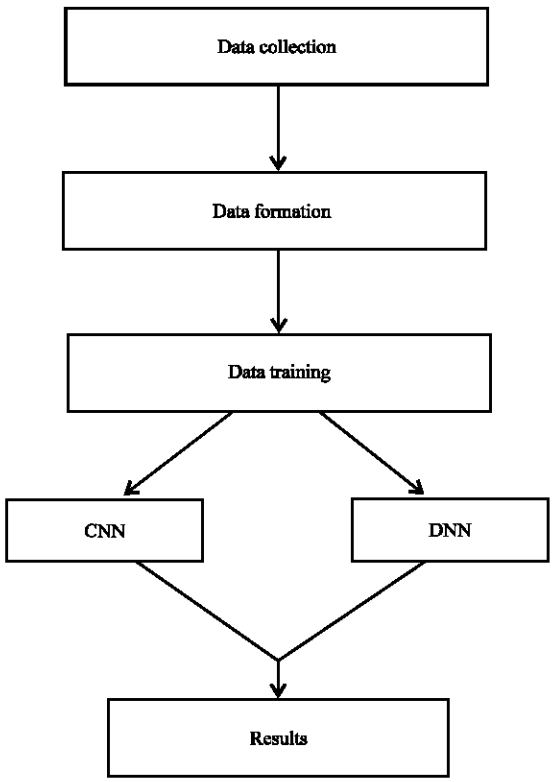

Fig. 1: Steps involved in research

Data collection: This study collected its dataset from OTCBVS benchmark-OSU thermal pedestrian database acquired by Raytheon 300 D thermal sensor (Anonymous, 2005, 2014) and pedestrian infrared/visible stereo video dataset acquired by FLIR thermo vision A40M (Anonymous, 2017).

OSU thermal pedestrian database: OSU thermal pedestrian database is a new and up-to-date benchmark dataset which is available publicly to test and evaluate high-tech computer vision algorithms. The benchmark consists of images and videos which are recorded not only in the visible spectrum but also beyond the visible spectrum (Fig. 2). Thermal image sensors are having a phantom sensitivity usually and this sensitivity ranges from 7-14 μ wavelength. The capacity of thermal image devices to capture images largely depends on their reflectivity and emissivity. The surface properties and material of entities control emissivity. And their reflectivity is affected by the amount of background radiations which are reflected by the entities. When multiple factors are involved in the image formation procedure then a number of distortions occur in thermal images, especially, radiometric distortions, hotspot areas and halo effect (Goodall *et al*., 2016). There are 10 sequences in OSU thermal pedestrian database which constitute 284 images in total. These images have 8-bit grayscale bitmap format. The sampling rate of these sequences is non-uniform. This database basically covers a diversity of environmental conditions such as sunny, rainy and cloudy days.

Pedestrian infrared/visible stereo video dataset: Pedestrian infrared stereo video datasets are the datasets for pedestrian detection which are fully visible by using both stereo and mono vision (Fig. 3). It may contain huge size human action video. Further, it includes manually interpreted figured data. In pedestrian infrared/visible stereo video dataset, there are four infrared-visible pairs between 100 and 4400 frames and 25819 ground truth point pairs. In total, this study used above 1500 images for the analysis.

Issues involved in data collection: The study encountered some issues while collecting the data which are as follows: the datasets collected during the research lacked the ground truth files, so, the study found it difficult to locate the bounded boxes. The thermal image datasets are not large. Most of the thermal image datasets were constituted on small number of images.

Dataset formation: By collecting the thermal images from the two datasets, the next step of the study is formatting the datasets according to the needs. The study divides the images in two categories that are positive images and negative images. The positive images are those which have human in trusion in them (Fig. 4). The yare considered as burg larpresence. The negative images are those which do not have any human in trusion they are considered as burglar absence (Fig. 5). The number of images in positive category is 1,589 and then umber of images in negative category is 1, 5 89 as well. The total number of images is 3, 178.

Training of dataset: The third step of the research is training of acquired dataset. For this purpose, the study uses two techniques named as generic DNN and CNN. Both of the techniques lie under the same umbrella which is deep learning. Deep learning theory was proposed by Professor Hinton *et al*. (2006). Deep learning technology gained a dramatic boost up in the machine learning field. In this situation, practical research findings and theoretical papers have created remarkable achievements, especially in the area of action recognition, image recognition and speech recognition (Hinton *et al*., 2006).

Fig. 2: a-d) OSU thermal pedestrian database (thermal images)

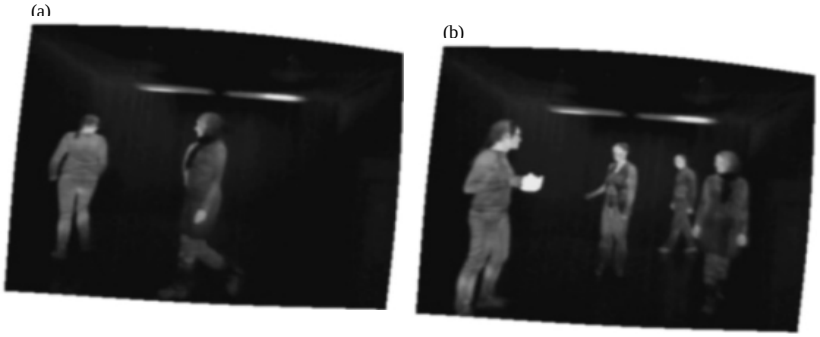

Fig. 3: a, b) Pedestrian infrared/visible stereo video dataset

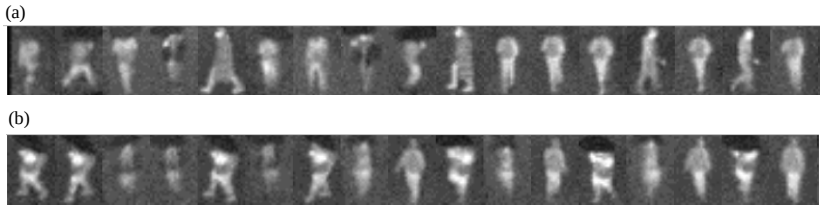

Fig. 4: a, b) Positive images (Burglar present)

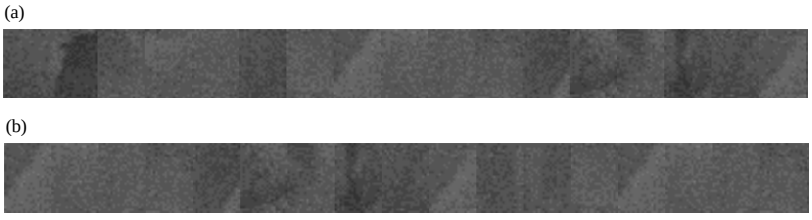

Fig. 5: a, b) Negative images (Burglar absent)

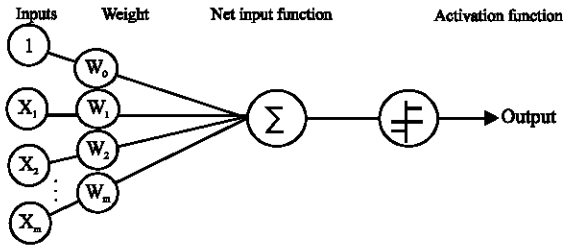

Fig. 6: Computation process inside a layer of neural network

Deep Neural Networks (DNNs): DNN is a deep learning technique with multiple hidden layers in between input and output layers. The layers are comprised of nodes. A node is only a room where calculation happens. A node has some coefficients or learnable weights hich are combined with the input using dot product. These coefficients or learnable weights either intensify or reduce the input, thus, allocating importance to the input. The dot product of all the nodes in a layer are summed and passed to another node's activation function and from there the result is obtained from an output laye (Fig. 6).

DNNs are usually feed forward neural networks in which data streams from input layer towards the output layer. It does not provide back propagation which means that this is a unidirectional network. DNN can be trained in an unsupervised means where the networks learn relevant features by learning to reproduce its input and it can also be trained in a supervised way that calibrates the features in order to classify.

DNNs can model hard non-linear relationships. DNN architectures generate compositional models where the object is articulated as a layered composition of primitives Fig. 7. The extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network. Because of abstraction layers the DNNs are disposed to over fitting. This over fitting allows a DNN to map the unusual dependencies in the training data. To overcome the over fitting effect of DNN, regularization methods are applied such as sparsity, weight decay and lvakhnenko's unit pruning. The other common issue with DNN is computational speed. Because of the deep nature of the network and its size, the DNN computational speed becomes less. To increase the computational speed batching trick is used. Batching is a process in which gradient is computed on numerous training samples at once rather than single samples.

Convolutional Neural Networks (CNNs/ ConvNets): CNN is a significant tool for deep learning and it is mostly suitable for image recognition. Two directional data flows are provided by CNN as compared to generic DNN which provides single directional data flow. Hence, in case of CNN, data can flow in both forward and backward direction but in case of generic DNN data can flow only in forward direction.

CNN is composed of neurons which are having learnable weights and biases Fig. 8. In the whole procedure of CNN a dot product is performed by a neuron after receiving inputs and then output/dot product is optionally followed with non-linearity. Finally, a fully linked output layer is achieved. Input which is received by the ConvNets is normally in form of the images. ConvNets convert 2-3D input volume into a 2-3D volume output. In ConvNets every layer not only contains parameters as well as supplementary hyper parameters. ConvNets consist of neurons which are arranged in three dimensions in contrast to regular neural network. Three dimensions are as follows: depth, height and width. Simple convolutional network is composed of layer's sequence. All layers of ConvNet converts input (volume of activation) to output (other volume of activation) via. differentiable function. To build ConvNet architecture researchers used three major kinds of layers after the input layer. These layers are as follows: convolutional layer, pooling layer and fully-connected layer.

Convolutional layer: Convolutional layer is the foundation of a CNN as it contributes to the whole structure. Layer parameters comprised of kernels which are basically a set of learnable filters. Convolutional layer figures out the neuron's output which is linked to the input's local region (Fig. 9). There are three basic characteristics of convolutional layer.

3D volumes of neurons: Neurons of the layers are arranged in 3 proportions named depth, height and width. Neurons within a layer are linked to a receptive field which is a small region of the layer at its starting point. Divergent kind of layers including completely connected and locally connected are piled together for the composition of CNN architecture.

Local connectivity: In CNN spatial locality is exploited with the help of a local connectivity pattern which is implemented between the neurons of the adjacent layers. Thus, structural design ensures strongest response to a

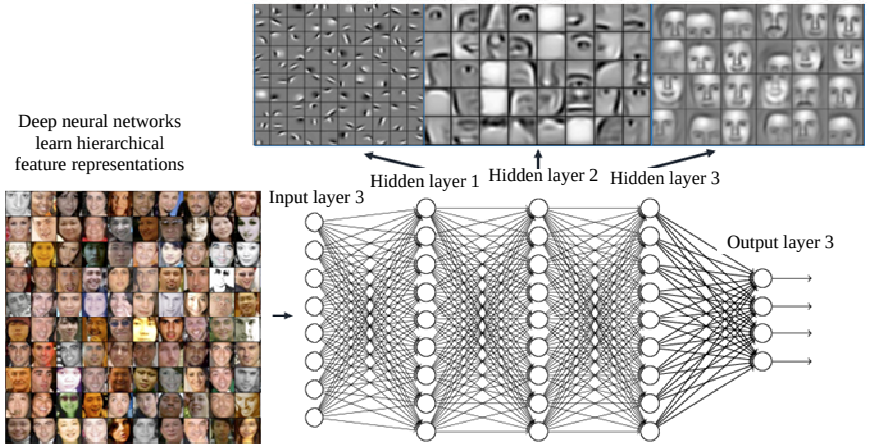

Fig. 7: How a deep neural network works

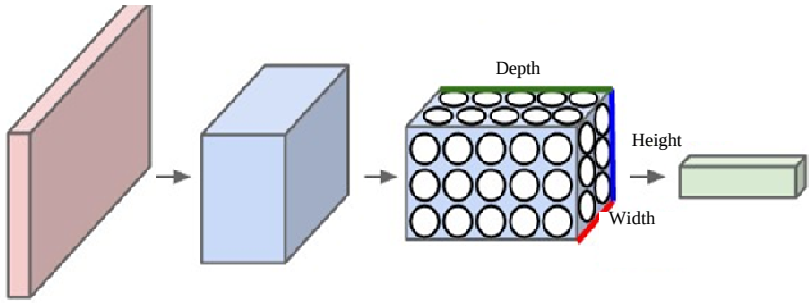

Fig. 8: A ConvNet arranges its neurons in three dimensions

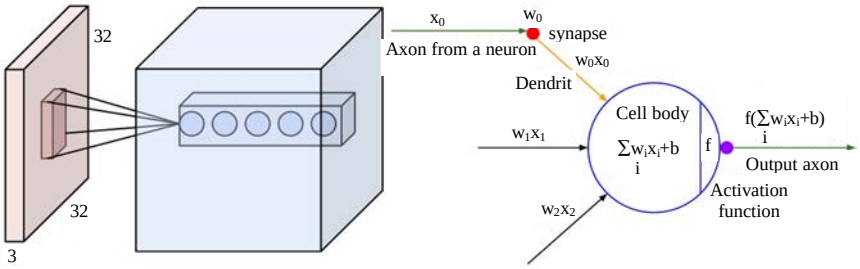

Fig. 9: Convolutional layer

spatial local input pattern which is produced by the learnt filters. Pile of these adjacent layers direct non-linear filters which are becoming progressively more universal. Local connectivity enables the network to assemble representations of larger areas by creating representations of small parts of input.

Shared weights: In CNN each filter is simulated across the whole optical field. Same parameterization; bias and weight vector is shared by these replicated units. Moreover, these replicated units form a feature map. It means that in a given convolutional layer all the neurons within their specific response field act in a response to the same feature. Therefore, regardless of the position all the replicating units allow the features to be detected in the visual field. Hence, replicating units constitutes the property of translation invariance.

Pooling layer: Pooling layer executes a down sampling operation by the side of spatial dimensions. And these spatial dimensions comprise of height and width. Down sampling operation results in volume such as [16×16×3] (Fig. 10).

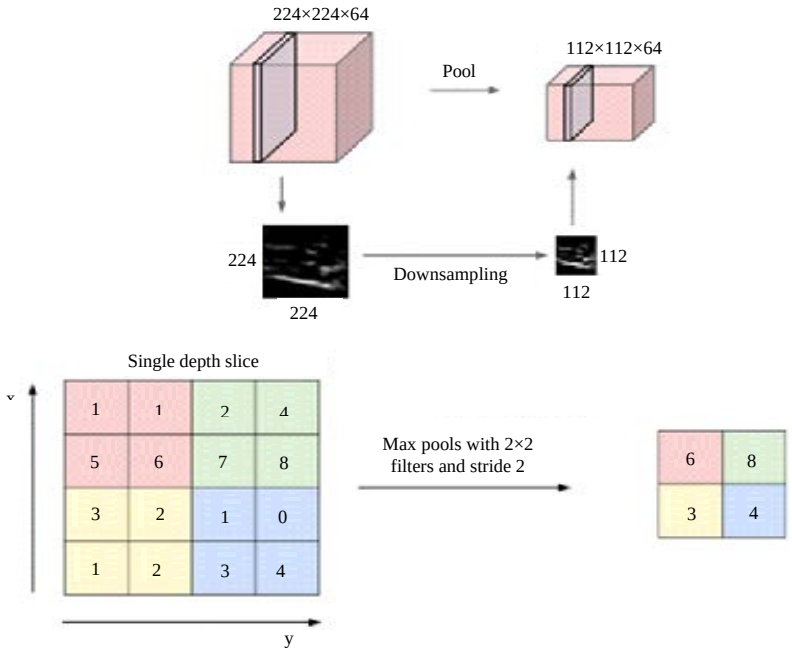

Fig. 10: Pooling or down sampling

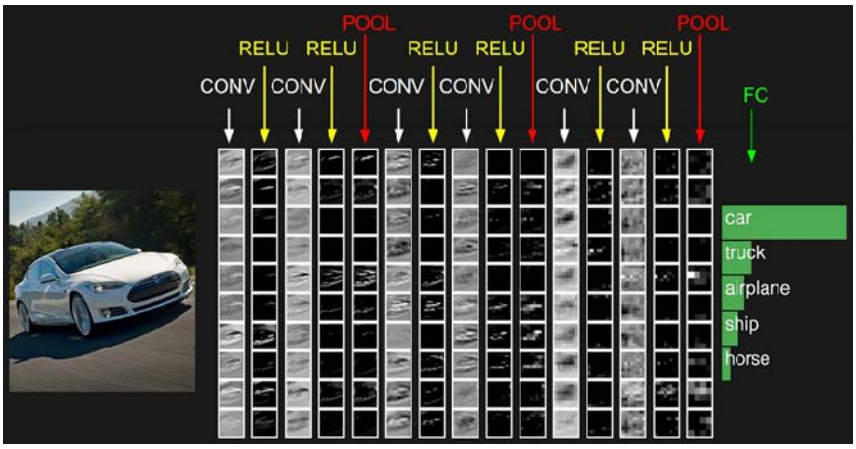

Fig. 11: Complete ConvNet view

Fully-connected layer: In the neural network, fully connected layer provides high level reasoning after some max pooling and convolutional layers. In fully connected layer neurons are connected to all previous layer activations and this characteristic is similar to the regular neural network. Activations of the neurons can be figured out with a matrix multiplication tracking out bias offset Fig. 11.

RESULTS AND DISCUSSION

This study presents the results of the current experimental study. First, this study covers the general results of DNN technique and then it provides the detailed results of R-CNN approach. Then, it provides the comparison between both DNN and CNN approach. Moreover, we also discuss later the supporting and contrary points from the previous studies to support our results and to show the differences in our research from other related studies. The aim of this section is to analyze obtained results and to discuss them with references to the previous studies from literature.

DNN training: The burglar detection was measured using two deep learning techniques that are generic DNN and CNN. The performance was evaluated by detecting the

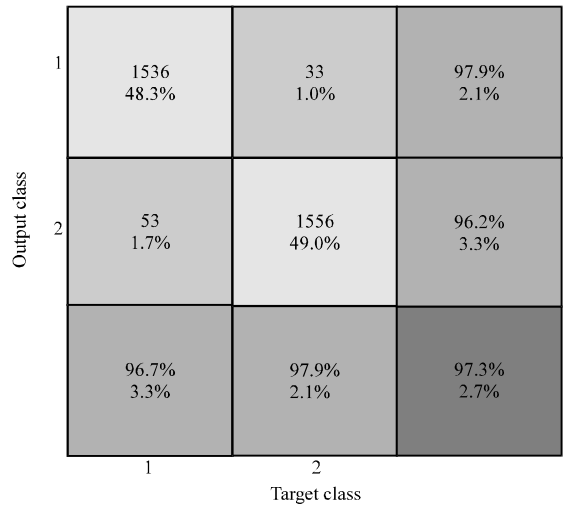

Fig. 12: Confusion matrix of DNN

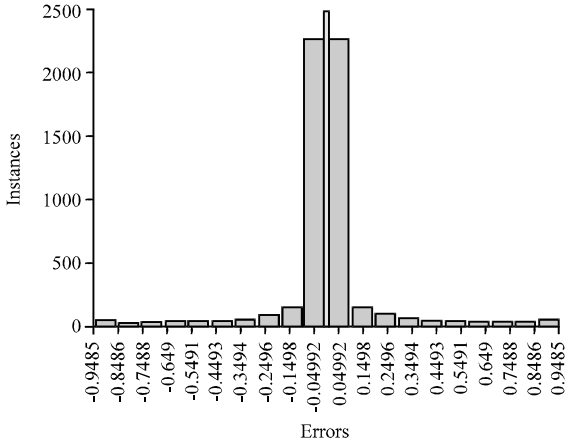

Fig. 13: Error histogram of DNN (Zero error)

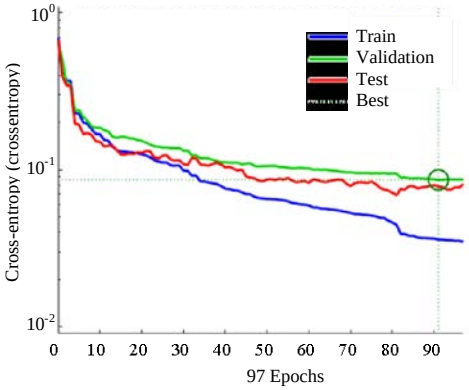

Fig. 14: Performance plot of DNN; Best validation performance is 0.086064 at epoch 91

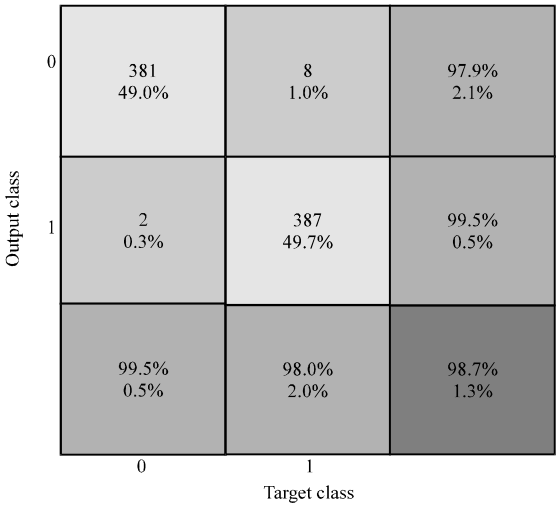

Fig. 15: Confusion matrix of CNN

accuracy. Generic DNN architecture is formed by using 2 hidden layers with number of neurons 75 for each. The confusion matrix of DNN training in the Fig. 12 shows that 1536 images are correctly classified as burglar presence which corresponds to 48.3% of all images. Similarly, 1556 images are correctly classified burglar absence which corresponds to 49.0% of all images. 53 of the burglar absence images are incorrectly classified as burglar presence which corresponds to 1.7% of all images in the data. Similarly, 33 of the burglar presence images are incorrectly classified as burglar absence and this corresponds to 1.0% of all data.

Out of 1569 predictions, 97.9% are correct and 2.1% are wrong. Out of 1609 predictions, 96.7% are correctly predicted as burglar absence and 3.3% are wrong. Out of 1589 burglar presence cases, 96.7% are correctly predicted as burglar presence and 3.3% are predicted as burglar absence. Out of 1589 burglar absence cases, 97.9% are correctly predicted as burglar absence and 2.1% are predicted as burglar presence. The overall training results of generic DNN are 97.3% accurate and there is 2.7% error in the results. Figure 13 shows the error histogram of DNN. The plot shows the error values of the network. Figure 14 shows the performance of DNN. The validation performance is best at epoch 91 where the validation error is lowest.

CNN training: CNN is designed by using input size 32×20×1 and the filter size is 5×5 and there are 50 activation maps in total. The max pooling layer is of size 2×2 with a stride of 2. The maximum number of epoch is 10. The initial learning rate is 0.0009. Figure 15 shows the training of the dataset using CNN. As discussed earlier, the base learning rate is 0.0009 which remains same throughout the training process. The mini-batch accuracy is less in the 1st epoch and the mini-batch loss value is

higher. With each passing epoch the mini-batch loss is decreasing and the mini-batch accuracy is increasing. At the 10th epoch, the network stops as it h as finished the training and computes the accuracy. With the current architecture of DNN, the network gives 98.7% accuracy and 1.3% error.

The confusion matrix of CNN training in Fig. 16 shows that 381 images are correctly classified as burglar presence which corresponds to 49.0% of all images. Similarly, 387 images are correctly classified burglar absence which corresponds to 49.7% of all images. The 2 of the burglar absence images are incorrectly classified as burglar presence which corresponds to 0.3% of all images in the data. Similarly, 8 of the burglar presence images are

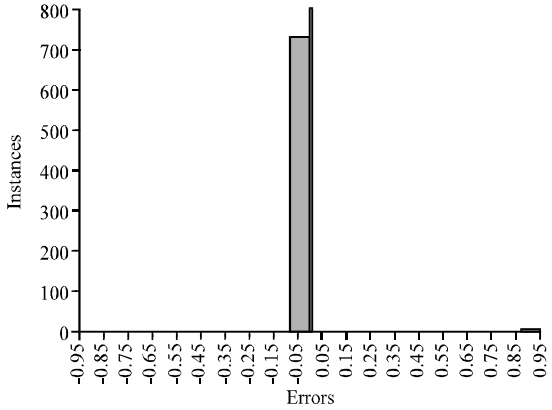

Fig. 16: Error histogram of CNN (Zero error)

Table 1: Training on single CPU and initializing image normalization; (Training of dataset using CNN)

Epoch	Iteration	Time elapsed (sec)	Mini-batch loss	Mini-batch accuarcy (%)	Base learning rate
1	1	2.66	0.8561	61.72	0.0009
3	50	48.71	0.1810	94.53	0.0009
5	100	95.45	0.2087	94.53	0.0009
8	150	142.22	0.0184	99.22	0.0009
10	200	188.74	0.0142	100.00	0.0009
10	210	198.24	0.1279	98.44	0.0009

Table 2: Techniques used in related studies and achieved accuracy

Title	Researchers	Technique used	Accuracy (%)
You only look once:unified, real time object detection	Redmon *et al.* (2015)	YOLO Model, Fast-CNN, combination of YOLO model and R-CNN	Fast R-CNN gets a 3.2% improvement from the combination with YOLO Model
Deep learning for underwater image recognition in small sample size situations	Jin and Liang (2017)	CNN	85.08
Finger motion estimation based on frequency conversion of EMG signals and image recognition using convolutional neural network	Asai1 and Takase (2017)	CNN	83
Food image recognition using deep convolutional network with pre-training and fine-tuning	Yanai and Kawano (2015)	DCNN	78.77
Image based facial micro-expression recognition using deep learning on small datasets	Madhumita and MinXu (2017)	CNN on CASME and CASMEII database	75.57
Faster R-CNN: towards real-time object detection with region proposal networks	Ren *et al.* (2015)	Region Proposal Network (RPN) and fast R-CNN	73.2
When face recognition meets with deep learning: an evaluation of CNN for face recognition	Hu *et al.* (2015)	CNN-FRS (Face Recognition System)	CNN-S = 0.7828 CNN-M = 0.7882 CNN-L = 0.7807
A CNN cascade for face detection	Li *et al.* (2015)	CNN	87.48
Object detection using CNN	McCann and Reeshman, 2016	CNN using ZCA whitened grayscale images	61.16
Implementation and evaluation of image recognition algorithm for an intelligent vehicle using heterogeneous multi-core SoC	Nau *et al.* (2015)	Heterogeneous multi-core SoC	64.4
Pedestrian detection with a large-field of-view deep network	Angelova *et al.* (2015)	Large Field Of View (LFOV) deep network was designed to detect faster detection problems	Pedestrian detection solution achieved 35.85 average miss rates at 280 msec per image on GPU
Deep face: closing the gap to human-level performance in face verification	Taigman *et al.* (2014)	DNN	97.35
Deep learning strong parts for pedestrian detection	Tian *et al.* (2015)	CNN	Results of experiments revealed that deep parts are highly effective with a miss rate of 11.89% and outperformed second best method by 10%
VoxNet: A 3D CNN for real-time object recognition	Maturana and Scherer (2015)	3D-CNN	0.92

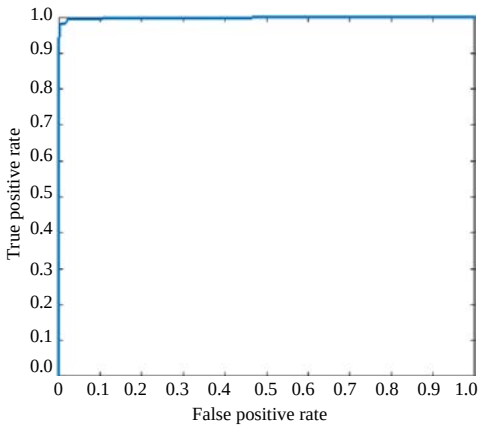

Fig. 17: ROC curve of CNN

incorrectly classified as burglar absence and this corresponds to 1.0% of all data. Out of 389 predictions, 97.9% are correct and 2.1% are wrong. Out of 389 predictions, 99.5% are correctly predicted as burglar absence and 0.5% is wrong. Out of 383 burglar presence cases, 99.5% are correctly predicted as burglar presence and 0.5% is predicted as burglar absence. Out of 395 burglar absence cases, 98.0% are correctly predicted as burglar absence and 2.0% are predicted as burglar presence. The overall training results of CNN are 98.7% accurate and there is 1.3% error in the results.

Figure 17 shows the error histogram of CNN. The plot shows the error values of the network. Figure 18 shows the ROC curve plot of CNN. The plot shows the relationship between the True Positive (TP) values and False Positive (FP) values. Results show that the TP rate is high as compared to FP rate which represents the higher accuracy rate. The area under the curve is 0.9982.

A summary table is presented to compare the previous techniques used for the object detection (Table 1 and 2). The presented table analyzes and interprets the findings or results of those techniques. Given source evidences and findings provided strong support to our findings. Moreover, it provides the comprehensive comparison of the different techniques and their findings.

CONCLUSION

The current study works on the development of human intruder tracking system by using pattern recognition in day and night time. For the training of dataset, this research used two techniques named Deep Neural Network (DNN) and Convolutional Neural Network (CNN). The findings of current study showed that CNN technique works really well to deal with a large dataset taken from real world and it gives more accurate results. The study found that DNN is suitable for smaller datasets. As the dataset increases, the accuracy of DNN starts decreasing because of it's over fitting problem. CNN technique works better than generic DNN although, both implicitly capture the structure of an image. The high rate of feature extraction makes the CNN distinct from the other neural networks. A fairly highly sensitive and positive projecting value can be achieved with CNN approach as CNN gives 98.7% accuracy as compared to DNN.

Results of the comparative study suggest that CNN boosts the system's performance and endow with better accuracy due to its exceptional characteristics for instance local connectivity and shared weight. CNN alleviates the traditional problem, so, it is a better option than other applications of deep learning methods relating to natural language processing and computer vision.

LIMITATIONS

There are certain limitations in the current study in spite of gaining high accuracy with CNN and the designed data set. Firstly, a medium sized data set was used in the study instead of large data set. Although, accuracy could be improved with the help of using large data set. Lessons learned from the study could be used further to get improved accuracy for object detection.

RECOMMENDATIONS

With this proposed framework, future research can be conducted on convolutional neural network for underwater image recognition especially with small sample size. This kind of research will be very helpful for commercial applications i.e., aquaculture and fisheries which can be used for marine defense. Moreover, proposed framework can also be implemented on mobile devices in future. The present study can also be extended to the analysis of food distribution. An improved accuracy can be achieved by employing CNN on micro-expression recognition. Further, only 2D data set was used for the face detection of the invaders. An improvement in this study can be done by the use of 2D and 3D data fusion. As 3D data set can eliminate all those problems which are encountered in 2D data set.

ACKNOWLEDGEMENTS

"This research was supported by Basic Science Research through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B04031440). This study was also supported by 2017 Research Grant from Kangwon National University".

REFERENCES

Angelova, A., A. Krizhevsky and V. Vanhoucke, 2015. Pedestrian detection with a large-field-of-view deep network. Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), May 26-30, 2015, IEEE, Seattle, Washington, USA., ISBN:978-1-4799-6924-1, pp: 704-711.

Anonymous, 2005. OTCBVS benchmark dataset collection. Visual Computing and Image Processing Lab (VCIPL), Oklahoma State University–Stillwater, Stillwater, Oklahoma.

Anonymous, 2014. Welcome to VCIPL. Visual Computing and Image Processing Lab (VCIPL), Oklahoma State University–Stillwater, Stillwater, Oklahoma. http://vcipl-okstate.org/.

Anonymous, 2017. [Laboratory of interpretation and image and video processing (LITIV)]. Polytechnique Montreal, Montreal, Canada. (In French) http://www.polymtl.ca/litiv/vid/BilodeauetAlInfraredDataset.zip.

Asai, K. and N. Takase, 2017. Finger motion estimation based on frequency conversion of EMG signals and image recognition using convolutional neural network. Proceedings of the 2017 17th International Conference on Control, Automation and Systems (ICCAS), October 18-21, 2017, IEEE, Jeju, South Korea, ISBN:978-1-5386-1025-1, pp: 1366-1371.

Bengio, Y., 2009. Learning deep architectures for AI. Found. Trends Mach. Learn., 2: 1-127.

Bengio, Y., A. Courville and P. Vincent, 2013. Representation learning: A review and new perspectives. IEEE. Trans. Pattern Anal. Mach. Intell., 35: 1798-1828.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, Berlin, Germany, ISBN:9780387310732, Pages: 738.

Ciregan, D., U. Meier and J. Schmidhuber, 2012. Multi-column deep neural networks for image classification. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-21, 2012, IEEE, Providence, Rhode Island, ISBN:978-1-4673-1226-4, pp: 3642-3649.

Deng, L. and D. Yu, 2014. Deep learning: Methods and applications. Found. Trends Signal Proc., 7: 197-387.

Farabet, C., B. Martini, P. Akselrod, S. Talay and Y. LeCun *et al*., 2010. Hardware accelerated convolutional neural networks for synthetic vision systems. Proceedings of the 2010 IEEE International Symposium on Circuits and Systems (ISCAS), May 30-June 2, 2010, IEEE, Paris, France, ISBN:978-1-4244-5308-5, pp: 257-260.

Fukushima, K., 1988. Neocognitron: A hierarchical neural network capable of visual pattern recognition. Neural Networks, 1: 119-130.

Gers, F.A. and E. Schmidhuber, 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE. Trans. Neural Networks, 12: 1333-1340.

Ghasemi, F., A.R. Mehridehnavi, A. Fassihi and H. Perez-Sanchez, 2017. Deep neural network in biological activity prediction using deep belief network. Appl. Soft Comput., 62: 275-278.

Gillick, D., C. Brunk, O. Vinyals and A. Subramanya, 2015. Multilingual language processing from bytes. Comput. Lang., 5: 1-11.

Girshick, R., 2015. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV'15), December 7-13, 2015, IEEE, Computer Society Washington, DC, USA., ISBN:978-1-4673-8391-2, pp: 1440-1448.

Goodall, T.R., A.C. Bovik and N.G. Paulter, 2016. Tasking on natural statistics of infrared images. IEEE. Trans. Image Proc., 25: 65-79.

Goodfellow, I., Y. Bengio, A. Courville and Y. Bengio, 2016. Deep Learning Cambridge. MIT Press, Cambridge, Massachusetts, USA., Pages: 778.

Hinton, G.E., S. Osindero and Y.W. Teh, 2006. A fast learning algorithm for deep belief nets. Neural Comput., 18: 1527-1554.

Hu, G., Y. Yang, D. Yi, J. Kittler and W. Christmas *et al*., 2015. When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition. Proceedings of the 2015 IEEE International Conference on Computer Vision Workshops, December 7-13, 2015, IEEE, Santiago, Chile, pp: 142-150.

Jin, L. and H. Liang, 2017. Deep learning for underwater image recognition in small sample size situations. Proceedings of the 2017 International Conference on OCEANS 2017-Aberdeen, June 19-22, 2017, IEEE, Aberdeen, UK., ISBN:978-1-5090-5279-0, pp: 1-4.

LeCun, Y., L. Bottou, Y. Bengio and P. Haffner, 1998. Gradient-based learning applied to document recognition. Proc. IEEE., 86: 2278-2324.

LeCun, Y., Y. Bengio and G. Hinton, 2015. Deep learning. Nat., 521: 436-444.

Li, H., Z. Lin, X. Shen, J. Brandt and G. Hua, 2015. A convolutional neural network cascade for face detection. Proceedings of the 2015 IEEE International Conference on Computer Vision and Pattern Recognition, June, 7-12, 2015, IEEE, Boston, Massachusetts, USA., pp: 5325-5334.

Maturana, D. and S. Scherer, 2015. Voxnet: A 3D convolutional neural network for real-time object recognition. Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), September 28-October 2, 2015, IEEE, Hamburg, Germany, pp: 922-928.

Mikolov, T., M. Karafiat, L. Burget, J. Cernocky and S. Khudanpur, 2010. Recurrent neural network based language model. Proceedings of the 11th Annual International Conference on the International Speech Communication Association, September 26-30, 2010, Makuhari, Chiba, Japan, pp: 1045-1048.

Olshausen, B.A. and D.J. Field, 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nat., 381: 607-609.

Ozaki, N., M. Uchiyama, Y. Tanabe, S. Miyazaki and T. Sawada *et al*., 2015. Implementation and evaluation of image recognition algorithm for an intelligent vehicle using heterogeneous multi-core SoC. Proceedings of the 2015 International 20th Asia and South Pacific Conference on Design Automation (ASP-DAC), January 19-22, 2015, IEEE, Chiba, Japan, pp: 410-415.

Redmon, J., S. Divvala, R. Girshick and A. Farhadi, 2016. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, IEEE, Las Vegas, Nevada, USA., ISBN:978-1-4673-8852-8, pp: 779-788.

Ren, S., K. He, R. Girshick and J. Sun, 2015. Faster R-Cnn: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems, Cortes, C., N.D. Lawrence, D.D. Lee. M. Sugiyama and R. Garnett (Eds.). Curran Associates, Inc., New York, USA., pp: 91-99.

Sainath, T.N., A.R. Mohamed, B. Kingsbury and B. Ramabhadran, 2013. Deep convolutional neural networks for LVCSR. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 26-31, 2013, IEEE, Vancouver, Canada, pp: 8614-8618.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Networks, 61: 85-117.

Sutskever, I., O. Vinyals and Q.V. Le, 2014. Sequence to Sequence Learning with Neural Networks. In: Advances in Neural Information Processing Systems, Ghahramani, Z., M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger (Eds.). Curran Associates, Inc., Red Hook, New York, USA., pp: 3104-3112.

Szegedy, C., A. Toshev and D. Erhan, 2013. Deep Neural Networks for Object Detection. In: Advances in Neural Information Processing Systems, Bartlett, P., F.C.N. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger, (Eds.). Curran Associates, New York, USA., pp: 2553-2561.

Taigman, Y., M. Yang, M.A. Ranzato and L. Wolf, 2014. Deepface: Closing the gap to human-level performance in face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 28, 2014, IEEE, New Jersey, USA., pp: 1701-1708.

Takalkar, M.A. and M. Xu, 2017. Image based facial micro-expression recognition using deep learning on small datasets. Proceedings of the 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), November 29-December 1, 2017, IEEE, Sydney, Australia, ISBN:978-1-5386-2840-9, pp: 1-7.

Tanabe, Y., M. Sumiyoshi, M. Nishiyama, I. Yamazaki and S. Fujii *et al*., 2012. A 464GOPS 620GOPS/W heterogeneous multi-core SoC for image-recognition applications. Proceedings of the 2012 IEEE International Solid-State Circuits Conference on Digest of Technical Papers (ISSCC), February 19-23, 2012, IEEE, San Francisco, California, USA., ISBN:978-1-4673-0376-7, pp: 222-223.

Tian, Y., P. Luo, X. Wang and X. Tang, 2015. Deep learning strong parts for pedestrian detection. Proceedings of the 2015 IEEE International Conference on Computer Vision, December 7-13, 2015, IEEE, Santiago, Chile, pp: 1904-1912.

Yanai, K. and Y. Kawano, 2015. Food image recognition using deep convolutional network with pre-training and fine-tuning. Proceedings of the 2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), June 29-July 3, 2015, IEEE, Turin, Italy, pp: 1-6.